



TECNOLOGÍA
en marcha

Quarterly journal
December 2022
Volume 35
ISSN-E 2215-3241



Special issue
**IEEE International Conference
on Bioinspired Processing**



TEC | Tecnológico
de Costa Rica

Publicación y directorio en catálogos

latindex

redalyc.org UAEM

Dialnet

melICA

SciELO

REDIB
Red Iberoamericana
de Investigadores y Científicos

DOAJ

Comisión Editorial

Felipe Abarca Fedullo. Director.
Editorial Tecnológica de Costa Rica

Juan Antonio Aguilar Garib
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León.
México

Carlos Andrés Arredondo Orozco
Facultad de Ingenierías
Universidad de Medellín. Colombia

Lars Köhler
Experimenteller Botanischer Garten
Georg-August-Universität Göttingen.
Alemania

Jorge Solano Jiménez
Instituto Costarricense del Cemento
y del Concreto

Edición técnica

Alexa Ramírez Vega

Revisión filológica

Esperanza Buitrago Poveda

Diseño gráfico

Felipe Abarca Fedullo

Diagramación

Leila Calderón Gómez

Diseño de cubierta

Ariana Sanabria García

Imagen de cubierta

<https://unsplash.com/>

Datos de catalogación en publicación

Tecnología en Marcha / Editorial Tecnológica de Costa Rica. - Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022– Trimestral ISSN-E 2215-3241

1. Ciencia y Tecnología –
Publicaciones periódicas CDD:600



Apdo 159-7050 Cartago, Costa Rica
Tel.:(506) 2550-2297, 2550-2618
Correo electrónico: editorial@itcr.ac.cr
Web: <https://www.tec.ac.cr/editorial>
https://revistas.tec.ac.cr/tec_marcha



La Editorial Tecnológica de Costa Rica es una dependencia especializada del Instituto Tecnológico de Costa Rica. Desde su creación, en 1978, se ha dedicado a la edición y publicación de obras en ciencia y tecnología. Las obras que se han editado abarcan distintos ámbitos respondiendo a la orientación general de la Institución.

Hasta el momento se han editado obras que abarcan distintos campos del conocimiento científico-tecnológico y han constituido aportes para los diferentes sectores de la comunidad nacional e internacional.

La principal motivación de la Editorial es recoger y difundir los conocimientos relevantes en ciencia y tecnología, llevándolos a los sectores de la comunidad que los requieren.

La revista *Tecnología en Marcha* es publicada por la Editorial Tecnológica de Costa Rica, con periodicidad trimestral. Su principal temática es la difusión de resultados de investigación en áreas de Ingeniería. El contenido de la revista está dirigido a investigadores, especialistas, docentes y estudiantes universitarios de todo el mundo.

Publicación y directorio en catálogos





TECNOLOGÍA *en marcha*

Contenidos

IEEE International Conference on Bioinspired Processing (BIP 2021)

Contribuciones especiales a la Iniciativa de Desarrollo de Investigación de la Conferencia

Juan Luis Crespo-Mariño, Mauricio Rodríguez-Calvo, Juan Esquivel-Rodríguez..... 2

Automatic social media news classification: a topic modeling approach

Clasificación automática de noticias en redes sociales: una aproximación desde el modelado de tópicos

Daniel Amador, Carlos Gamboa-Venegas, Ernesto García, Andrés Segura-Castillo..... 4

Automatic image segmentation using Region-Based convolutional networks for Melanoma skin cancer detection

Segmentación automática de imágenes mediante redes convolucionales basadas en regiones para la detección del cáncer de piel tipo melanoma

Karen Dayana Tovar-Parra, Luis Alexander Calvo-Valverde, Ernesto Montero-Zeledón, Mac Arturo Murillo-Fernández, Jose Esteban Perez-Hidalgo, Dionisio Alberto Gutiérrez-Fallas..... 14

Benchmarking the NXP i.MX8M+ neural processing unit: smart parking case study

Evaluando la unidad de procesamiento neuronal NXP i.MX8M+: el caso de estudio de parqueos inteligentes

Edgar Chaves-González, Luis G. León-Vega..... 26

Analysis of a cooling system for robotic joints using a computational fluid dynamics study

Análisis de un sistema de enfriamiento para articulación robótica mediante un estudio de dinámica de fluidos computacional

María Fernanda Madriz-Ramírez, Kevin Alberto Solano-Núñez, Mauricio Rodríguez-Calvo..... 31

Accelerating machine learning at the edge with approximate computing on FPGAs

Acelerando aprendizaje de máquina en el *Edge* con computación aproximada en FPGAs

Luis Gerardo León-Vega, Eduardo Salazar- Villalobos, Jorge Castro-Godínez..... 39

Design of a microkernel-based services manager for IoT with isolated processing

Diseño de un gestor de servicios IoT basado en *microkernel* con procesamiento aislado

Jose Antonio Ortega-González, Luis G. León-Vega..... 46

IEEE International Conference on Bioinspired Processing (BIP 2021)

Contribuciones especiales a la Iniciativa de Desarrollo de Investigación de la Conferencia

Juan Luis Crespo-Mariño¹, Mauricio Rodríguez-Calvo², Juan Esquivel-Rodríguez³

Crespo-Mariño, J.D.; Gamboa-Venega, C.; Rodríguez-Calvo, M.; Esquivel-Rodríguez, J. IEEE International Conference on Bioinspired Processing (BIP 2021). *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 2-3.

 <https://doi.org/10.18845/tm.v35i9.6496>

1 Área Académica de Ingeniería Mecatrónica. Instituto Tecnológico de Costa Rica. Coeditor invitado del número especial.

Correo electrónico: jcrespo@itcr.ac.cr

2 Carrera de Licenciatura en Ingeniería Electrónica, Universidad Técnica Nacional. Coeditor invitado del número especial

3 Escuela de Computación. Instituto Tecnológico de Costa Rica. Coordinador General de la Conferencia BIP 2021

En Costa Rica existe un importante núcleo de investigación científica relacionada con la intersección entre la ingeniería, las ciencias computacionales, las ciencias de la salud y las ciencias naturales. No solo en lo que se refiere al uso de técnicas matemáticas y/o computacionales en la resolución de problemas en, por ejemplo, estudios de poblaciones, biodiversidad, ambiente, etc. sino asimismo en Biología Molecular, Virología, Microbiología, Biomedicina o Agronomía, entre otras.

Estas áreas, que son clásicamente estudiadas desde perspectivas analíticas, se pueden ver muy beneficiadas al introducir técnicas y recursos procedentes de las ciencias exactas y la ingeniería como el modelado matemático, la computación de alto desempeño, procesamiento de imágenes o el uso de softwares de simulación especializados para la validación de diseños; lo que constituye un componente altamente multidisciplinar en los tipos de estudios y resultados de los grupos de investigación que existen en el país.

Estos estudios requieren un entorno multidisciplinario que englobe a los diferentes grupos de investigación para facilitar el intercambio de conocimientos, problemas y experiencia, de tal forma que se fomente la producción científica y que, a su vez, proyecten al exterior las capacidades científicas de Costa Rica y el alineamiento de los objetivos de los grupos de investigación costarricenses con el estado del arte a nivel mundial en cuanto a problemas a resolver y metodologías para ello, con el fin de potenciar las relaciones colaborativas con otros grupos de investigación alrededor del mundo.

La necesidad de generar este espacio motivó a que, profesores de las 5 universidades públicas costarricenses: UCR, TEC, UNA, UNED y UTN; así como investigadores del Centro Nacional de Alta Tecnología (CeNAT) se unieran para crear la Comunidad BIP, que organiza conferencias internacionales con el auspicio de IEEE bajo el nombre “IEEE International Conference on Bioinspired Processing (BIP)”, y celebró su tercera edición en 2021 (<https://www.bipconference.org/home-2021>).

La conferencia tuvo lugar durante los días 4 y 5 de noviembre de forma virtual y se desarrollaron distintas sesiones con presentaciones de artículos científicos, ponencias y presentaciones de posters de parte de estudiantes de maestría y doctorado de distintas universidades, con el fin de acercar a los profesionales en formación para que tuvieran el contacto y vivieran la experiencia que significa participar en una conferencia científica. Esto permitió evidenciar el esfuerzo que dedicaron a cada uno de sus respectivos trabajos de investigación y recibieron retroalimentación de expertos en el área, además que tuvieron la posibilidad de hablar con investigadores que podrían fungir luego como posibles tutores guía, si piensan optar por estudios más avanzados o por algún tipo de estancia o posdoctorado.

En este número especial, se muestran algunos de esos trabajos que reflejan el compromiso de la Comunidad BIP de comunicar el quehacer en investigación. Además, pone en evidencia que las actividades de investigación de calidad y de alto valor realmente forman a los nuevos profesionales de nuestro país a un nivel muy alto, con la capacidad de abordar problemas no resueltos en el mundo en que vivimos.

Automatic social media news classification: a topic modeling approach

Clasificación automática de noticias en redes sociales: una aproximación desde el modelado de tópicos

Daniel Amador¹, Carlos Gamboa-Venegas²,
Ernesto García³, Andrés Segura-Castillo⁴


Amador, D.; Gamboa-Venegas, C.; García, E.; Segura-Castillo, A. Automatic social media news classification: a topic modeling approach. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 4-13.

 <https://doi.org/10.18845/tm.v35i9.6477>

1 Centro Nacional de Alta Tecnología. E-mail: damador@cenat.ac.cr

 <https://orcid.org/0000-0003-1197-4313>

2 Centro Nacional de Alta Tecnología. E-mail: cgamboa@cenat.ac.cr

 <https://orcid.org/0000-0001-9712-0575>

3 Universidad de Costa Rica. E-mail: luis.garciaestrada@ucr.ac.cr

4 Universidad Estatal a Distancia. E-mail: asegurac@uned.ac.cr

 <https://orcid.org/0000-0001-5647-1176>

Keywords

Automatic news classification; social media; topic modeling.

Abstract

Social media has modified the way that people access news and debate about public issues. Although access to a myriad of data sources can be considered an advantage, some new challenges have emerged, as issues about content legitimacy and veracity start to prevail among users. That transformation of the public sphere propels problematic situations, such as misinformation and fake news. To understand what type of information is being published, it is possible to categorize news automatically using computational tools. Thereby, this short paper presents a platform to retrieve and analyze news, along with promising results towards automatic news classification using a topic modeling approach, which should help audiences to identify news content easier and discusses possible routes to improve the situation in the near future.

Palabras clave

Clasificación automática de noticias; Redes sociales; Modelado de tópicos.

Resumen

Las redes sociales virtuales han modificado significativamente la forma en la que las personas acceden a contenido noticioso y, por ende, el debate en la esfera pública. Aunque el acceso a múltiples y diversas fuentes puede considerarse una ventaja, a su vez genera situaciones problemáticas relacionadas con la legitimidad y veracidad del contenido circulante, por ejemplo, desinformación y noticias falsas. Para lograr entender qué tipo de información se está publicando, se puede llevar a cabo una categorización de las noticias por tema, con ayuda herramientas computacionales para realizar este proceso de forma automática. Así, este artículo corto presenta una plataforma para recuperar y analizar noticias, así como resultados prometedores del uso de modelado de tópicos para la clasificación automática de contenido noticioso, en aras de facilitar a la audiencia la categorización del contenido. Asimismo, discute las rutas posibles a seguir para mejorar la propuesta a futuro.

Introduction

Social media has modified the ways people consume information and engage in political participation [1, 2, 3]. Although easy access to a myriad of data sources is a considerable advantage, it has brought some new challenges to the table. One of them has to deal with the amount of news that circulates and how to automatically and accurately classify them. For instance, [4], has pointed out that, given the vast amount of news circulating on social media, it has become more difficult for readers to identify topics according to their content, therefore, making fake news proliferation more possible and thus an issue. Research from the London School of Economics [5] insists that this situation has created a “trust crisis” that leads to a permanent social uncertainty. Moreover, Livingstone [5] argues that politicians, institutions, platforms, news media and activists have not found a viable way to mitigate the issue.

From another perspective, Waisbord [6] points out that the main challenge we have to deal with, is the transformation of the public sphere. Traditional news media legitimacy is at stake, the proliferation of digital spaces to express opinions publicly without any previous editorial filter contributes to a wider debate, but at the same time poses a problem for content validity

and veracity. At the end of the day, behind issues such as massive media content and fake news, there is a dispute between different ideas about legitimacy and power, an “epistemic democracy” has emerged.

The relevance of understanding and mitigating this situation of interest is clear for the scientific community and contributions from computational perspectives to this matter are highly expected. Therefore, a topic modeling approach is proposed as a means to contribute to automatic social media news content detection, thus reducing the uncertainty for the user. The article presents the bases of topic modeling, then gives a description of the implemented approach, shows some preliminary results and finally discusses possible routes for future development and improvement.

Topic modeling

Amongst modern techniques of computational and automatic processing applied to digital humanities and social sciences, topic modeling is an unsupervised machine learning methodology that provides a suitable approach to problems that involve clustering, semantics and discourse analysis. In broad terms, a topic model algorithm receives an input of unclassified texts and outputs a set of topics that are common throughout a corpus. Additionally, it provides a net weight that reflects the degree to which the text belongs to a possible topic.

The topic modeling algorithm to be discussed in this article is a Latent Dirichlet Allocation (LDA). This model was proposed as a bioinspired approach to study population genetics by Prichard, Stephens and Donnelly in 2000, and later applied to computational science by Blei, NG and Jordan in 2003.

According to Blei [7] there are two key assumptions that an LDA model requires in order to be successful:

1. All documents in the corpus share similar word patterns that are called topics.
2. All the documents in the corpus can be allocated within any of this limited amount of topics.

Biel [7] states that the generative process of LDA begins once a corpus is feeded into the system. First, the process chooses an amount of topics, given as a parameter, to distribute the documents. The distribution distributes the contents among the topics using a Dirichlet distribution and then a multinomial one. Afterwards, each document is assigned with a weight to describe what topics have major presence in it. Finally, for every word composing the document the system will choose a topic assignment and will generate a new document with a given probabilistic weight of similarity towards the previous document inputted. As stated by Blei, NG and Jordan (2003), given the parameters α and β , the combined distribution of a set of topics θ , a set of N topics z , and a set of N words w is:

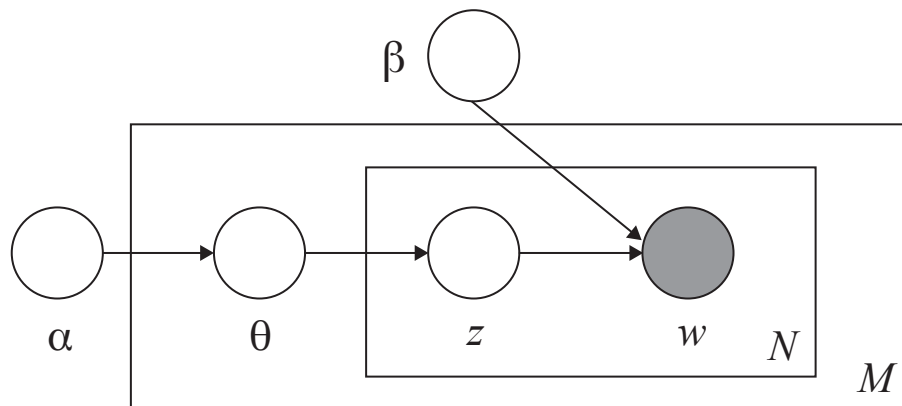


Figure 1. Graphical model representation of LDA retrieved from [8].

Blei, NG and Jordan [8], indicate that the boxes are “plates” representing replicates. The outerplate represents documents, while the inner plate represents the repeated choice of words within a document.

In the wider sense, an LDA model defines topics by two major sources of probabilistic outcomes within an operation: Dirichlet distribution and multinomial distribution. The Dirichlet distributions relate documents-topics and topics-words. These topics are decided by the user to further be computed by the system which will then cluster the words by topic related to the training corpus provided.

Currently, for ease of researching teams in the realm of semantic analysis and Natural Language Processing (NLP), there are open sourced libraries that can be implemented to projects. In the present article GenSim by Řehůřek and Sojka [9] was utilized. GenSim is a pure Python library that provides a memory-efficient and scalable out-of-the-box solution.

Methodology: Implemented approach

Based on previous work from Soto-Rojas et al. [10], Facebook posts from local news pages were automatically retrieved and stored on a database. It is important to mention that the platform that allowed our data retrieval, also includes an interface to request information for public consulting purposes, a Content Management System for expert content validation and administration and some visualization tools to explore the contents available on the platform. Having the benefit of such a platform, allowed us to explore the implementation of a Topic Modeling Approach to automate the classification of the news. At the moment, only 10% of the data, i.e, 91725 posts, have been manually tagged by experts into several categories. The pages from local media selected for this study are: La Nación (the reference newspaper in the country), CR Hoy (the most popular online newspaper), Telenoticias (a long time running news TV program), Repretel (a high rating news TV program), Semanario Universidad (an academic newspaper), El Financiero (a theme based newspaper, mostly focusing on economical issues), Noticias Monumental (a well known news radio program), Prensa Libre (the oldest newspaper), Diario La Extra (a tabloid newspaper) and Amelia Rueda (a popular news radio program). We have collected all 91795 posts for the topic modeling experimentation.

Figure 2 shows a general view of the complete solution, including the steps for data retrieval, storage and the Application Programmers Interface available for public request. Moreover, the process for topic modeling with the visual tools is also included.

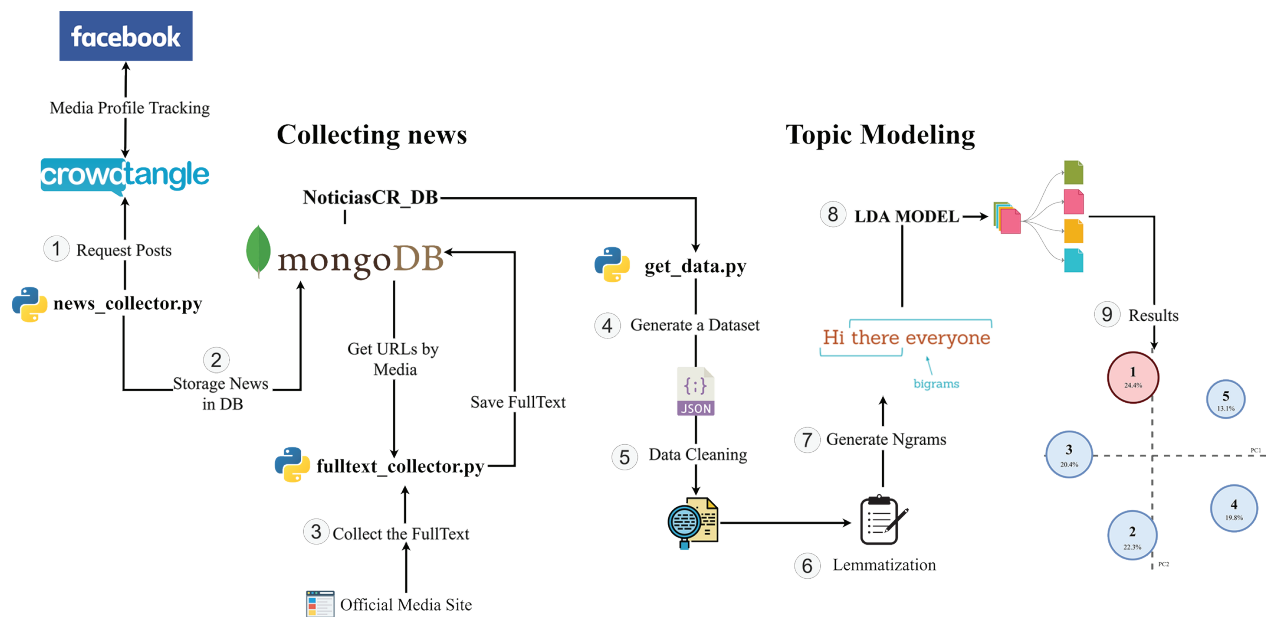


Figure 2. Diagram of the current infrastructure, with the complete execution pipeline of the solution, ending with the analysis of news using Top Modeling with LDA.

Data retrieval and storage

To store retrieved data we selected MongoDB, a document-based database that allows storage of data with different structures with high query efficiency. This database model makes it possible to have different collections in the same storage space. The database is hosted in the Kabre supercomputer at Centro Nacional de Alta Tecnología (CeNAT), and currently stores more than 1 million posts from Costa Rican news sources on Facebook, ranging from 2018 to the moment.

It is very important to mention that we obtained legal access to Facebook post content thanks to an agreement with CrowdTangle, their official source for academic purposes. Thus, we are able to retrieve the posts with all its contents and several aggregated statistics that show their respected engagement. No user data is ever manipulated or stored and the agreement strictly prevents non aggregated data publication [11].

Collecting news takes two steps. The first one gathers posts and their provided data from Facebook profiles of Costa Rican news media and stores them in the database. The second part takes the external link provided with the retrieved data and extracts the complete text from the original news source at their respective official website.

However, using CrowdTangle as a tool to consult Facebook posts, has its limitations. There is a rate limit for requests, when this is overpassed, indexes are restarted, thus causing duplication of data. Usually, it is possible to extract 10,000 entries before this happens. To avoid this issue, the implemented script gathers news daily only.

Then, to feed the database, a Python script is in charge of querying the list of news pages accessed through CrowdTangle, a library called Pytangle allowed to obtain the publications by a date range, in this module the filtering is applied to prevent the rate limit of the requests from being exceeded. In addition, a log was added to prevent duplicity of data by saving the last date recorded in the database, when an error occurs in the collection, the last date is loaded and the execution of the collector is restarted.

News texts retrieval

Each collected post from Facebook has a link to the original news source. Next in the process, that link is used to extract the original full text from the news media publication. Since every website is created with a different technology, several scripts were implemented to extract the text from each one, these programs were unified as a single scrapper. When the program starts, it runs nine scripts, each connecting to the database to add contents to the field "fullText". Each execution retrieves 200 news from each news source.

Additionally to the nine scrappers, the tool includes four scripts written in Python. Two containing information about configuration and connection to the database. And two more with the structure of the data to be read and the main code with the necessary flow to run each scrapper separately.

Topic modeling implementation

With all the posts and full content stored in the database, we proceeded with an exploration of the more topics prevalent in a specific set of news. To do so, an LDA model was used to perform the experiments with 55815 news from four evenly distributed previously tagged categories.

For the controlled experiments, a pipeline was implemented that is responsible for data cleaning, lemmatization, N-gram creation, corpus generation and subsequently the application of the LDA model. Figure 3 illustrates the steps of the data processing pipeline. Using this pipeline, the following cases were executed: identification of 5 topics, identification of 7 topics and identification of 10 topics.

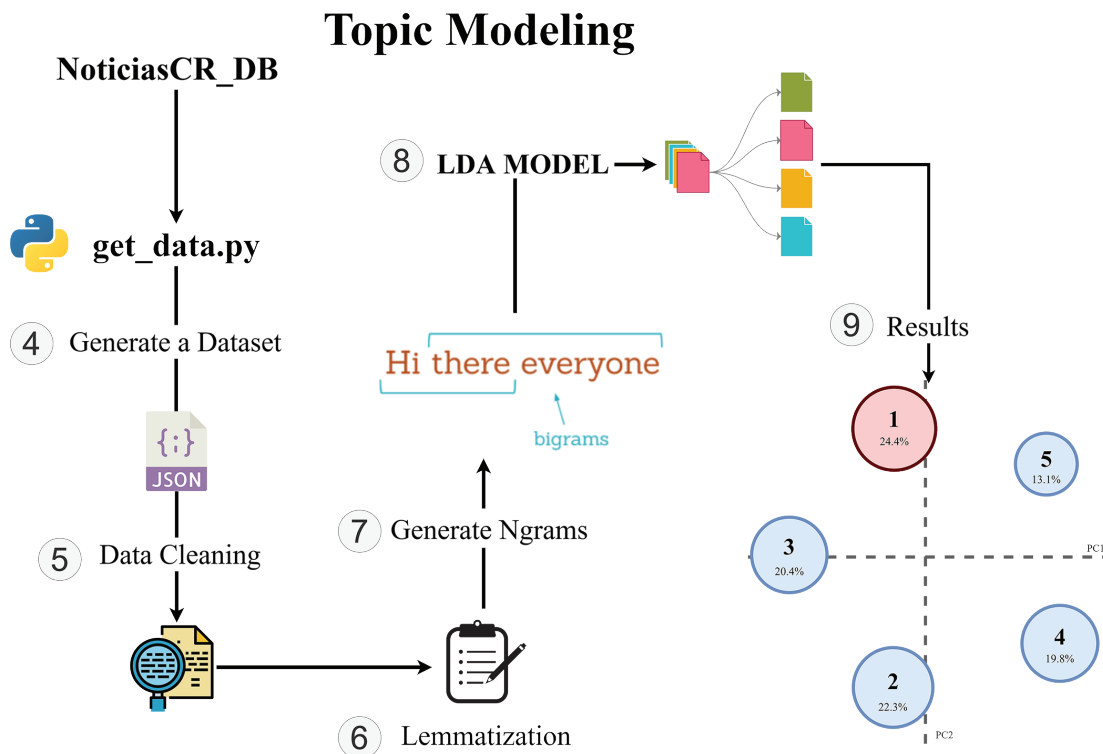


Figure 3. Diagram of the LDA pipeline, with the complete execution of data processing.

For each of the cases a series of data preprocessing steps must be followed. The first step is the selection of the data from the dataset, as mentioned above, an attempt was made to balance the data by selecting a balanced amount of news associated with each topic. With the dataset selected, the second step is the cleaning of the data, where we applied the elimination of stopwords, accents and converted all texts to lowercase. Approximately 1700 words were used as stopwords, which were removed from the texts since they have no significant relevance for the analysis. At the end of the data cleaning, the third step is the generation of the lemmatization, which consists of converting all conjugated words into their base form (e.g. “Robó”, “Robando”, “Robar”) and thus avoid duplication of the meaning of the same word. When the lemmatization is finished, the N frames are created, which correspond to the association of N number of words, in this case, bigrams and trigrams were used. Once the N frames are generated, a corpus is generated with all the word unions already filtered and cleaned to be processed by the model. Finally, the LDA model is applied using the gensim library, which contains a series of models for the study of natural language processing.

The LDA model receives the following elements as parameters: corpus, id2word, num_topics, random_state, update_every, chunksize, passes, alpha. The corpus represents the sparse shape matrix (num_documents, num_terms) and is generated prior to execution, as well as, the id2word which serves to determine the size of the vocabulary. The number of topics was set according to each experiment, so three separate runs were made with 5, 7 and 10 numbers of topics to be identified. For random state a value of 100 was assigned and a value of 1 was added to the update every which represents the number of documents to use in each update, the chunksize which represents the number of documents to use in each training chunk was defined as 100, and the passes which is the number of passes through the corpus during training was set to 10. Finally, the alpha was defined as *auto* which means that the learning strategy is based on the asymmetric prior from the corpus.

At the end of the model training, a result is obtained with the different topics and the most relevant words for each one of them. To visualize these results, a library called pyLDAvis [12] was used to generate a diagram of the relevance of the words and the incidence of the topics in the total number of news items in the dataset. The obtained visualizations are presented in the next section.

Evaluation

To evaluate the model we use the coherence metric [13]. For the experiments we know that words will be grouped into a *topic*. This *topic* group is said to be coherent when the words of the group support each other in the same semantic meaning. For this reason a topic group can be interpreted as such in a context that embraces all or most of the words.

Complimentary, expert criteria were consulted. In terms of the expert criteria, elicited to determine a gold standard to classify the models, the researchers utilized The Semantic Fields Theory (Trier. 1931 as cited by Gliozzo [14]). This theory establishes that any language lexicon is structured into *Semantic Fields*, which is defined as semantic relations among concepts belonging to the same field. In other words, semantic relations between words of the same field have a very close semantic relationship while words that do not belong to the same field are very distant between each other. This can be exemplified by giving a lexicon of words: bird, dog, space, cat, math; where we can identify that bird, dog and cat belong to an “animal’s field” while the other entries could belong to a “scientific field”.

Based on this theory, a consulted expert proposed a classification, when possible, to the topics outputted by the model. As mentioned before, this will serve as a human inputted gold standard from the field of discourse analysis to cross check the performance of the model.

Results

In this section we show preliminary results and visualization of the LDA model with intertopic distance maps. The goal is to explore and analyze if the topics generated by the model can represent a similar distribution of subjects already labeled in the news of the complete dataset. The intertopic distance map is a visualization of the topics in a two-dimensional space. Each topic is represented by a circle, and its size is proportional to the amount of tokens (words) that belong to each topic across the dictionary.

Figure 4 illustrates the results of LDA with 5 topics. Topic 1 has the highest percentage and includes in its top 5 words like: *país*, *mundo*, *precio*, *nivel*, *nacional*. Following with words such as *mercado*, *mundial*, *internacional* in the top 10. This might be evidence that the subject in question is related with International and National news, and maybe economy and market. The smaller topic is number 5 which includes words like *educativo*, *proyecto*, *educación*, *nacional*. Suggesting that the education topic is the least covered in the dataset.

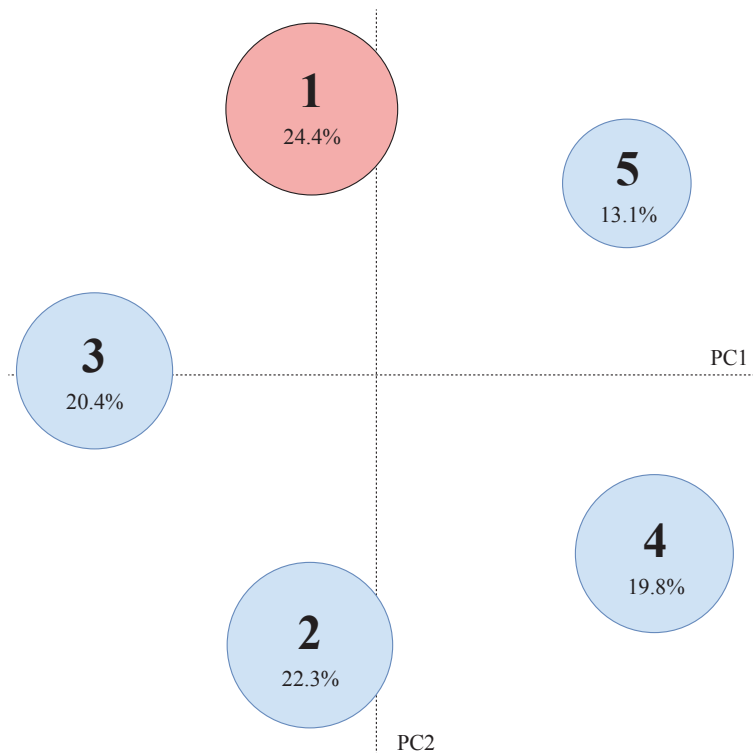


Figure 4. Intertopic Distance Map of LDA with 5 topics, and the percentage of tokens per group.

After running the three experiments, the coherence metric is returned by the algorithm, resulting in 79% for the 5 topics, 70% for the 7 topics and 64% for the 10 topics. This means that the 5 and 7 topic models scored high enough to be considered good candidates for the classification. At this point, expert criteria is needed to determine which model fits best for the purposes of the research.

In table 1, the expert criteria for the three different models are identified, if the words of the topic have not a clear semantic relationship then it will be tagged as unclassifiable.

Table 1. Expert criteria consulted for topic model experiments with 5, 7 and 10 topics. Topics numbers are independent, they do not indicate any relation among experiments.

Topic #	Semantic Field		
	5 topics	7 topics	10 topics
1	International	Unclassifiable	Unclassifiable
2	Health	Unclassifiable	Unclassifiable
3	Unclassifiable	International	International
4	Crime and judiciary	Government affairs	Government affairs
5	Education	Health	Health
6		Unclassifiable	Unclassifiable
7		Crime and judiciary	Crime and judiciary
8			Unclassifiable
9			Unclassifiable
10			Unclassifiable

By expert criteria suggestion, the model that best fits the semantic content of the data is the 5 topic one.

Discussion and conclusions

Results show an effective topic modeling approach towards automatic news classification, which could help news media to classify the topic without depending on an editorial intervention. Given the vast amount of news that circulates on social media, that would also mean a significant advantage for audiences, who would have an accurate classification of their content of interest in advance.

Furthermore, this approach could contribute to the reduction of the amount of time social media researchers have to invest for tagging purposes during their projects. To our knowledge, and particularly in Costa Rica, this would be the first tool available for that goal.

On the other hand, there are still some limitations to work around. First, Costa Rican news outlets on social media tend to focus on crime and judiciary issues, which could possibly create a bias for future automatic classification. As our database grows, more of these cases are included, therefore, the proposed topic modeling approach still needs to adjust the algorithm to efficiently avoid such an issue without losing accuracy.

Secondly, our gold standard approach derived from an expert consultation has several limitations. Without a clear quantitative parameter to classify the topics based on the semantic field they belong to, there exists room for error in the decision the expert took for classifying the topic. Furthermore, the distance between words that belong to a certain topic is very narrow for this article as they can be different parts-of-speech that, in the right context, could be part of any topic.

In the near future, it would be interesting to compare the approach with another unsupervised classification paradigm, for instance, convolutional neural networks. Given the amount of data available, it would be a feasible and interesting comparison. As a result, a mixed approach could be developed as a contribution to the field.

Finally, it is important to bear in mind that social media news consumption is a complex and evolving situation of interest. It unfolds in a specific context, therefore, special care needs to be taken before any generalization of findings. We have presented an effective topic modeling approach for the Costa Rican case, but other contexts might require particular adjustments that are out of our scope at the moment.

References

- [1] T. Highfield, "Social media and everyday politics", Cambridge: Polity Press, 2016.
- [2] H. Margetts, P. John, S. Hale, & T. Yasseri, "Political turbulence: How social media shape collective action", Princeton: Princeton University Press, 2016.
- [3] N. Newman, et al, "Reuters Institute Digital News Report 2019", Oxford: Reuters Institute, 2019.
- [4] A. Marwick, "Why Do People Share Fake News? A Sociotechnical Model of Media Effects", *Georgetown Law Technology Review*, 2(2), pp:474-512, 2018
- [5] S. Livingstone, "Tackling the Information Crisis: A Policy Framework for Media System Resilience", Foreword. In LSE 2018. p: 2. London: LSE, 2018.
- [6] S. Waisbord, "Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism Studies*", 19(13), pp:1866–1878, 2018.
- [7] D. Blei, "Topic Modeling and Digital Humanities". *Journal of Digital Humanities*, 2(1), pp: 8-11. 2021.
- [8] D. M. Blei, A. Y. Ng & M. I. Jordan, "Latent dirichlet allocation". *Journal of machine Learning research*, 3(Jan), pp: 993-1022, 2003.
- [9] R. Řehůřek, & P. Sojka, "Gensim—statistical semantics in python" Retrieved from genism.org, [Accessed May. 2, 2011]
- [10] C. Soto-Rojas, C. Gamboa-Venegas, A. Céspedes-Vindas, "MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media". *Tecnología en Marcha. Edición especial 2020*. 6th Latin America High Performance Computing Conference (CARLA). pp: 18-24. 2020.
- [11] CrowdTangle Team. CrowdTangle. Facebook, Menlo Park, California, United States. [List ID: 1510711], 2020.
- [12] C. Sievert and K. Shirley. "LDAvis: A method for visualizing and interpreting topics". In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp: 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics, 2014.
- [13] M. Röder, A. Both, and A. Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures". In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Association for Computing Machinery, New York, NY, USA, pp: 399–408. 2015.
- [14] A. Gliozzo, "Semantic domains and linguistic theory". In *Proceedings of the LREC 2006 workshop Toward Computational Models of Literary Analysis*, Genova, Italy. 2006.

Automatic image segmentation using Region-Based convolutional networks for Melanoma skin cancer detection

Segmentación automática de imágenes mediante redes convolucionales basadas en regiones para la detección del cáncer de piel tipo melanoma

Karen Dayana Tovar-Parra¹, Luis Alexander Calvo-Valverde², Ernesto Montero-Zeledón³, Mac Arturo Murillo-Fernández⁴, Jose Esteban Perez-Hidalgo⁵, Dionisio Alberto Gutiérrez-Fallas⁶

Tovar-Parra, K.D.; Calvo-Valverde, L.A.; Montero-Zeledón, E.; Murillo-Fernández, M.A.; Perez-Hidalgo, J.E.; Gutiérrez-Fallas, D.A. Automatic image segmentation using Region-Based convolutional networks for Melanoma skin cancer detection. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 14-25.

 <https://doi.org/10.18845/tm.v35i9.6479>

- 1 Universidad de Costa Rica. Costa Rica. E-mail: karen.tovar@ucr.ac.cr
 <https://orcid.org/0000-0003-1201-7195>
- 2 Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: lcalvo@tec.ac.cr
 <https://orcid.org/0000-0003-3802-9944>
- 3 Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: emontero@tec.ac.cr
 <https://orcid.org/0000-0002-4545-5805>
- 4 Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: mamurillo@tec.ac.cr
 <https://orcid.org/0000-0002-6913-3472>
- 5 Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: jose.perez@tec.ac.cr
 <https://orcid.org/0000-0001-7184-9974>
- 6 Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: dgutierrez@itcr.ac.cr
 <https://orcid.org/0000-0001-7190-8858>

Keywords

Melanoma; segmentation; machine learning; neural networks; medical image.

Abstract

Melanoma is one of the most aggressive skin cancers, however, its early detection can significantly increase probabilities to cure it. Unfortunately, it is one of the most difficult skin cancers to detect, its detection relies mainly on the dermatologist's expertise and experience with Melanoma. This research deals with targeting most of the common Melanoma stains or spots that could potentially evolve to Melanoma skin cancer. Region-based Convolutional Neural Networks were used as the model to detect and segment images of the skin area of interest. The neural network model is focused on providing instance segmentation rather than only a box-bounding object detection. The Mask R-CNN model was implemented to provide a solution for small trained datasets scenarios. Two pipelines were implemented, the first one was with only the Region-Based Convolutional Neural Network and the other one was a combined pipeline with a first stage using Mask R-CNN and then getting the result to use as feedback in a second stage implementing Grabcut, which is another segmentation method based on graphic cuts. Results demonstrated through Dice Similarity Coefficient and Jaccard Index that Mask R-CNN alone performed better in proper segmentation than Mask R-CNN + Grabcut model. In both models' results, variation was very small when the training dataset size changed between 160, 100, and 50 images. In both of the pipelines, the models were capable of running the segmentation correctly, which illustrates that focalization of the zone is possible with very small datasets and the potential use of automatic segmentation to assist in Melanoma detection.

Palabras clave

Melanoma; segmentación; aprendizaje automático; redes neuronales; imagen médica.

Resumen

El Melanoma es uno de los cánceres de piel más agresivos, sin embargo, su diagnóstico en una etapa temprana aumenta significativamente las opciones y el éxito en el tratamiento. Desafortunadamente, el Melanoma es uno de los cánceres de piel más difíciles de detectar, pues depende principalmente de la pericia y experiencia del dermatólogo. Esta investigación se enfoca en las manchas comunes que podrían evolucionar potencialmente a Melanoma. Se utilizaron redes neuronales convolucionales basadas en regiones como modelo para detectar y segmentar imágenes del área de la piel de interés. El modelo de red neuronal se centra en proporcionar segmentación de instancias en lugar de solo una detección de objetos delimitados por cajas. Se implementó el modelo Mask R-CNN con el propósito de proporcionar una solución para escenarios de pequeños conjuntos de datos entrenados. Inicialmente, solo se utilizó el modelo Mask R-CNN, luego se implementó Mask R-CNN y Grabcut: otro método de segmentación basado en cortes gráficos. Los resultados demostraron a través del coeficiente DSC y el índice de Jaccard que Mask R-CNN se desempeñó mejor en la segmentación que el modelo Mask R-CNN + Grabcut. En ambos modelos, la variación de los resultados fue muy pequeña cuando el tamaño del conjunto de datos de entrenamiento cambió entre 160, 100 y 50 imágenes. En ambas canalizaciones, los modelos fueron capaces de ejecutar la segmentación correctamente, lo que ilustra que la focalización de la zona es posible con conjuntos de datos muy pequeños y el uso potencial de la segmentación automática en la detección temprana de Melanoma.

Introduction

Melanoma is cancer in the skin that occurs when the cells that give color to the skin (melanocytes) begin growing in an uncontrolled manner. Melanoma can develop in almost any part of the body, and it can spread very fast to other body areas [1]. Melanoma cancer incidence has grown considerably in the last few years worldwide, this might be concerning since it is one of the most aggressive forms of skin cancer [2].

Like many other diseases, its early detection and removal can significantly make a difference in the patient, greatly increasing the survival probability. However, recent medical advances identified Melanoma as a complex and heterogeneous disease that has a variety of genotypes and phenotypes (the way it looks), which represents a challenge for detecting early Melanoma [2].

This skin cancer detection relies upon the physician's expertise and experience through a total body skin examination. During an examination is necessary a very bright light and different positions since this cancer have different features according to the anatomic site and the growth pattern such as symmetry, irregularity of borders, which varies by sex, race, and age. In this way, Melanoma detection depends mainly on qualitative observations [3].

This study aims to provide a complementary tool that supports the specialists whenever they are in the parameter recognition phase to detect Melanoma in the early stages. This research deals with targeting most of the common stains or spots that could potentially evolve to Melanoma skin cancer, more automatically by using Region-based Convolutional Neural Networks and segmentation methods based on graph cuts working directly on the picture of the stain to diagnose.

Region-based Convolutional Neural Networks:

To understand Region-based Convolutional Neural Networks, first, Convolutional Neural Networks (CNNs) should be defined. CNN's are neural networks fed forwardly, that have several layers, in each layer occurs a transformation and calculation to the outputs of the previous layer. The neurons that make up the neural network have biases and weights, however, in CNN's these are also organized in dimensions such as channels, width, height as well as the number of filters it has [4].

One of the most important characteristics of CNN is that its architecture expects an image as an input, this allows the network to have different layers such as the convolution layer, pooling layer, and fully connected layers and assign the weights to different parts of the image to differentiate one object from other [4] a diagram of a basic CNN is shown in the figure 1. Convolutional Neural Networks' functionality relies on working in feature extraction and feature classification in the same learning body, they can process large inputs and they are flexible enough to work with different input sizes [5]. The CNN architecture represents an advantage in efficiency over previous models [5].

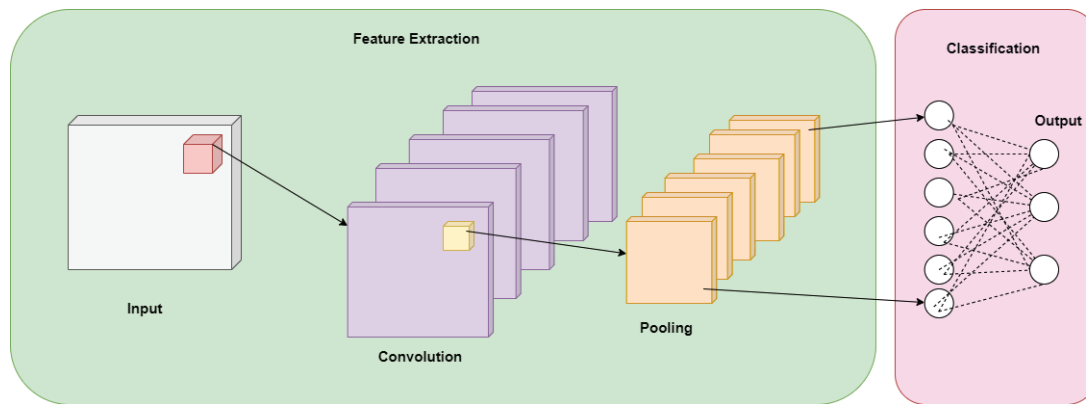


Figure 1. Basic Convolutional Neural Network (CNN) diagram.

CNN's are usually used to classify the image as a whole, nonetheless, if the interest is to detect specific objects in the image, Region-based Convolutional Neural Networks (R-CNN) are the ones designed for this purpose [6]. R-CNN uses convolutional neural networks and adds a method using selective search to divide the image into two thousand different regions that act as proposals. Contrary to image classification, object detection requires locations in the image to work on, in this path R-CNN operates in these regions to classify each one using linear Support Vector Machines until the objects of the training are detected and recognized [7].

Materials and methods (methodology)

For Melanoma detection and segmentation, a total of 200 images from different patients with malignant melanoma were used. The images were obtained from The International Skin Imaging Collaboration (ISIC) archives specifically from the HAM10000 [8]. Image dimensions are 600 x 450 pixels in jpg format. Melanoma stains are from both women and men, and they are in different parts of the body such as the posterior torso, upper extremity, lower extremity, head, and neck. The number of images chosen correlates with the objective of demonstrating that few images can properly train the models to have accurate results in test images. Desirable results with a small training dataset open the door to the use of images from medical centers in which there are hundreds but not thousands of Melanoma study cases.

Dataset was divided randomly into two parts: the training dataset and the validation dataset. The training dataset had 80% of the images, this dataset was used to train the model. The 20% left was assigned to the validation dataset to make the tests and see how the model behaved.

For the training dataset, it was necessary to manually identify and label the images manually to have a ground truth, the manual segmentation and labeling were performed image by image using the Make Sense tool [9]. Labeling and segmentation of images had the supervision of field experts, images with the melanoma stain were masked by the clinician's contour and were labeled as "Melanoma".

In this investigation, two models were used: Mask R-CNN which uses Region-based Convolutional Neural Networks, and Grabcut which is a method that uses graph cuts. In the first part of the study, only the mask region-based convolutional neural network model was used.

For the first model, COCO weights [10] were used in h5 format. COCO stands for “Common Objects in Context”, it is a dataset that is used for object detection, segmentation, and captioning datasets. This dataset has weights that are the parameters used in the layers of the neural network, such as filter weights and biases. COCO has more than 330K images and more than one million object instances [10].

During the process, all images were dimensioned to 512 x 512. The model was set up with 2 classes: “Melanoma” and “Background”, the configuration also included 500 steps per epoch in the training, 5 validation steps during training, and the backbone used was Residual Network: ResNet50. The backbone is the network selected to extract the features from the model’s input. The output of the backbone is the input for the rest of the convolutional networks. The training dataset was transformed before the training into a Coco dataset alike since this is the one that the original neural network uses. The first output is the image with the color mask over the stain, however, to obtain the stain segmented the mask is applied to the original image using computer vision as can be observed in figure 2.

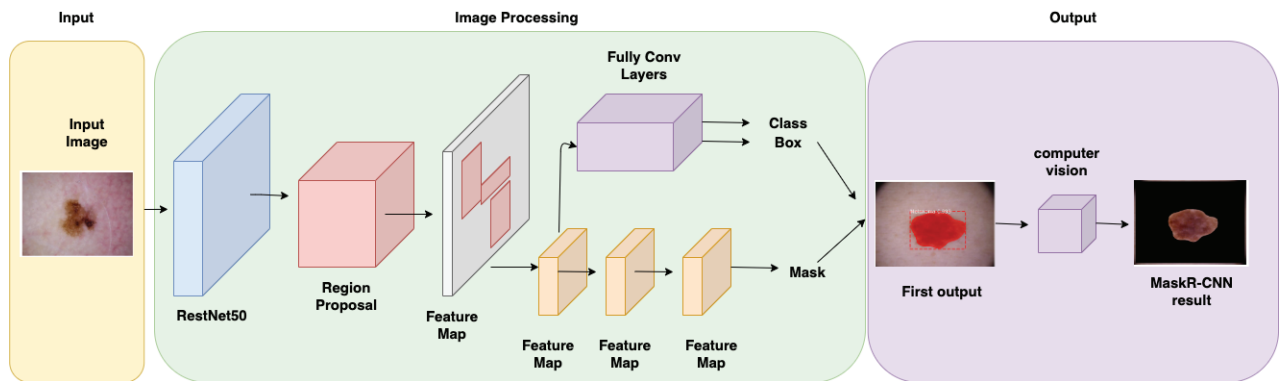


Figure 2. Mask R-CNN general architecture used in our research. RestNet50 was chosen as the backbone network and computer vision image processing is used in the output.

In the second phase of the investigation, Grabcut was introduced, as shown in figure 3. The algorithm was added to the pipeline to work with the outputs from the first model Mask R-CNN to make a foreground extraction and see if this could improve the pipeline results. Output from the first model was the segmented image with a black background, using Python’s computing vision library CV2, the mask was separated from the image leaving a mask with a black background and a white zone corresponding to the Melanoma shape. This was used as input to Grabcut which segmented a second time the image and generated a new mask using color similarity mapping.

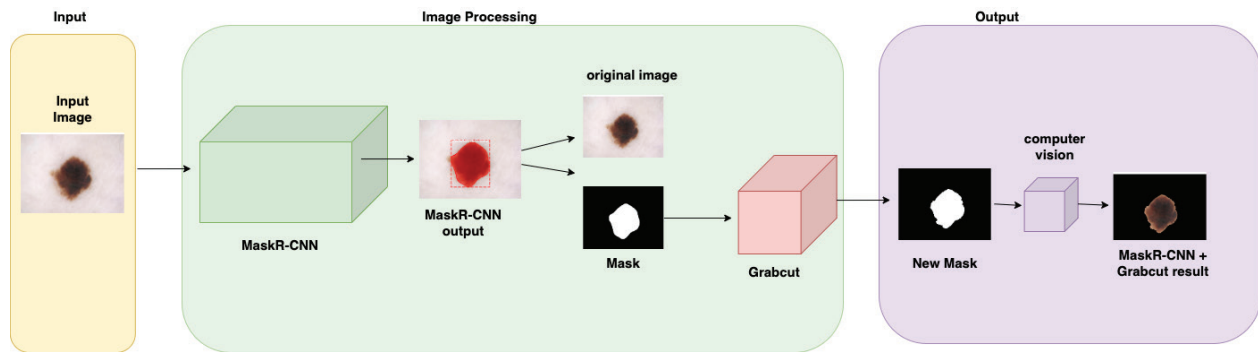


Figure 3. Pipeline architecture implementing Grabcut receiving Mask R-CNN mask output as an input.

Grabcut is a segmentation algorithm based on making graph cuts using the Gaussian mixture model to detect the foreground and the background on an image. The Gaussian Mixture Model translates the image to a graph that has both a source point and a sink point; these points are useful to recognize the pixel and the correlation between the pixels and the background and foreground [11]. Figure 4 shows the graph diagram.

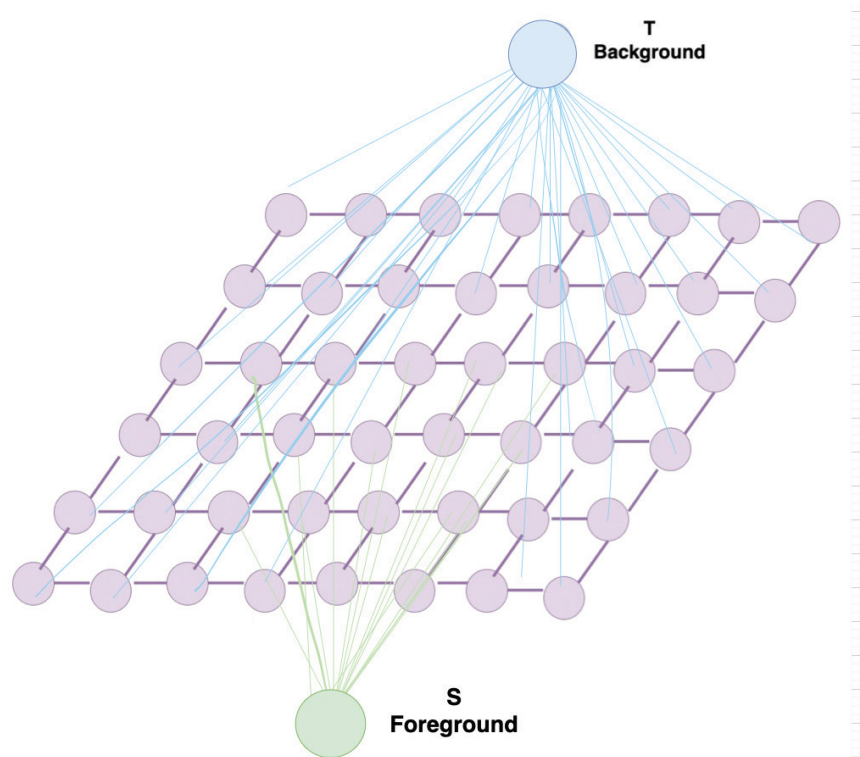


Figure 4. Graph diagram that Grabcut algorithm designs using the Gaussian Mixture Model.

The graph can be then converted to a function that receives a vertex, if the function has a result of 1, then the vertex is the source point, on the contrary, if the function has a result of 0, it means the vertex is a sink point [12]. Grabcut algorithm requires the user to provide as an input a selection of the area of interest in the image, the region outside the selected area is mapped as the background, whereas the region of the selected area is mapped as the unknown area, which

later is classified in two sub-regions: background in the area of interest and the intended region. These two subregions are divided according to the maximum and minimum of the Gaussian Mixture Model, and in parallel, the parameters for the model are updated. The graph that was initially created is divided when the algorithm converges with the right parameters [11].

For both pipelines, three different training sessions took place by varying the quantity of the images in the training dataset. The three different stages were carried on with 50, 100, and 160 images, to see how affected the training dataset to the models and their behavior. This as mentioned, works in the line with the objective to demonstrate good performance with few training images. Having different groups varying the training dataset size allows this investigation to measure the impact of decreasing training images and the acceptable minimum limit of the training scope to make both models work successfully by making the automatic segmentation a support tool for medical professionals. This, for example, would also give an idea of the image quantity needed for training another melanoma type with this model.

Evaluation Metrics

The metrics used to evaluate both implementations were the Dice similarity coefficient (DSC) and the Jaccard Index. Both metrics work on overlapping to make a comparison between the resulting image with the ground truth. Equation (1) defines DSC:

Where:

- X: The ground truth. This is the image segmented manually in the zone the segmentation is wanted.
- Y: The result obtained. This is the image with the automatic segmentation performed by the model.

$$\text{DSC} = \frac{2 * |X \cap Y|}{|X| + |Y|} \quad (1)$$

On the other hand, equation (2) defines the Jaccard Index:

$$\text{Jaccard Index} = \frac{|X \cap Y|}{|X| + |Y| * |X \cap Y|} \quad (2)$$

Both metrics are in a range between 0 and 1, the closest to 1 the similar to the ground truth.

Results

Results obtained using Mask R-CNN alone and Mask R-CNN + Grabcut are shown in Table 1 and Table 2. Table 1 shows the middle value of both the Dice similarity coefficient and the Jaccard Index for the 40 images that were used to test both models, in Table 2 other measurements can be observed with both models and both evaluation metrics.

Overall, results show that Mask R-CNN alone has a DSC and a Jaccard Index closest to 1, and the standard deviation is highest in the model that implements Grabcut. The highest metric is the Dice Similarity Coefficient in Mask R-CNN with 0.923.

On the other hand, Table 3 presents the evaluation metrics obtained using different sizes in the training dataset in both models Mask R-CNN and Mask R-CNN + Grabcut. The dataset sizes are 50, 100, and 160 images. For Mask R-CNN the greatest difference between the different training datasets is 0.01 points in DSC and 0.018 in Jaccard Index. As well, for Mask R-CNN + Grabcut the biggest difference is 0.012 in the DSC and 0.014 points in Jaccard Index.

Table 1. General results of segmentation of Melanoma lesions.

Model	Dice Similarity Coefficient	Jaccard Index
Mask R-CNN	0.923	0.857
Mask R-CNN + Grubcut	0.729	0.573

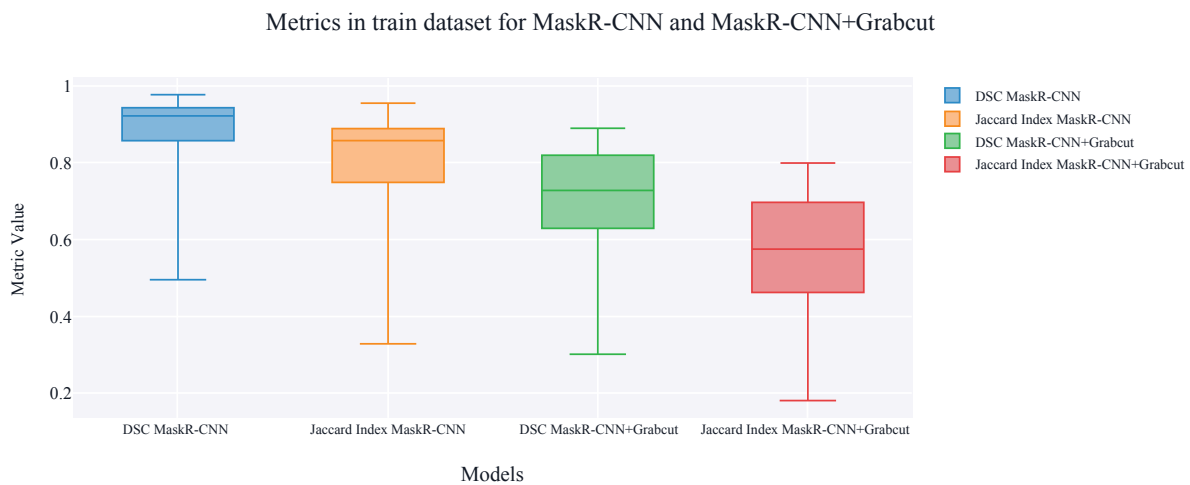


Figure 5. Results of Melanoma Segmentation for Mask R-CNN and MasR-CNN+Grabcut

Table 2. Detailed results of segmentation of Melanoma lesions including Mean, Median, and the Standard Deviation in the implementation of Mask R-CNN and Mask R-CNN + Grabcut

Model	Dice Similarity Coefficient		
	Mean	Median	Standard Deviation
Mask R-CNN	0.876	0.923	0.114
Mask R-CNN + Grubcut	0.698	0.729	0.154
Model	Jaccard Index		
	Mean	Median	Standard Deviation
Mask R-CNN	0.795	0.857	0.153
Mask R-CNN + Grubcut	0.554	0.573	0.163

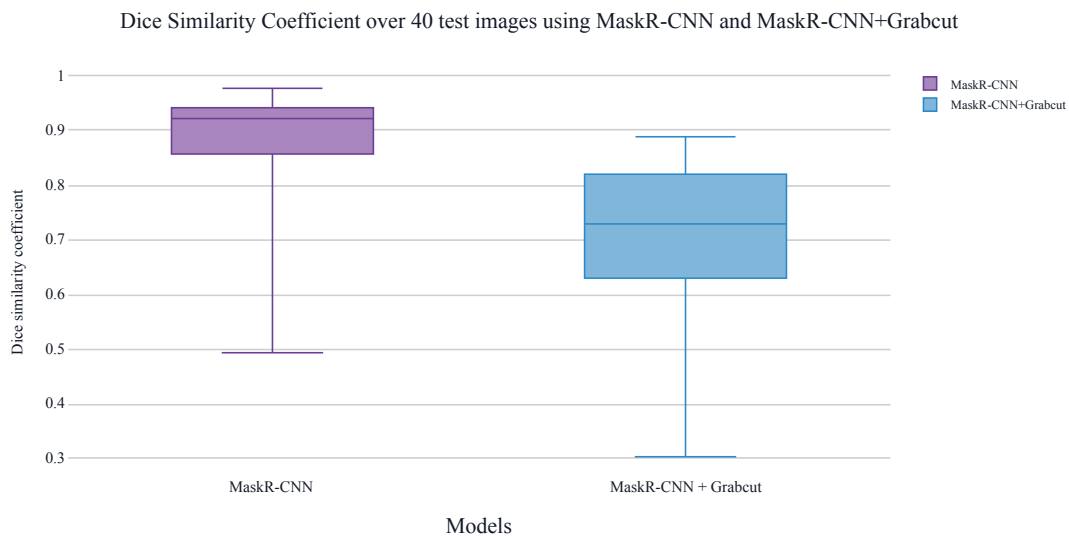


Figure 6. Dice Similarity Coefficient results from a comparison between Mask R-CNN and Mask R-CNN + Grabcut.

Table 3. General results of Melanoma segmentation using different-sized training datasets.

Model	Dice Similarity Coefficient	Jaccard Index
Mask R-CNN 50 images	0.913	0.839
Mask R-CNN 100 images	0.922	0.856
Mask R-CNN 160 images	0.923	0.857
Mask R-CNN + Grabcut 50 images	0.717	0.559
Mask R-CNN + Grabcut 100 images	0.718	0.560
Mask R-CNN + Grabcut 160 images	0.729	0.573

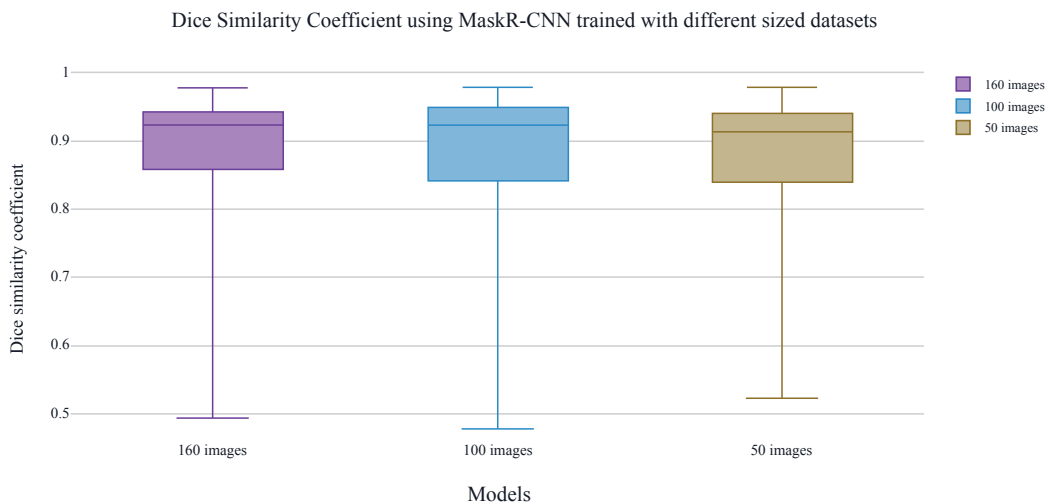


Figure 7. Results DSC metric in the segmentation using Mask R-CNN alone with different sizes in the training datasets (160, 100, 50 images).

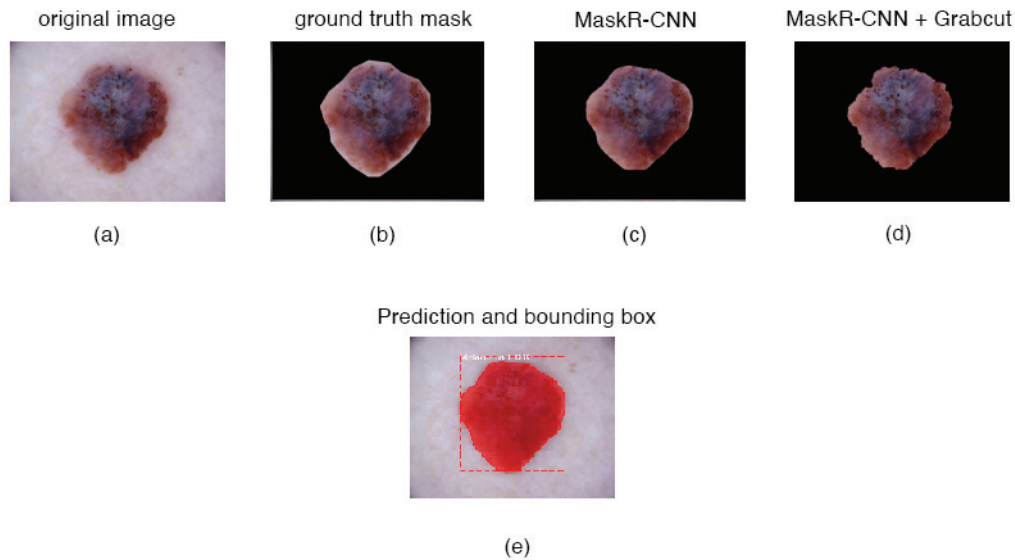


Figure 8. Melanoma segmentation results from one patient. (a) The original image; (b) ground truth segmentation; (c) mask region-based convolutional neural network segmentation; MaskR+CNN using Grabcut is shown in (d); Mask R-CNN prediction and bound boxing are shown in (e). For this Melanoma, DSC was 0.953 in Mask R-CNN and 0.821 with Mask R-CNN + Grabcut.

Discussion and conclusions

Results have showed that adding another layer to the pipeline using the Grabcut algorithm does not give better results, since as is observed in figure 4, DSC and Jaccard index metrics are higher for the pipeline that implements only Mask R-CNN, considering as well as presented in figure 7, the extra layer tends to take away the stain's border in the segmentation, which is one of the ABCDs of melanoma used to properly detect a Melanoma cancer stain [2]. Mask R-CNN using ResNet50 was able to present a smoother segmentation, maintaining the borders and closest to the ground truth. Although in other experiments such as [11] Grabcut is used as an interactive method that could improve the segmentation, in this experiment the results showed that images segmented by a second layer using Grabcut algorithm have less border resolution, and for some cases, the border was eliminated by the segmentation, it also segmented the image with a more irregular frontier, which leads to a bigger non-common area with the ground truth mask of the images, that affected negatively the metrics. Less background in other case studies might be beneficial however it might not always be desired as the stain shape and border color are important in Melanoma detection.

In Figure 3 can be also observed how the results for DSC are dispersed in a bigger range in the case of the implementation with Mask R-CNN + Grabcut rather than the implementation only using Mask R-CNN, which points to less consistency when adding Grabcut.

Furthermore, with different training dataset sizes (50, 100, and 160 mages) the difference is not relevant as observed in figure 6, where the changes are over ± 0.015 points in the evaluation metrics, this states that Mask R-CNN with ResNet50 as the backbone can work properly even with few images in training such as 50 images and the segmentation keeps a high Dice Similarity Coefficient and Jaccard Index.

Both models were able to avoid the wrong detection in most of the cases by ignoring hair, small moles, and other pigmentation not related to the Melanoma, herewith it demonstrated the potential use of the automatic segmentation to assist the clinician in the Melanoma detection and delineation to extract it in pertinent cases.

This research results open up possibilities for future work such as using the segmentation to predict and classify different types of Melanoma and skin cancer based on shape and color structure. In addition, more work could be performed on looking for the right place in the pipeline to use Grabcut algorithm, including the creation of a new layer inside Mask R-CNN network that uses Grabcut instead of placing it at the end of the neural network results, so that the network can feed the Grabcut algorithm in the way it feeds Mask R-CNN results.

As well, an interesting approach would be to design the experiment to replace Grabcut with a second neural network especially another region-based convolutional network based on a U-Net architecture, making the output of the first RCNN the input of the second RCNN. Improving the results obtained in the experiment could also be approached from the preprocessing stage by applying contrast limited adaptive histogram equalization, applying the Single Image Haze Removal Using Dark Channel, and applying pseudo labeling to the test data. On the other hand for the loss function other metrics further than DSC and JSC could be used, such as the Weighted Boundary Loss which looks to reduce the distance between ground truth and the result segmentation, Lovasz which performs an optimization of the mean intersection over union and Center loss which penalizes distances between deep features and the class centers.

Acknowledgements

The authors would like to thank the Vicerrectoría de Investigación y Extensión at Instituto Tecnológico de Costa Rica for providing the occasion for this research.

References

- [1] T. C. Mitchell, G. Karakousis and L. Schuchter. "Melanoma" in *Abeloff's Clinical Oncology*, 6th ed. J. E. Niederhuber, J. O. Armitage, J. H. Doroshow, M. B. Kastan MD and J. E. Tepper, Eds. Elsevier, 2020, ch. 66, sec. 3, pp. 1034-1051
- [2] C. Garb et al., *Cutaneous Melanoma*. Springer, 2020, doi: 10.1007/978-3-030-05070-2
- [3] H. L. Kaufman y J. M. Mehnert, Eds. *Melanoma*. Cham: Springer International Publishing, 2016, doi: <https://doi.org/10.1007/978-3-319-22539-5>
- [4] J. Teuwen, N. Moriakov, "Convolutional neural networks" in *Handbook of Medical Image Computing and Computer-Assisted Intervention*. S. Zhou, D. Rueckert, G. Fichtinger, Eds. Academic Press, 2020, ch. 20, pp. 481-501, doi: <https://doi.org/10.1016/C2017-0-04608-6>
- [5] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj y D. J. Inman, "1D convolutional neural networks and applications: A survey", *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021, doi: <https://doi.org/10.1016/j.applthermaleng.2014.05.008>
- [6] L. Chenning, Y. Ting, Z. Qian, and X. Haowei, "Object-based Indoor Localization using Region-based Convolutional Neural Networks," *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2018, pp. 1-6, doi: <https://doi.org/10.1109/ICSPCC.2018.8567795>.
- [7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, doi: 10.1109/CVPR.2014.81
- [8] Tschandl, P. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", vol 3. *Harvard Dataverse*, 2018, doi: <https://doi.org/10.7910/DVN/DBW86T> (2018)
- [9] "Make Sense". Make Sense. Available: <https://www.makesense.ai/>

- [10] T. Lin *et al.* "Microsoft COCO Common objects in Contexts". D. Fleet, T. Pajdla, B. Schiele B., T. Tuytelaars, Eds. Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham, 201, doi: https://doi.org/10.1007/978-3-319-10602-1_48
- [11] G. Yao, S. Wu, H. Yang, S. Li. "GrabCut Image Segmentation Based on Local Sampling." in *Business Intelligence and Information Technology. BIIT 2021. Lecture Notes on Data Engineering and Communications Technologies*, Vol 107. A. Hassanien, Y. Xu, Z. Zhao, S. Mohammed, Z. Fan, Eds. Springer, Cham, 2022, ch. 5, pp. 356-365, doi: https://doi.org/10.1007/978-3-030-92632-8_34
- [12] D. Ren, Z. Jia, J. Yang and N. K. Kasabov, "A Practical GrabCut Color Image Segmentation Based on Bayes Classification and Simple Linear Iterative Clustering," in *IEEE Access*, vol. 5, pp. 18480-18487, 2017, doi: [10.1109/ACCESS.2017.2752221](https://doi.org/10.1109/ACCESS.2017.2752221).

Benchmarking the NXP i.MX8M+ neural processing unit: smart parking case study


Evaluando la unidad de procesamiento neuronal NXP i.MX8M+: el caso de estudio de parqueos inteligentes

Edgar Chaves-González¹, Luis G. León-Vega²


Chaves-González, E.; León-Vega, L.G. Benchmarking the NXP i.MX8M+ neural processing unit: smart parking case study. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 26-30.

 <https://doi.org/10.18845/tm.v35i9.6487>

1 Embedded Software Engineer & Graduate Student Intern. RidgeRun LLC, Costa Rica. E-mail: edgar.chaves@ridgerun.com

 <https://orcid.org/0000-0002-0269-526X>

2 Senior Embedded Software Engineer. RidgeRun LLC, Costa Rica. Master's in Electronics Student. Instituto Tecnológico de Costa Rica. Costa Rica. PhD fellow. Università degli Studi di Trieste. Italy. E-mail: luis.leon@ridgerun.com

 <https://orcid.org/0000-0002-3263-7853>

Keywords

Computer vision; AI accelerator; embedded software; smart cameras; convolutional neural networks; TinyYOLO; Rosetta.

Abstract

Nowadays, deep learning has become one of the most popular solutions for computer vision, and it has also included the Edge. It has influenced the System-on-Chip (SoC) vendors to integrate accelerators for inference tasks into their SoCs, including NVIDIA, NXP, and Texas Instruments embedded systems. This work explores the performance of the NXP i.MX8M Plus Neural Processing Unit (NPU) as one of the solutions for inference tasks. For measuring the performance, we propose an experiment that uses a GStreamer pipeline for inferring license plates, which is composed of two stages: license plate detection and character inference. The benchmark takes execution time and CPU usage samples within the metrics when running the inference serially and parallel. The results show that the key benefit of using the NPU is the CPU freeing for other tasks. After offloading the license plate detection to NPU, we lowered the overall CPU consumption by 10x. The performance obtained has an inference rate of 1 Hz, limited by the character inference.

Palabras clave

Visión por computador; Acelerador de AI; software empotrado; cámaras inteligentes; convolutional neural networks; TinyYOLO; Rosetta.

Resumen

Actualmente, el aprendizaje profundo se ha convertido en una de las soluciones más populares para la visión por computador y ha incluido también el *Edge*. Esto ha influenciado a los productores de *System-on-Chip* (SoC) a integrar aceleradores para tareas de inferencia en sus SoC, incluyendo a NVIDIA, NXP y Texas Instruments. En este trabajo se explora el rendimiento de la unidad de procesamiento neuronal (NPU) de la NXP i.MX8M Plus como una de las soluciones de tareas de inferencia. Para las mediciones de desempeño, proponemos un experimento basado en un *pipeline* de GStreamer para la inferencia de placas de vehículos, el cual consta de dos etapas: la detección de placas y la inferencia de sus caracteres. Para ello, este *benchmark* toma muestras de tiempos de ejecución y uso de CPU dentro de sus muestras cuando se corre la ejecución de la inferencia de manera serial y paralela. Los resultados obtenidos demuestran que el beneficio en el uso del NPU radica en la liberación del CPU para otras tareas. Después de descargar la detección de placas al NPU, reducimos el consumo total de CPU 10 veces. El desempeño obtenido tiene una tasa de inferencia de 1 Hz, limitado por la inferencia de caracteres.

Introduction

Recently, there have been efforts for accelerating machine learning (ML) inference at the Edge. In 2020, NXP launched the i.MX8M Plus, an ARM-based SoC equipped with a Neural Processing Unit. It provides up to 2.3 TOP/s, supports 8-bit and 16-bit fixed-point arithmetic and implements the Winograd algorithm [1] for convolutions.

NVIDIA, another SoC vendor, proposes other similar solutions in the market. The NVIDIA Jetson Nano is the closest board in memory, CPU, and peripherals to the i.MX8M Plus. It is equipped with an NVIDIA Maxwell GPU that facilitates and accelerates algorithms using CUDA [2] and

delivers up to 472 GOP/s (4.87x less than the theoretical of i.MX8M Plus) [3]. According to our previous measurements, we can run a TinyYoloV2, a deep neural network architecture for object detection on images, on a Jetson Nano, delivering up to 15 frames per second (fps) [4].

The application fields of these embedded systems are vast, going from home control and security to artificial intelligence, automated driving, healthcare, and others. They are preferred over cloud solutions because of their reliability in inference latency and power consumption, which does not reach more than 20W [1].

This work explores the performance benefit of offloading deep learning (DL) models to the NPU included in the i.MX8M Plus [1]. For this study, the system runs a license plate inference (LPI) pipeline implemented in the RidgeRun GstInference framework [5], utilizing TensorFlow Lite as DL backend and NNAPI as an accelerator-backend interface [6]. The key contribution of this work is a case study of the usage of the i.MX8M Plus in real-world applications, in particular, smart parking.

Benchmark architecture

Our work is based on a smart parking use case. It is an industry-coming sample to evaluate boards with an actual application environment. It consists of a camera recording the license plates of the cars coming into the parking lot. The camera is connected directly to the board, where an LPI task is executed on the captured images, and the results are uploaded to the Internet for data logging.

Our LPI architecture is two-fold. The first stage performs license plate detection (LPD) to recognise the license plate as an object within the image. It is based on an in-house trained TinyYOLOV3 model architecture [3]. The model delivers fixed bounding boxes quantifying their confidence and classifying it into several classes. The result is transferred to the license plate recognition (LPR) stage, based on Rosetta [7], to get the inference in a character string format.

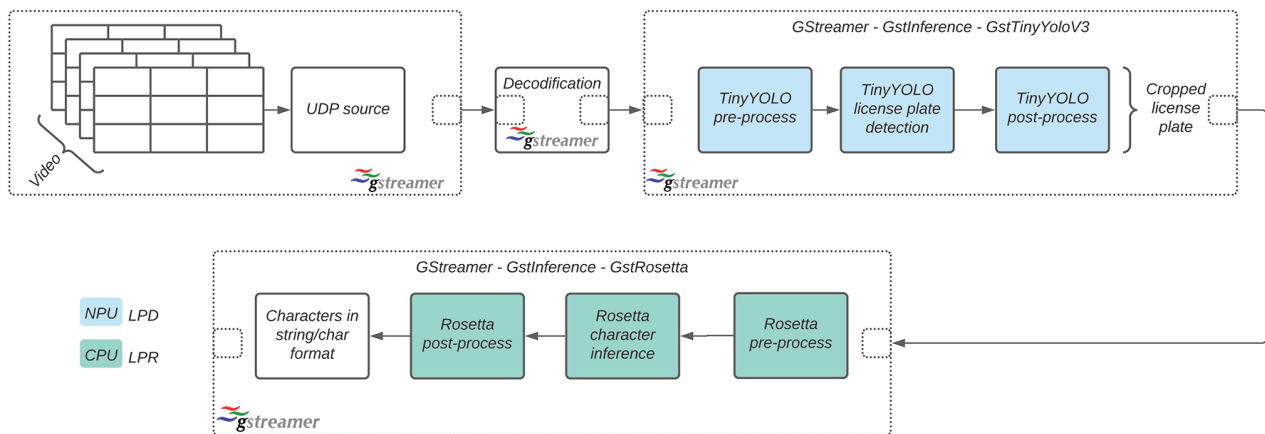


Figure 1. Dataflow of the LPD and LPR algorithms within a GStreamer pipeline based on GstInference

The inference data flow is implemented in a GStreamer pipeline, as presented in Figure 1. We have chosen GStreamer because of its support, portability and performance [8]. To add GStreamer support to the models, we wrapped TinyYOLOv3 and Rosetta into two GStreamer elements for isolating the LPD from LPR algorithms. Thus, we manage to place each algorithm in a separate thread using a RidgeRun’s solution called GstInference, which makes our solution

portable amongst several inference backends and devices (CPU, GPU, and TPU). According to Figure 1, our pipeline starts with a UDP source to receive the data from a network camera; then, the image is decoded and transformed into a suitable format for our GStreamer wrappers. Within our GStreamer elements (GstTinyYoloV3 and GstRosetta), we preprocess the data, perform the inference and post-process to place the results as metadata circulating within the pipeline.

Regarding the measurements, the benchmark takes samples of the CPU usage using the Linux PS at 2 Hz for 100 iterations to measure the results. Our setup is capable of running in two modes: i) individually, executing the two models isolated, and ii) in parallel, running both inference tasks concurrently. The time is estimated using the high-precision clock from the C++ Standard Template Library (STL).

Results

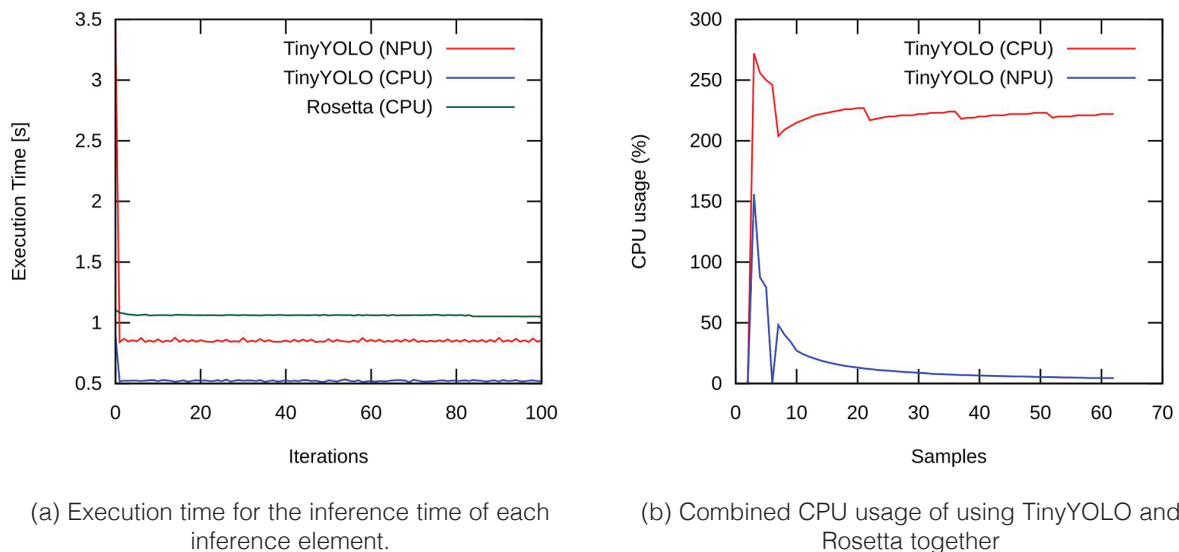


Figure 2. Inference performance metrics. The NPU offloading benefits in freeing 10x the CPU usage. Rosetta was running on the CPU all the time.

After running our benchmark, we obtained the results presented in Figure 2. Figure 2.a shows the execution time of each algorithm running isolated (one at a time). Figure 2.b presents the results when running the whole system together using two configurations: i) TinyYOLOv3 on CPU, and ii) TinyYOLOv3 on NPU. In the first iterations, the inference is slow in the beginning because of a warmup process (context and resource allocation).

TinyYOLOv3 can run on both the NPU and the CPU regarding the individual runs. With NPU, the inference rate was 1.18 Hz, whereas the CPU was 1.69x faster, with 1.99 Hz. Rosetta could not be accelerated because the model does not support quantisation. It managed to run at approximately 1 Hz.

Figure 2.b presents the CPU consumption of running the LPD in the NPU and the CPU while executing the LPR parallel on the CPU. Running both algorithms on the CPU leads to consumption above 200%, which means that more than two cores are fully occupied during the inference tasks. On the other hand, when running the LPD on NPU and the LPR on the CPU,

the CPU consumption drops below 20% (10x gained), freeing the CPU for other tasks. In both cases, the inference rate is 1 Hz. Hence, using the NPU does not benefit the inference rate, and it contributes to lowering the CPU utilization. According to our results for TinyYOLOv2 in the Jetson Nano, we expected to have around 15 Hz but the actual rate achieved was 2 Hz. It also highlights the resource underutilisation of the Rosetta model.

Conclusions

In this work, we have exposed a benchmark using a smart parking application for evaluating the NXP i.MX8M Plus for deep learning inference tasks, particularly for license plate detection and recognition. For this case study, we offloaded the LPD (TinyYOLOv3) to the NPU, freeing 10x of the CPU resources for other tasks, performing similarly and without impacting the final inference rate. We managed to get TensorFlow running at 1.99 Hz maximum and Rosetta at 1 Hz, posing the bottleneck of our application. In future work, we plan to extend this application to other platforms with similar power consumption and capabilities, such as NVIDIA Jetson Nano and Xilinx ZYNQ 7000.

Acknowledgements

This project was possible thanks to RidgeRun LLC and their Master's scholarship and internship programmes.

References

- [1] "i.MX8M Plus", Nxp.com, 2021. [Online]. Available: <https://www.nxp.com/products/processors-and-microcontrollers/arm-processors/i-mx-applications-processors/i-mx-8-processors/i-mx-8m-plus-arm-cortex-a53-machine-learning-vision-multimedia-and-industrial-iot:IMX8MPLUS>. [Accessed: 22- Sep- 2021].
- [2] NVIDIA. 2014. DATA SHEET NVIDIA Jetson Nano System-on-Module. [Online]. Available: <https://developer.nvidia.com/embedded/jetson-nano> [Accessed: 22- Mar- 2022]
- [3] H. Bouzidi, H. Ouarnoughi, S. Niar, and S. Ait El Cadi. 2022. Performances Modeling of Computer Vision-based CNN on Edge GPUs. ACM Trans. Embed. Comput. Syst. DOI: <https://doi.org/10.1145/3527169>
- [4] "GstInference Benchmarks", RidgeRun Embedded Solutions LLC. [Online]. Available: <https://developer.ridgerun.com/wiki/index.php?title=GstInference/Benchmarks>. [Accessed: 14- Sep- 2021].
- [5] "Introduction to GstInference", developer.ridgerun.com. [Online]. Available: <https://developer.ridgerun.com/wiki/index.php?title=GstInference/Introduction>. [Accessed: 20- Oct- 2021].
- [6] "Neural Networks API | Android NDK | Android Developers", Android Developers. [Online]. Available: <https://developer.android.com/ndk/guides/neuralnetworks>. [Accessed: 09- Sep- 2021].
- [7] V. Sivakumar, A. Gordo and M. Paluri, "Rosetta: Understanding text in images and videos with machine learning", Facebook Engineering, 2021. [Online]. Available: <https://engineering.fb.com/2018/09/11/ai-research/rosetta-understanding-text-in-images-and-videos-with-machine-learning/>. [Accessed: 20- Sep- 2021].
- [8] "GStreamer: open source multimedia framework", GStreamer.freedesktop.org. [Online]. Available: <https://gstreamer.freedesktop.org/>. [Accessed: 14- Sep- 2021].

Analysis of a cooling system for robotic joints using a computational fluid dynamics study

Análisis de un sistema de enfriamiento para articulación robótica mediante un estudio de dinámica de fluidos computacional

María Fernanda Madriz-Ramírez¹, Kevin Alberto Solano-Núñez², Mauricio Rodríguez-Calvo³

Madriz-Ramírez, M.F.; Solano-Núñez, K.A.; Rodríguez-Calvo, M. Analysis of a cooling system for robotic joints using a computational fluid dynamics study. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 31-38.

 <https://doi.org/10.18845/tm.v35i9.6488>

- 1 Universidad Técnica Nacional. Costa Rica. E-Mail: mamadriz@est.utn.ac.cr
 <https://orcid.org/0000-0001-9922-9398>
- 2 Universidad Técnica Nacional. Costa Rica. E-Mail: knsolano@est.utn.ac.cr
 <https://orcid.org/0000-0002-2805-1091>
- 3 Universidad Técnica Nacional. Costa Rica. E-Mail: mrodriguez@utn.ac.cr
 <https://orcid.org/0000-0002-3737-2866>

Keywords

Heat exchanger; cooling system; humanoid robot; fluid dynamic simulation.

Abstract

The cooling system is one of the most important parts of a robot, a correct design allows it to increase its capacity while reducing the chances of a malfunction. In general, cooling systems for humanoid robots require special characteristics to not interfere with the optimal operation; aspects such as weight, size, materials, and positioning of this system are crucial for maximum efficiency. The current problem in this design is that the cooling system needs to maintain its power to perform tasks when lifting heavy weights without consequences such as reduced mobility of the joint, and therefore decrease the robot's efficiency. In this paper, we studied, analyzed, and discarded existing options of electric motor actuators. Once the viable options were obtained, so we proceed to make a proper design, then was made a computational fluid dynamics study (CFD) to get heat extraction measurements as well as fluid velocity, essential to make the final decision, solving the initial problem.

Palabras clave

Intercambiador de calor; sistema de refrigeración; robot humanoide; termodinámica; simulación dinámica de fluidos.

Resumen

El sistema de refrigeración es una de las partes más importantes de un robot, un correcto diseño le permite aumentar su capacidad a la vez que reduce las posibilidades de un mal funcionamiento. En general, los sistemas de refrigeración para robots humanoides requieren de características especiales para no interferir en su óptimo funcionamiento; aspectos como el peso, el tamaño, los materiales y el posicionamiento de este sistema son cruciales para lograr la máxima eficiencia. El problema actual en este diseño es que el sistema de enfriamiento necesita mantener su potencia para realizar tareas al levantar pesos pesados sin consecuencias como la reducción de la movilidad de la articulación, y por lo tanto disminuir la eficiencia del robot. En este trabajo estudiamos, analizamos y descartamos las opciones existentes de actuadores de motores eléctricos. Una vez que se obtuvieron las opciones viables, por lo que se procede a hacer un diseño adecuado, luego se realizó un estudio de dinámica de fluidos computacional (CFD, por sus siglas en inglés) para obtener las medidas de extracción de calor y la velocidad del fluido, esenciales para tomar la decisión final, resolviendo el problema inicial.

Introduction

Currently, at the Universidad Técnica Nacional, the program (Design of a proprioceptive robotic arm for robotic limbs with practical applications in humanoid robots) is under development in the degree course in Electronic Engineering, which consists of the creation of a robotic arm with qualities of handling heavy objects in everyday life, however, it has encountered some difficulty regarding an optimal solution for its cooling system. Therefore, the contribution of engineering in industrial production in this research focuses on the need to investigate, study, and analyze improvements about liquid cooling systems in motors for robotic joints, which adapt to the specifications of the joint of the humanoid robot.

For the optimal operating of any robot, a set of various important factors are required, one of the most important is the cooling system. The main objective of the cooling system is to prevent the joints from overheating, especially when working under high-demand tasks. When an adequate temperature is maintained, higher efficiency is obtained, and therefore a lower exposure to a possible failure is achieved [9]. To enable robots as effective assistants they need a certain power to perform tasks that require lifting a lot of weight generating a high energy demand, current solutions have been based on adding very large mechanical reductions, this is not a viable solution, since it adds more weight and decreases mobility [14][15].

Humanoid robots pretend to simulate a human body and its functions, many of them use electric motors to generate this joints movement, some problems related to robots, especially when we talk about dynamic ones, are the overheating of their joints, which is why, at the University of Tokyo, Japan, they built a humanoid robot called KENGO, which has a mechanism that simulates sweating to control the temperature of its electric motors [1][11][13].

The humanoid Kengoro robot is powered by 108 motors, but what is truly amazing about this robot is that it can sweat, thanks to its innovative cooling system. Like all human beings, robots generate heat when they perform their tasks, so the mechanism used by the robot to simulate artificial perspiration is in question a cup of deionized water, a type of purified water that lacks salts and minerals. dissolved. Water bubbles through the porous framework and is replaced every 12 hours or so. What they do is take advantage of the metal structure of the robot, which normally uses it for support but this time it is used to fill the metal structure with water and kept it cold, so the clever layers of the perforated structures allow the water would run through channels inside the skeleton, pass to an outer level and evaporate, all without a wet robot dripping onto the floor [9].

Stanford scientists and researchers at Stanford University are also exploring new methods to keep robots cool with tiny holes, they showed a new type of breathable plastic fabric with nanometer-sized pores that could absorb body radiation [8][12].

In short, a cooling system is a mechanism based on the principles of thermodynamics and fluid mechanics, they are designed to transfer thermal energy between two sources. These are also designed as a thermodynamically closed system, they prevent the motor from overheating; first, the liquid allows to absorb the heat generated in the motor, then this heat is transferred into the air. These are some of the reasons why water is normally used because first, it has a great heat transfer coefficient without increasing the system temperature, in addition, it has a lower pollution impact, on top of that's it is inexpensive [3].

Methodology

The following section will show the 3D design of the heat exchanger, a vital part of the cooling system, variables of weight, mass, and volume are considered, in addition to the other parts that are assembled such as the robotics joints, copper pipes, cables, the pump, and mechanical support. In this part, the design methodology that is applied, efficiency is conformed of six steps [2].

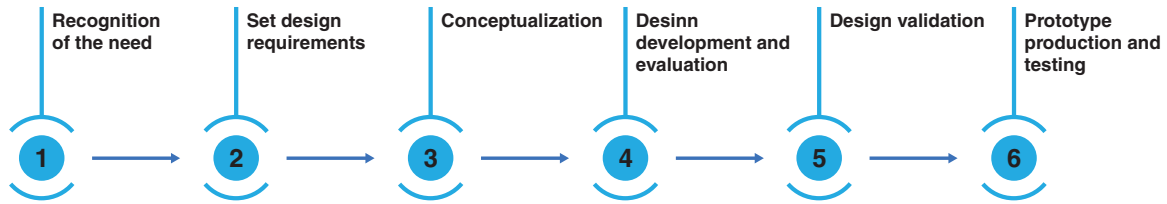


Figure 1. Methodology

As it was mentioned before, the central part of a cooling system is the heat exchanger, which acts as a regulator of temperature on the liquid used, additionally the pump is the center of the cooling system, it moves the liquid through all the internal conductors, as if it were an inner cardiovascular system. For this part, the calculations to determine the dimensions of the radiator and its design was made with CAD software, all the process is shown hereunder. To obtain the heat exchanger dimension was necessary to calculate the calorific power, which is represented with $\dot{Q}_{dissipate}$, this data shows how much heat must be dissipated by the cooling system. Using the $P_{max} = 880.88$ W, the engine maximum's power and the motor efficiency have a value of 0.91. The nomenclature used on these formulas will be shown above [2].

$$\dot{Q}_{dissipate} = P_{max} * (1 - motor\ efficiency) \quad (1)$$

The heat obtained to dissipate from the system was 79.27 W, now we proceed to calculate the calorific power of de refrigerant ($\dot{Q}_{refrigerant}$), using $\dot{Q}_{dissipate}$ and a value of 1.1 that was applied as a safety factor to prevent the continuous fouling of the radiator [2].

$$\dot{Q}_{refrigerant} = \dot{Q}_{dissipate} * 1.1 \quad (2)$$

Now to calculate the radiator area (A_{rad}) using the $\dot{Q}_{refrigerant} = 87.21$ W, the value of the heat transfer coefficient of the oil (h) = W/m² K and the oil inlet temperature ($T_{in,oil}$) and the air temperature T_{air} being 313.15 K and 293.15 K, respectively. And The value obtained of A_{rad} is 2.18 m²

$$A_{rad} = \frac{\dot{Q}_{refrigerant}}{h * (T_{in,oil} - T_{air})} * 10^2 \quad (3)$$

Next the radiator area (A_{rad}) was divided into 35 plates to reduce the size of the heat exchanger, the area of each plate is 0.062 m².

$$A_{plate} = \frac{A_{rad}}{Number\ of\ plates} \quad (4)$$

Now to be able to verify this the following formula is applied, where U represents the aluminum heat transfer coefficient, A is equal to the plate area and must show a result value close to the output power of the motor, which is the dissipated heat, $T_{in,oil}$ is the temperature entering the radiator, and T_{air} is for the ambiental temperature.

$$Q = U * A * (T_{in,oil} - T_{air}) \quad (5)$$

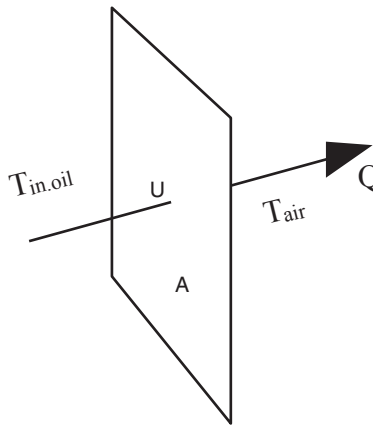


Figure 2. Calculate the heat (Q) to dissipate

Next, the heat transfer coefficient of aluminum was calculated with the following formula:

$$U = \frac{1}{\left(\frac{1}{h_{ci}} + \frac{s}{k} + \frac{1}{h_{co}}\right)} \quad (6)$$

Where, U is the total heat transfer coefficient ($\frac{W}{m^2 * K}$), k is thermal conductivity of the material ($\frac{W}{mK}$), h_{ci} and h_{co} are the individual fluid convection heat transfer coefficient of the interior or exterior wall ($\frac{W}{m^2K}$) and s is the plate thickness (m) [5][6]. The value obtained of U is 54,58 W/m² K, then proceed to calculate the heat to dissipate, a single plate was used, this was represented with a Q, the result was Q = 2368.77 W. Next step was to obtain the same heat to dissipate with the difference that this time 35 plates was used, represented with the same Q, the result was Q = 2368.77 W. Finally, with a 2.18 m² heat exchanger, it is possible to dissipate the necessary heat established by the formulas. In this analysis the use of a fan is not considered, to have a more compact and lighter design.

The heat exchanger was designed with SolidWorks, this software allows modeling, simulation, and management of 3D solids. In addition, through simulations is possible to observe the way the models would work in real circumstances. Meanwhile to obtain the simulation first was necessary to sketch the plates and tubes using a 2D simulation, then using different commands the modeling of the solid took the shape needed, at the same time the design of the radiator was based on a type of heat exchanger with smaller area and volume called compact. The plates have a square shape of length of 249 mm and width of 249 mm, with an area of 62001 mm², which is the area of the plates according to formula number 4. When multiplied by 35 (number of plates) an area of 2170035 mm² was obtained which is also the area of the radiator according to formula number 3. In addition, each plate has 30 holes of 9.55 mm diameter, distributed in three columns of five holes each, with a distance between them of 48 mm, between 98.5 mm columns, 5 mm thick, and with a total mass of 3,084 kg. Thus, the tubes have an internal diameter of 8 mm and an external diameter of 9.55 mm, a length of 340 mm, a height of 192 mm, and a total mass

of 4,593 kg. Finally, by joining all the pieces through the assembly function, a radiator with the following dimensions was obtained, with a mass of 7.68 kilograms, and a surface area of 4.86m², all conformed with aluminum plates and copper tubes.

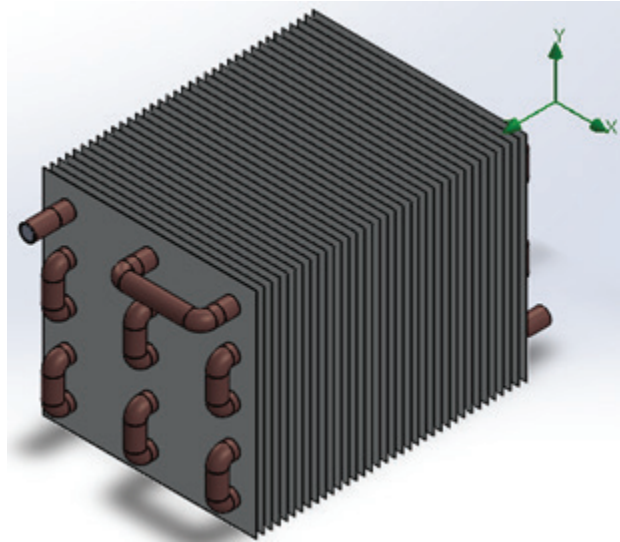


Figure 3. The final design of the heat exchanger.

Results

For the first CFD simulation, the system was analyzed based on the following conditions: environmental conditions a temperature of 20.05 °C and a static pressure of 101.325 Pa, for solid conditions they have an initial temperature of 19.00 °C and the liquid used in the simulation is olive oil that shares properties with dielectric oil. Now for the input conditions, there was a temperature of 45 °C, an ambient pressure of a temperature of 20.05 °C and static pressure of 101.325 KPa and a heat transfer coefficient of $55.45 \frac{W}{m^2 K}$ for the outer wall. The following results

were obtained: The liquid temperature decreases when passes through the radiator, it enters at around 50 °C and gets out approximately at 30 °C, a difference of 20 °C.

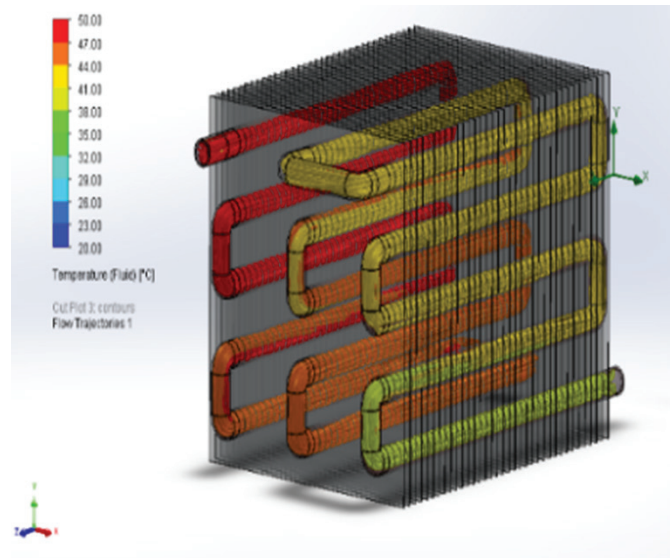


Figure 4. Flow trajectory and temperature.

Finally, the temperature of the plates and tubes at the end of the simulation had an average temperature of 21 °C and 37 °C respectively when they were initially 19 °C and 17 °C, showing an increase of temperature of 2 °C and 22 °C respectively.

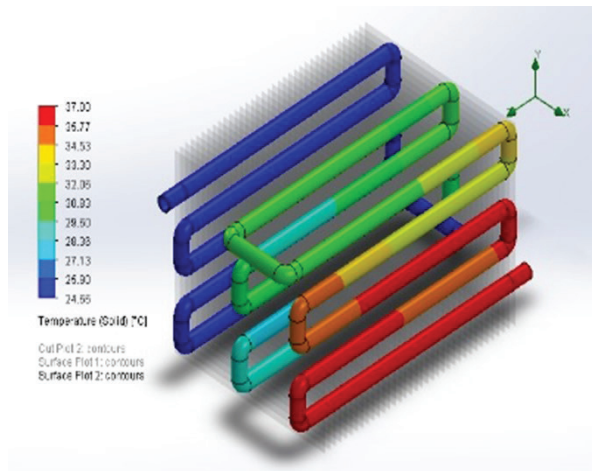


Figure 5. Plates temperature

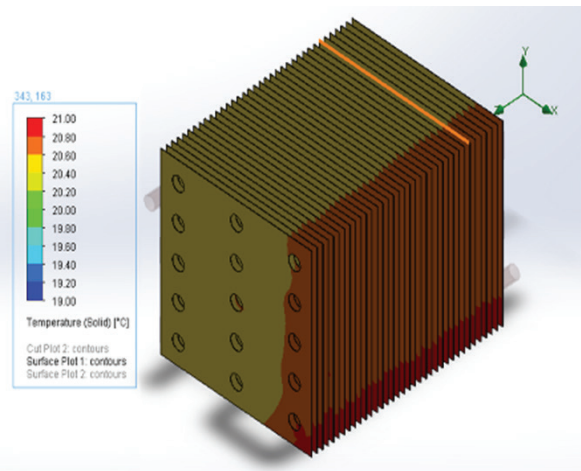


Figure 6. Tubes temperature

In various experimental studies of refrigeration systems, the difference temperature when measuring the inlet and outlet temperatures is in one of the studies approximately 15 °C to 18 °C, in another a little lower than 6 °C to 10 °C [4][7]. Therefore, in this analysis, when performing the simulation in the computer, it generates a result that, according to the design of the heat exchanger, can reduce the refrigerant temperature by approximately 20 °C, being a little higher than what can be seen in the studies mentioned, but it can be expected that it will be lower when doing the experimental test with the radiator or exchanger already built, even so, the objective of dissipating a good amount of heat while keeping the engine cooled is achieved.



Conclusions

In conclusion, thanks to the 3D design and its results it was possible to find the dimensions of the heat exchanger that can supply the necessities of the cooling system; therefore, it will be able to adapt the prototype that best illustrates the needs of the robotic joint. Additionally, using CFD simulation was the key to proving that the designed prototype of the heat exchanger can extract the heat efficiently since the first simulation does achieve the objective of extracting a significant amount of heat, reducing the temperature of the oil from 50 °C to almost 30 °C.

References

- [1] Ackerman, E. (2016). This Robot Can Do More Push-Ups Because It Sweats. Recuperated of <https://spectrum.ieee.org/automaton/robotics/humanoids/this-robot-can-do-more-pushups-because-it-sweats>.
- [2] Cabrera, Franklin. Tigre, E. (2016). Diseño y construcción de los sistemas de refrigeración de un vehículo fórmula SEA eléctrico [Tesis de Ingeniero Mecánico Automotriz, Universidad Politécnica Salesiana]. Repositorio institucional - Universidad Politécnica Salesiana.
- [3] Cengel, Y. Boles, M. (2014). Termodinámica. México. Octava Edición. McGraw Hill.
- [4] Martín, H. (2019). Diseño del sistema de refrigeración de un vehículo de fórmula SAE. [Tesis de Ingeniería Mecánica, Universidad Carlos III de Madrid]. Repositorio institucional - Universidad Carlos III de Madrid.
- [5] Engineers Edge. (2021). Convective heat transfer coefficients. Recuperated of https://www.engineersedge.com/heat_transfer/convective_heat_transfer_coefficients_133.
- [6] Engineering ToolBox. (2003). Overall Heat Transfer Coefficient. [online] Recuperated of https://www.engineeringtoolbox.com/overall-heat-transfer-coefficient-d_434.html.
- [7] Gavilema, H. (2014). Estudio teórico y experimental de los parámetros de funcionamiento de un motor de combustión interna a gasolina a diferente concentración de agua-refrigerante [Tesis de Ingeniero Mecánico Automotriz, Escuela Superior Politécnica de Chimborazo]. Repositorio institucional - Escuela Superior Politécnica de Chimborazo.
- [8] Guarino, B. (2016). The amazing sweating robot. EU: The Washington Post. Recuperated of <https://www.washingtonpost.com/news/morning-mix/wp/2016/10/17/the-remarkable-sweating-robot-researchers-cool-overheating-machines-in-human-style/>
- [9] Hernández, R., Fernández, C., Baptista, P. (2010). Metodología de la investigación. McGRAW-HILL. (Original publicado en 1991).
- [10] Jaramillo, O. (2007). Intercambiadores de calor. Centro de Investigación en Energía. Universidad Nacional Autónoma de México. <http://www.cie.unam.mx/~ojs/pub/HeatExchanger/Intercambiadores.pdf>.
- [11] Kakiuchi, Y. ET. AL. (2015). Development of humanoid robot system for disaster response through team NEDO-JSK'S approach to DARPA. Corea del Sur: IEEE.
- [12] Kubota, T. (2017). Stanford researchers develop a new type of soft, growing robot. Stanford, EU: Stanford News. Recuperated of <https://news.stanford.edu/2017/07/19/stanford-researchers-develop-new-type-soft-growing-robot/>
- [13] Pia, M. ET. AL (2016). Direct Liquid Cooling in Low-Power Electrical Machines: Proof-of-Concept. Japan: IEEE.
- [14] Rodríguez, M. Ruiz, F. (2019). Comparative efficiency study of two proposed designs tested in water- and air-cooling conditions for a high-power humanoid robot hollo joint. Costa Rica. INII.
- [15] Rodríguez, M. Ruiz, F. (2019). Torque sensor geometry study on 3D simulated conditions for a Hollow robotic joint. Costa Rica. INII.

Accelerating machine learning at the edge with approximate computing on FPGAs

Acelerando aprendizaje de máquina en el *Edge* con computación aproximada en FPGAs

Luis Gerardo León-Vega¹, Eduardo Salazar-Villalobos², Jorge Castro-Godínez³

León-Vega, L.G.; Salazar-Villalobos, E.; Castro-Godínez, J. Accelerating machine learning at the edge with approximate computing on FPGAs. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 39-45.

 <https://doi.org/10.18845/tm.v35i9.6491>

- 1 Master's in Electronics Student. Instituto Tecnológico de Costa Rica. Costa Rica. PhD fellow. Università degli Studi di Trieste. Italy. E-Mail: lleon95@estudiantec.cr  <https://orcid.org/0000-0002-3263-7853>
- 2 Electronics Engineering Student. Instituto Tecnológico de Costa Rica. Costa Rica. E-Mail: eduardosalazar@estudiantec.cr
- 3 Assistant Professor. School of Electronics Engineering. Instituto Tecnológico de Costa Rica. Costa Rica. E-Mail: jocastro@tec.ac.cr

Keywords

Approximate computing; edge computing; machine learning; neural networks; linear algebra.

Abstract

Performing inference of complex machine learning (ML) algorithms at the edge is becoming important to unlink the system functionality from the cloud. However, the ML models increase complexity faster than the available hardware resources. This research aims to accelerate machine learning by offloading the computation to low-end FPGAs and using approximate computing techniques to optimise resource usage, taking advantage of the inaccurate nature of machine learning models. In this paper, we propose a generic matrix multiply-add processing element design, parameterised in datatype, matrix size, and data width. We evaluate the resource consumption and error behaviour while varying the matrix size and the data width given a fixed-point data type. We determine that the error scales with the matrix size, but it can be compensated by increasing the data width, posing a trade-off between data width and matrix size with respect to the error.

Palabras clave

Computación aproximada; computación periférica; aprendizaje por computador; redes neuronales; álgebra lineal

Resumen

La inferencia en algoritmos complejos de aprendizaje automático (ML, por sus siglas en inglés) en *edge computing* está tomando importancia para desvincular la funcionalidad de un sistema de la nube. Sin embargo, los modelos de ML incrementan en complejidad más rápido que los recursos de hardware disponibles. Esta investigación tiene como objetivo acelerar el aprendizaje por automático al delegar el cálculo computacional a FPGAs de baja gama y usar computación aproximada para optimizar el uso de los recursos aprovechando la naturaleza inexacta de los modelos de aprendizaje automático. En este artículo, proponemos un diseño de un elemento de procesamiento genérico de multiplicación-suma de matrices, parametrizado en el tipo de dato, tamaño de la matriz y ancho del dato. Evaluamos el comportamiento del consumo de recursos y del error mientras variamos el tamaño de la matriz y el ancho del dato un tipo de datos de punto fijo. Determinamos que el error escala con el tamaño de la matriz pero que puede ser compensado al incrementar el ancho del dato, representando un compromiso entre el ancho del dato y el tamaño de la matriz con respecto al error.

Introduction

Deep neural network (DNN) models constantly increase in complexity over time, while edge computing systems tend to be steady in their capabilities [1]. Power consumption is a common concern when moving inference to the edge, adding complexity to dealing with the trade-off between resources and inference time. One possibility to add more control over this trade-off is to use Field Programmable Gate Arrays (FPGA), which are devices capable of reconfiguring their logic and implementing custom hardware designs. This implies exploring the synthesis of inference algorithms described with Hardware Description Languages (HDL) or using High-Level Synthesis (HLS), tools capable of converting C/C++ untyped code to Register Transfer Level

(RTL) [2,3]. There are also pre-built alternatives to leverage inference to FPGAs using vendor IP cores, such as Vitis AI and OpenVINO [4,5]. However, most alternatives focus on mid-end to high-end FPGAs, leaving aside the low-end solutions that exhibit less power consumption.

One downside of the pre-built alternatives is that they limit the optimisation opportunities beyond pruning and fixed-point representation. This research aims to address the use of FPGAs to accelerate machine learning operations rather than represent the whole model in hardware, precisely placing the accelerators required for the model execution, improving hardware utilisation, and achieving better energy consumption. The key contribution of this research is the proposal of a framework for creating custom architectures according to the model requirements, in particular, tuned in arbitrary numerical representation, number of processing elements (PE), and quantisation techniques.

Background and Related-Work

According to the state-of-the-art, there are implementations for accelerating matrix-matrix multiplications and convolutions at the industry level. Intel and NVIDIA propose AMX and Tensor cores, respectively, which accelerate matrix fused multiply-add (FMA) [6,7]: $D = AB + C$ integer arithmetic. Convolutions have been optimised through mathematical manipulation. For instance, the Winograd algorithm reduces the number of multiplications involved in the convolution by using a domain transformation [8]. It presents multiple advantages, such as memory footprint and better performance than the spatial convolution, achieving more than 3x in speedup [8].

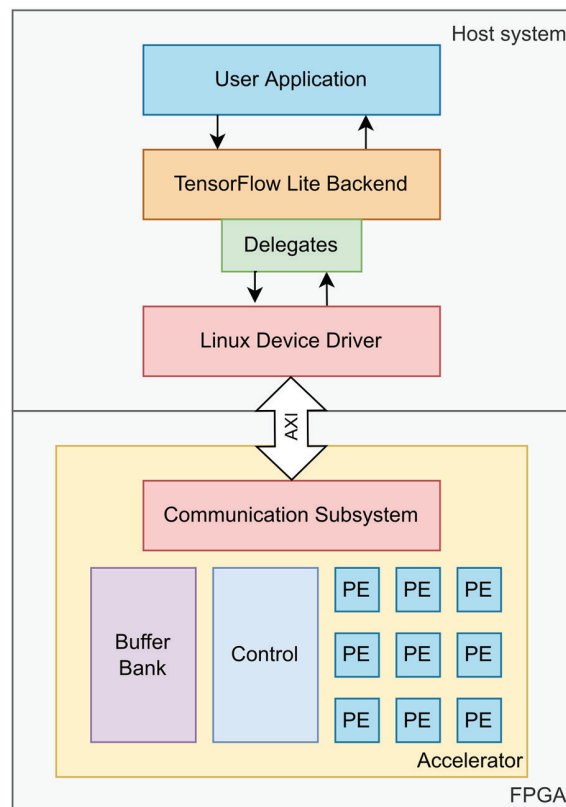


Figure 1. Structure of the research proposal.

Our proposal is presented in Figure 1. It considers the implementation of 1) a set of generic accelerators implemented in standard C++, synthesised using Vivado HLS, and parameterised in datatype, the number of bits, arithmetic operands and number of PEs. 2) A device driver to communicate the accelerators to Linux as the host Operating System and sample implementations using Tensorflow for benchmarking. These implementations are part of a framework for accelerating inference in low-end FPGAs.

The research analyses common mathematical operations performed during model inference, mainly matrix-matrix multiplications, convolutions and non-linear activations. It leads to having a baseline for the accelerators library, which is parameterised to make the designs more flexible than the typical implementations offered in CPUs, GPUs and TPUs, allowing the designer to explore optimisation opportunities. After the first set of implementations in HLS, the PEs will be analysed to model their behaviour when tuning the numerical precision and varying the operand sizes, observing their impact on resource utilisation, numerical errors and energy consumption. It will help users create and evaluate accelerator candidates according to the complexity of the ML model and their design constraints.

For this work, we will focus on the first stage of our framework (the set of accelerators), particularly with a fixed-width matrix fused multiply-add (FMA) PE, parameterised in datatype, data width, and matrix size. In this case, the FMA operator implementation keeps the data width in the output, involving a scaling factor that depends on the matrix dimensions to mitigate the overflow:

$$d' = \frac{ab}{2N} + \frac{c}{2N}, d = 2Nd' \quad (1)$$

where a , b , and c are the input operands, N is the number of rows considering a square matrix, and d is the output.

Methodology

We start the implementation of the matrix FMA by using standard C++ to define the algorithm itself. After having a correct implementation using floating-point numbers, we start by parameterising the implementation at the data type, matrix size, and the number of bits level. It will lead to inaccuracies in the results that must be evaluated before applying the implementation to an actual application. In this case, we evaluate the errors by using a normalised mean error

$\underline{\xi}$ and its standard deviation $\sigma(\underline{\xi})$, computed as

$$\underline{\xi} = \frac{1}{\alpha K} \sum_{i=1}^K \xi_i, \sigma(\underline{\xi}) = \frac{1}{\alpha K} \sqrt{\sum_{i=1}^K |\xi_i - \underline{\xi}|^2}, \xi_i = |y_i - \hat{y}_i| \quad (2)$$

where K is the number of elements of the matrix evaluated (including 10 iterations with different seeds), y_i is the i -th floating-point result (matrix entries), \hat{y}_i is the measured value in a custom data type and a number of bits, and α is the normalisation factor (set to 2 given the numerical range from $]-1, 1[$).

After evaluating the numerical impact, we proceed with the iterative design optimisation, adding pipelining, registers, and concurrency to the implementation. These optimisations are added through Vivado HLS directive files. In this case, we will present our Pareto's optimal solution.

Results

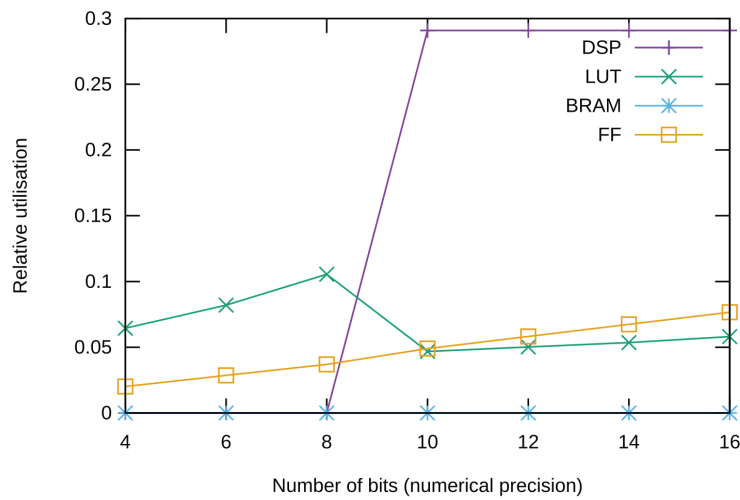
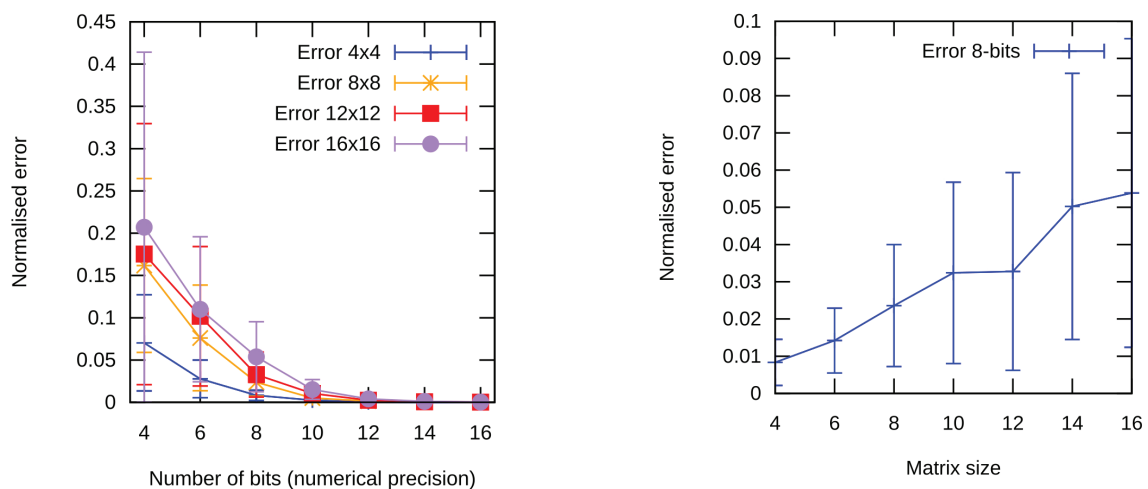


Figure 2. Resource consumption of the current best FMA processing element for 8x8 matrices. The resources analysed are Digital Signal Processors (DSPs), Look-Up Tables (LUTs), Block Random Access Memories (BRAMs), and Flip-Flops (FFs).

Figure 2 presents the results of our current FMA implementation (see [9]). It scales linearly with respect to the data width concerning the FFs used in the design. It has a discontinuity in the LUTs, switching the logic from LUTs to DSPs when the number of bits reaches more than 10 bits, consuming almost 30% of an Avnet Zedboard (based on the Xilinx XC7Z020). Neither of the implementations consumes BRAMs since the PE corresponds to an operator, and no memory is needed. However, the consumption depends on the design mapping when implementing the logic into RTL and the user requirements.



Normalised error while varying the number of bits of the fixed-point numerical representation for four matrix sizes

Normalised error while varying the matrix size for an 8-bit fixed-point representation

Figure 3. Resource consumption of an 8-bit fixed-point 8x8 matrix MAC. The FMA with registers results is the intermediate case in resource utilisation.

Since our FMA can specialise in arbitrary precisions, Figure 3 explores the numerical errors when tuning the numerical representation using (2). Figure 3.a shows how the error evolves as the number of bits invested increases, leading to an exponential decreasing behaviour.

The preservation of the number of bits of the output with respect to the input through the scaling of one of the registers, as described in (2), introduces a dependency on the matrix size in the error as presented in Figure 3.b. In this case, the bigger the matrix, the greater the error. It can also be noticed in Figure 3.a, showing that the most compressed precisions are heavily affected by the matrix size. A 4-bit representation scales the mean relative error from 7.5% up to 21%, having more than 40% peaks. As the number of bits increases, the error brought by the matrix size is compensated.

Conclusions

This work presented our matrix FMA as the first accelerator analysis of our deep learning inference framework for FPGAs. This framework aims to automate the generation of PEs for accelerating inference tasks while adding approximate computing to address issues concerning resource and energy consumption on low-end FPGAs. Our matrix FMA implementation shows a linear scaling in the resource consumption before less than 10 bits in the data width, showing for an 8-bit 8x8 matrix FMA a consumption of up to 11%. When having more bits, the consumption becomes constant at almost 30%, where the scaling in resource consumption becomes agnostic to the number of bits from 10 up to 16 bits. In terms of errors, we noticed the influence of the matrix size on the error, where large matrix sizes lead to high variance and coarse errors, which can be compensated by increasing the numerical precision.

In future work, we are extending the analysis in the error characterisation to get models without the need for simulations. Moreover, more PEs are being developed.

Acknowledgements

This work is supported by RidgeRun LLC, through its Master's scholarship for collaborators, and by the Instituto Tecnológico de Costa Rica.

References

- [1] C. J. Wu, D. Brooks, K. Chen, D. Chen, et al., "Machine learning at facebook: Understanding inference at the edge" Proceedings 25th IEEE International Symposium on High Performance Computer Architecture, HPCA 2019, pp. 331–344, 2019. <https://doi.org/10.1109/HPCA.2019.00048>
- [2] B. C. Schafer and Z. Wang, "High-Level Synthesis Design Space Exploration: Past, Present, and Future" in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 10, pp. 2628–2639, Oct. 2020, <https://doi.org/10.1109/TCAD.2019.2943570>.
- [3] Z. Wang and B. C. Schafer, "Learning from the Past: Efficient High-Level Synthesis Design Space Exploration for FPGAs" in ACM Transactions on Design Automation of Electronic Systems, vol. 27, no. 4, Jul. 2022, <https://doi.org/10.1145/3495531>.
- [4] T. Liang, J. Glossner, L. Wang, S. Shi and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey", *Neurocomputing*, vol. 461, pp. 370–403, 2021. <https://doi.org/10.1016/j.neucom.2021.07.045>
- [5] T. González, J. Castro-Godínez. "Improving Performance of Error-Tolerant Applications: A Case Study of Approximations on an Off-the-Shelf Neural Accelerator" in V Jornadas Costarricenses de Investigación en Computación e Informática (JoCICI 2021), Virtual Event, Oct. 2021.
- [6] Intel, "Intel® Architecture Instruction Set Extensions and Future Features", Intel Corporation, May 2021. [Online]. Available: <https://software.intel.com/content/www/us/en/develop/download/intel-architecture-instruction-set-extensions-programming-reference.html>

- [7] NVIDIA Corporation, "NVIDIA TESLA V100 GPU ARCHITECTURE" 2017. [Online]. Available: <https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf>
- [8] Andrew Lavin and Scott Gray. 2016. Fast algorithms for convolutional neural networks. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 4013–4021. <https://doi.org/10.1109/CVPR.2016.435>
- [9] Salazar-Villalobos, Eduardo, and Leon-Vega, Luis G. (2022). Flexible Accelerator Library: Approximate Matrix Accelerator (v1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.6272004>

Design of a microkernel-based services manager for IoT with isolated processing

Diseño de un gestor de servicios IoT basado en *microkernel* con procesamiento aislado

Jose Antonio Ortega-González¹, Luis G. León-Vega²

Ortega-González, J.A.; León-Vega, L.G. Design of a *microkernel*-based services manager for iot with isolated processing. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 46-52.

 <https://doi.org/10.18845/tm.v35i9.6492>

- 1 Computer Engineering Student, Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: j.ortega98@estudiantec.cr
 <https://orcid.org/0000-0002-2924-7037>
- 2 Master's in Electronics Student, Instituto Tecnológico de Costa Rica. Costa Rica. E-mail: lleon95@estudiantec.cr
 <https://orcid.org/0000-0002-3263-7853>

Keywords

Internet of things; scheduling algorithms; protocols; high-performance computing; web services; distributed computing.

Abstract

The development of projects related to the Internet of Things has generated a significant impact on the growth of the software industry. Although there are several alternatives to their development, users are limited by inherent factors, such as vendor lock-in, software development flexibility, and security. This article shows a platform design that provides a deployment and management system for IoT projects avoiding some of the limitations of the current tools, such as vendor lock-in and limitations in the development technologies.

Palabras clave

Internet de las cosas; algoritmos de calendarización; protocolos; computación de alto rendimiento; servicios web; computación distribuida.

Resumen

El desarrollo de proyectos relacionados con el Internet de las Cosas (IoT, por sus siglas en inglés) han generado un impacto significativo en el crecimiento de la industria del software. Si bien existen alternativas para su desarrollo, los usuarios se ven limitados por factores inherentes, tales como la dependencia de un proveedor, flexibilidad de desarrollo de software y seguridad. Este artículo muestra el diseño de una plataforma que facilita el despliegue y manejo de proyectos de IoT, solventando algunos de los retos actuales del mercado como los bloqueos de proveedor y limitaciones en las tecnologías de desarrollo.

Introduction

Web services development has been rapidly evolving, moving from monolithic and microservices architectures until arriving at the serverless approach, where a provider offers the option to only upload programmed functions. In particular, the serverless architecture came as a solution to reduce operational costs and has proven to be a good fit for IoT applications [1], with quick set-up times and focusing the development on the application functionality.

Serverless is a paradigm that provides processing resources based on uploading and setting individual functions executed on demand and includes subtasks to other services without having them in the same monolithic environment, such as running a database query or sending a message through the network. The serverless provider fully manages server maintenance tasks, software installation, and security.

Amazon, Google, and Microsoft offer several alternatives for serverless applications. However, most of these products have vendor lock-in or technology limitations that restrict users to migrate from one service to other, as mentioned in [2], tying customers to a single cloud provider and making it difficult and expensive to move to a different vendor in the future. Moreover, there are restrictions to the functionality and the software that can run on the serverless instance. Therefore, there are concerns about vendor lock-in when adopting a cloud service in new projects [3].

There are open-source approaches such as Apache OpenWisk, OpenFaaS, Knative, and Kubeless, that offer an alternative to avoid vendor lock-in [4]. While these solutions provide an advance, there are also remaining challenges to attend to, in particular, the security risk due to the usage of containers [5]. Also, this project integrates services such as storage, communication, and databases, with the computational power of the serverless approach.

This work aims to create an open-source platform that handles automated infrastructure creation and deployment to speed up IoT projects development avoiding vendor lock-in and improving security. Our target is a *high-performant, robust, flexible, scalable, reliable, and easy to use* system.

General System Design

The system follows a microkernel-like architecture, which provides communication and scheduling functionalities and delegates other functionalities as services connected to it [6]. In our work, the services comply with a serverless infrastructure for delivering basic web service functionality.

Figure 1 shows a top-level architecture for the system where services are connected to the *Kernel* using Inter-process communication (IPC). The Kernel controls the operation providing an agnostic communication mechanism between the services. The *Doer* is an isolated environment that manipulates *Workers* for process execution. The *DBriver* is a database administrator that handles SQL and NoSQL database entities. The *CommServer* handles the communication services based on HTTP and MQTT network protocols. The *Resadmin* is responsible for the management of the resources and the users. The Control Gate dashboard allows users to create and control their IoT Projects entities.

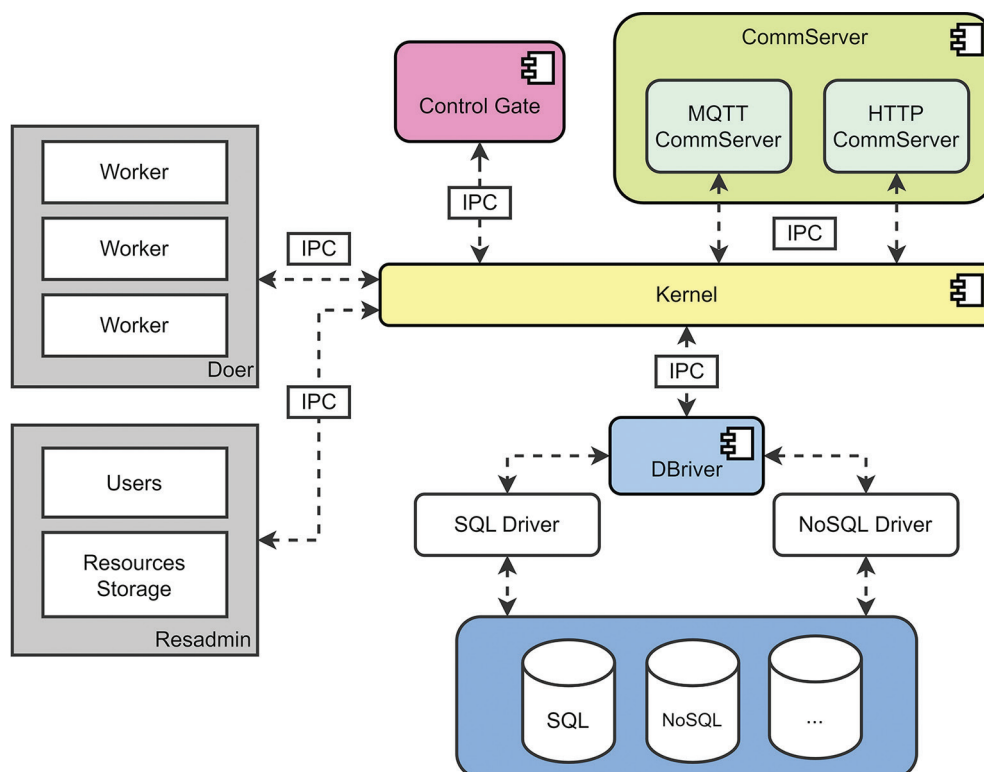


Figure 1. Top-level platform architecture.

Kernel

Figure 2 shows the Kernel design. The Kernel consists of four main modules and one module for interface purposes. We start with the *IPC Module*, which handles all the communication amongst the services. This communication is implemented over an IPC protocol provided by ZeroMQ [7], a high-performance message passing library that provides Operating System (OS)-agnostic support for multiple protocols and communication patterns to create an infrastructure for several services. The IPC Module connects to the *IPC Connections* that provide the mechanism for each service to send and receive a message from the Kernel and other services. It can be seen as an abstracted software interface.

We propose the *Process Controller* for the processes and their management, which handles all the requests that arrive at the system and manages them as processes. Besides, we propose a priority-based queue as *Scheduler*. It handles the message (process) dispatching for each service. Each queue is a Red-Black Binary Search Tree (BST) where the message with the highest priority will be at the left, similar to the Completely Fair Scheduler (CFS) from Linux. Last, we propose a *Services Controller* to handle the information related to the services connected to the Kernel.

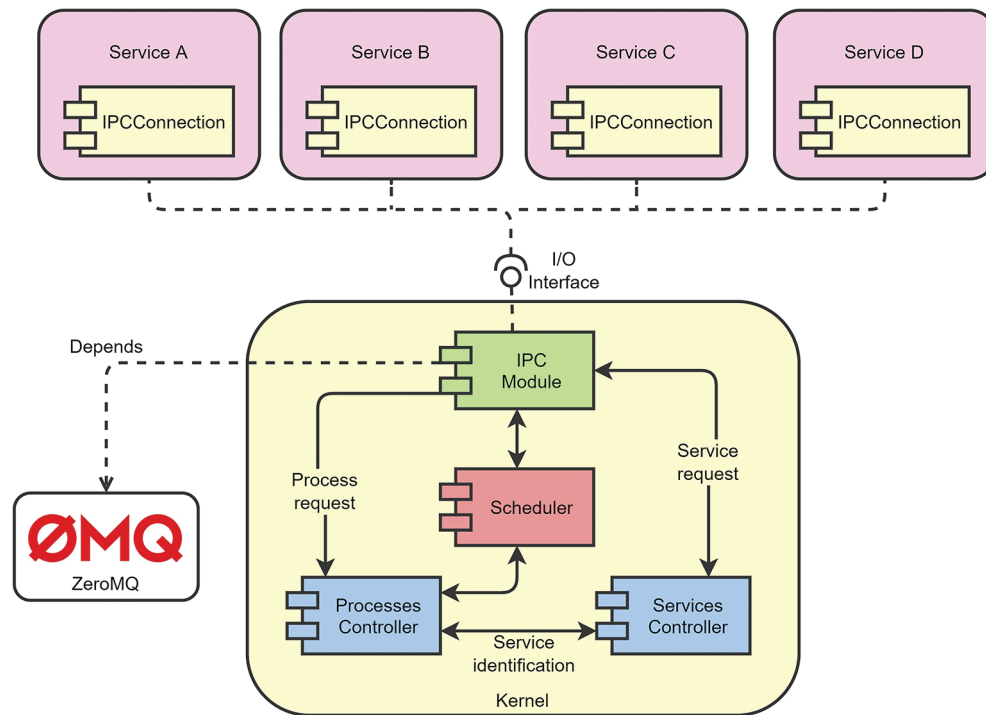


Figure 2. Kernel design overview.

Doer

Figure 3 shows a design overview of the Doer. The Doer is a serverless implementation based on Singularity [8] as a virtualization tool also used in FuncX [9] for Python serverless execution, which could be used in local and cloud environments, provides better security than other tools such as Docker, and focuses on native integration of high-performance computing tools.

The logic for each project is executed using the Doer, which includes three main modules: 1) The *Doer Admin* handles the requests that arrive at the Doer and holds them as *tasks*. It manages three queues depending on the state of each task. 2) The *Worker Handler* invokes and controls the *Workers*, isolated environments based on containers for executing the programmed functions (called in this project as *Doees*). 3) The *Worker*: is the environment where the functions are executed. We are using Singularity containers to provide a safe and isolated environment. Furthermore, 4) The *Doee Executor* handles the execution of the *Doees* within the containers. It resides in each worked as a server daemon and is triggered when the *Worker* receives an execution request.

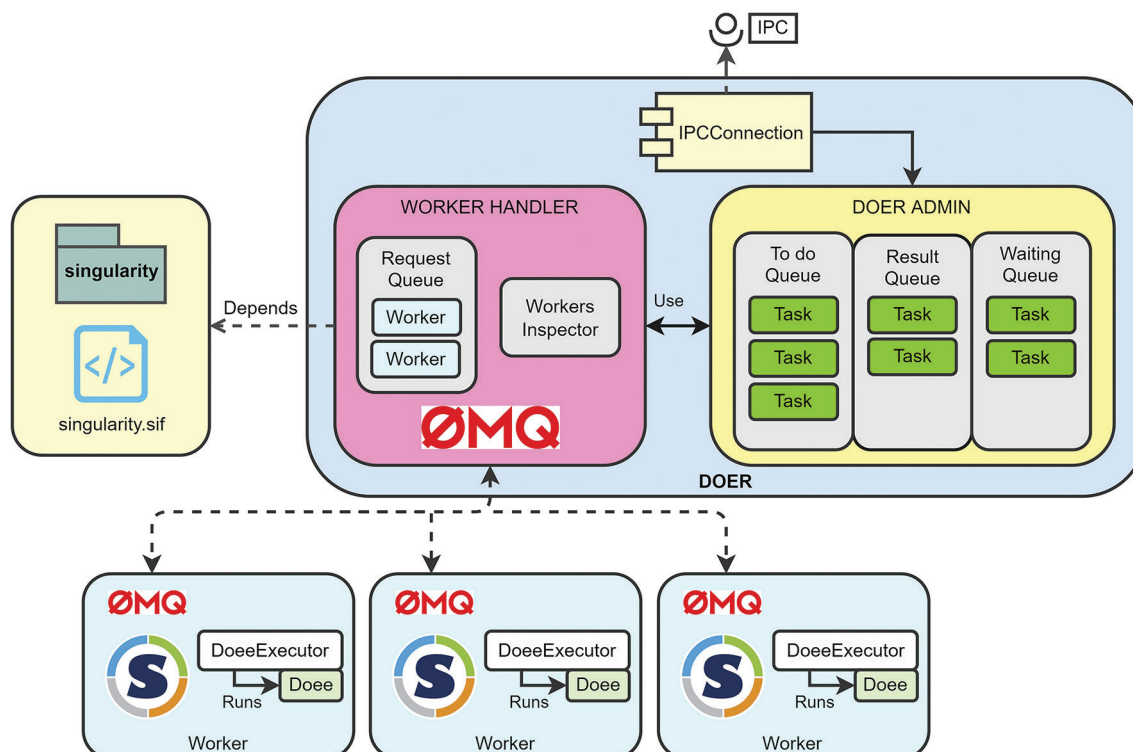


Figure 3. Doer design overview

Since the Doer involves the communication of a host server with isolated Singularity containers, we propose using of ZeroMQ as the communication bridge.

Doee

The functions executed by the *Workers* are called *Doees*. They are custom pre-compiled shared objects, compiled just after the source code of the *Doee* is added to the platform. A *Worker* executes the *Doee* after a trigger, which can be an action specified like an HTTP GET request. It adds support for other programming languages by using object bindings and interfaces.

The *Doee Executor* requests the *Resadmin* to provide the pre-compiled binaries of a single *Doee*, so the *Worker* will only contain one *Doee* at a time. Figure 4 shows the structure of the *Doees* and how they are created.

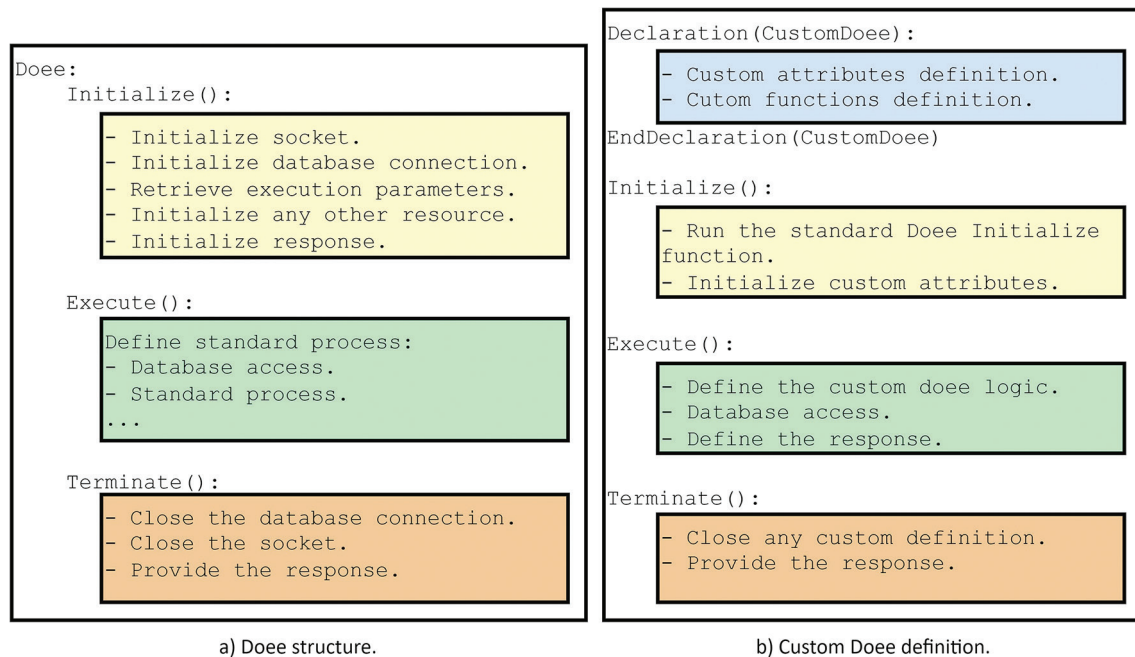


Figure 4. Example of the Doee usage.

Figure 4.a is the structure of the Doee interface class; it requires three main functions: 1) *Initialize* sets the Doee resources required to execute the process, initializes the socket for external communication, the database connection for queries, and retrieves the parameters to be used in the process. 2) *Execute* runs the Doee standard process and a specific logic implemented by the actual user implementation (concrete Execute method). Then, we have 3) *Terminate*, which closes the connections and resources initialized by the Initialize function.

The Doee class is a pure virtual class, and it cannot be executed like any other function. It shall be implemented by a Custom Doee defined by the developer through inheritance, where the specific logic of the function is implemented. Figure 4.b shows how the custom Doee is defined:

1. *Declaration*: The Custom Doee is declared, and the inheritance is automatically added. Also, between the declaration, the developer can define attributes and functions to be used for this Custom Doee.
2. *Initialize*, *Execute* and *Terminate* works as same as for the Doee class.
3. Custom Doee is the function that is executed in the Worker.

The current development of the project can be found in [10]. Updates and new modules will be added there.

Conclusions

Our proposal addresses the vendor lock-in by offering an open-source solution capable of being installed in any computing premise proposed by Apache Open Wisk, Open FaaS, Knative, and Kubeless. Moreover, our contribution to safety is based on Singularity, as applied by FuncX, which differs from many current cloud-based methods in the market.

The platform has more services under development that will be addressed in future work, such as the Resadmin, DBriver, CommServers, and the Control Gate. These modules provide a more integrated platform that offers storage and management to the overall infrastructure. Besides, we plan to compare FuncX and our platform to validate our implementation, executing Python functions.

References

- [1] G. McGrath and P. R. Brenner, "Serverless Computing: Design, Implementation, and Performance," *IEEE*, vol. 37th, pp. 405-410, 2017.
- [2] J. Opara-Martins, R. Sahandi and F. Tian, "Critical review of vendor lock-in and its impact on adoption of cloud computing," *International Conference on Information Society*, pp. 92-97, 2014.
- [3] J. Opara-Martins, R. Sahadi and F. Tian, "Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective," *Journal of Cloud Computing*, pp. 1-18, 2016.
- [4] A. Palade, A. Kazmi and S. Clarke, "An Evaluation of Open Source Serverless Computing Frameworks Support at the Edge," *2019 IEEE World Congress on Services (SERVICES)*, pp. 206-211, 2019.
- [5] S. Sultan, I. Ahmad and T. Dimitriou, "Container security: Issues, challenges, and the road ahead," *IEEE Access*, vol. 7, pp. 52976-52996, 2019.
- [6] A. S. Tanenbaum, *Modern operating systems*, Boston: Pearson, 2015.
- [7] ZeroMQ, "ZeroMQ," 2021. [Online]. Available: <https://zeromq.org/>. [Accessed 2021].
- [8] "SingularityCE," Sylabs, 2022. [Online]. Available: <https://sylabs.io/singularity>. [Accessed 2022].
- [9] R. Chard, T. J. Skluzacek, Z. Li, Y. Babuji, A. Woodard, B. Blaiszik, S. Tuecke, I. Foster and K. Chard, "Serverless Supercomputing: High Performance Function as a Service for Science," 2019.
- [10] J. A. Ortega and L. León, "IoT Services Management System [Source Code]," 2022. [Online]. Available: <https://gitlab.com/klooid/isms/isms-kernel>. [Accessed 2022].