Estimación de incertidumbre para un sistema de reconocimiento de voz

Uncertainty estimation for a speech recognition system

Walter Morales-Muñoz[1], Saúl Calderón-Ramírez[2]

Morales-Muñoz, W; Calderón-Ramírez, S. Estimación de incertidumbre para un sistema de reconocimiento de voz. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 97-103.

https://doi.org/10.18845/tm.v37i7.7305

1 Instituto tecnológico de Costa Rica. Costa Rica.
wmorales@itcr.ac.cr
https://orcid.org/0000-0002-3888-4951
2 Instituto Tecnológico de Costa Rica. Costa Rica.
sacalderon@itcr.ac.cr
https://orcid.org/0000-0001-9993-4388

Palabras clave

Incertidumbre; Reconocimiento de voz; ASR; Whisper; Monte Carlo Dropout.

Resumen

Whisper es un sistema de reconocimiento de voz diseñado por la compañía OpenAI, dicho sistema ha sido entrenado con 680,000 horas de datos supervisados multilingües y multitarea recopilados de la web. La siguiente investigación tiene como objetivo adaptar y emplear la técnica de Monte Carlo Dropout utilizando datos audios etiquetados en español y contaminados con una cantidad de ruido y la distancia de Levensthein para estimar la incertidumbre de dicho sistema. Resultados preliminares muestran que existe una relación lineal entre la estimación de la incertiumbre utilizando la distancia Levensthein y el medoide respecto al Word Error Rate (WER) de las transcripciones, además se observa que la cantidad de inserciones u omisiones en las transcripciones tiende a ser bajo.

Keywords

Uncertainty; Speech Recognition; ASR; Whisper; Monte Carlo Dropout.

Abstract

Whisper is a voice recognition system designed by the company OpenAI, which has been trained with 680,000 hours of multilingual and multitask supervised data collected from the web. The following research aims to adapt and employ the Monte Carlo Dropout using audio data labeled in Spanish and contaminated with a certain amount of noise and Levensthein distance to estimate the score uncertainty of this system.Preliminary results show that there is a linear relationship between uncertainty estimation and the Word Error Rate (WER) of the transcriptions. Furthermore, it is observed that the number of insertions or omissions in the transcriptions tends to be low.

Introducción

Los modelos de Deep Learning se centran en generar aplicaciones del mundo real, abarcando diversos ámbitos como la visión por computadora, el procesamiento del lenguaje natural, las finanzas, la robótica, el reconocimiento de voz y más. Independientemente del dominio de la aplicación, los modelos de Deep Learning utilizan redes neuronales profundas para aprender autónomamente cómo realizar tareas complejas a partir de grandes cantidades de datos. Para garantizar la fiabilidad de estos modelos en las distintas tareas específicas que realizan en aplicaciones cotidianas, es necesario evaluar cuán seguras son las predicciones que realizan estos modelos. Si los datos de prueba difieren significativamente de los datos de entrenamiento, es posible que el rendimiento del modelo no sea óptimo.

Como se menciona en [1], el tema de medir la calidad de los datos a través de diferentes métricas requiere más investigación, especialmente en el caso de datos no estructurados, que se utilizan comúnmente en la mayoría de las aplicaciones de Deep Learning. Este es el caso de aplicaciones que se centran en el reconocimiento automático de voz (ASR), donde las señales de audio (datos no estructurados) se utilizan como entrada para producir transcripciones precisas de voz a texto. Es importante comprender cómo se puede estimar la pérdida de confianza que el modelo puede experimentar en sus predicciones, ya sea debido a la variabilidad inherente en la adquisición de datos de entrada o a la arquitectura del modelo.

En el caso de los sistemas de Reconocimiento Automático de Voz (ASR), existen varios tipos, uno de los cuales es el sistema Whisper diseñado por la empresa OpenAI. Como se menciona en [2], este sistema ha demostrado ser uno de los sistemas ASR más potentes en la actualidad. Ha sido entrenado con 680,000 horas de datos supervisados multilingües recopilados de la web. Los autores afirman que el uso de un conjunto de datos tan grande y diverso conduce a una mayor robustez del sistema, incluso si los datos de entrada utilizados para las predicciones incluyen acentos específicos, ruido de fondo y otras variaciones.

Con esto en mente, el objetivo es evaluar la confiabilidad del sistema de reconocimiento de voz Whisper, donde se utilizará la técnica de Monte Carlo Dropout para capturar la incertidumbre del modelo. La importancia de resolver este problema radica en la necesidad de establecer técnicas que promuevan el uso seguro de aplicaciones impulsadas por IA, que ahora son ampliamente utilizadas por el público a diario. Esta investigación tiene como objetivo fomentar el uso seguro de varios modelos de acceso público y proporcionar ideas valiosas para empresas que dependen del sistema de reconocimiento Whisper y requieren un análisis detallado de su rendimiento.

La importancia de abordar este problema propuesto es evidente en un contexto en el que la Inteligencia Artificial ha experimentado un crecimiento exponencial, lo que ha llevado a la automatización de tareas que antes se consideraban imposibles de automatizar. Evaluar el rendimiento de estos modelos promueve la adopción segura de estas tecnologías.

En términos generales, para cualquier modelo de Deep Learning como Whisper, la incertidumbre del modelo, como se propone en [3], puede expresarse de acuerdo con la ecuación (1), y puede aproximarse como se muestra en la ecuación (2).

$$p(y^* \mid x^*, X, Y) = \int p(y^* \mid W, x^*, X, Y)\, p(W \mid x^*, X, Y)\, dW \tag{1}$$

$$\approx \int p(y^* \mid W, x^*)\, p(W \mid X, Y)\, dW \tag{2}$$

Según lo anterior, para calcular la incertidumbre de diferentes modelos de Deep Learning, es necesario modelar las distribuciones posteriores dadas por la ecuación (2). En términos generales, para cualquier modelo de Deep Learning como Whisper, la incertidumbre del modelo, como propone [3], puede expresarse de acuerdo con la ecuación (1), donde y^* es una posible transcripción dada por los parámetros W del modelo obtenidos a partir de un clip de audio de prueba x^* y los datos de entrenamiento previos X, Y. La ecuación (1) puede aproximarse como se muestra en la ecuación (2). Sin embargo, como afirma [4], para los modelos de Deep Learning, la distribución $p(W \mid X, Y)$ es intratable, y nuestro caso no es una excepción. La ecuación (4) depende de $p(Y \mid X)$, lo que significa que se basa en el cálculo de la verdadera distribución de probabilidad del conjunto de datos, y dada la naturaleza del problema, esto no se puede hacer analíticamente. Por lo tanto, se debe utilizar un enfoque para aproximar la ecuación (2). Hay varios enfoques informados en la literatura, pero un método comúnmente utilizado que no implica modificar la arquitectura del modelo es aproximar la distribución $p(W \mid X, Y)$ y así cuantificar la incertidumbre como se indica en la ecuación (2). Como sugiere [4], una manera de hacer esto es utilizando la aproximación mostrada en la ecuación (3) donde ϕ representa el dropout, un concepto ampliamente utilizado en aplicaciones de aprendizaje profundo. El término dropout puede entenderse como las tasas de Bernoulli que permiten la desactivación de neuronas específicas dentro de la arquitectura del modelo. El significado de dropout está determinado por la variable aleatoria de Bernoulli que decide si la entrada conectada en cada capa de la red neuronal debe ser descartada o no. Dado lo anterior, como se muestra en [4] y demostrado por [5], con la aproximación realizada, se puede mostrar que la incertidumbre

epistémica del modelo, como se indica en la ecuación (2), puede aproximarse mediante (4). Donde $\{y\}_{t=1}^{T}$ es un conjunto de T muestras de salida obtenidas de los parámetros W del modelo cuando se activa el dropout, y $\overline{y} = \frac{1}{T}\sum_{t=1}^{t} y_t$ es la media de la salida.

$$p(W \mid X, Y) \approx q(W; \Phi) = Bernoulli(W; \Phi) \tag{3}$$

$$p(y^* \mid x^*, X, Y) \approx Var_{p(y|x)}^{model}(y) = \sigma_{model} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2 \tag{4}$$

Esta técnica es conocida como Monte Carlo Dropout, se utiliza dropout durante la inferencia para obtener una estimación de la incertidumbre epistémica del modelo y es una técnica muy utilizada en el área de visión computacional. Esta técnica se puede adecuar para problemas de reconocimiento de voz como veremos más adelante, la idea general es que el dropout se aplica múltiples veces en diferentes capas del modelo de red neuronal durante la inferencia. Al generar múltiples transcripciones para entradas de audio específicas, se obtienen múltiples transcripciones y múltiples probabilidades de selección de palabras, lo que conduce a una transcripción dada. Utilizando estas transcripciones, se puede calcular una medida de variabilidad de las transcripciones, proporcionando una así una medida de incertidumbre del modelo.

Es importante destacar que dado que el modelo predecirá una lista de palabras correspondientes al audio en cada iteración y, dada la arquitectura del decodificador basado en Transformer del modelo, como afirma [6], en cada iteración se pueden obtener transcripciones de longitudes variables. La forma en que se genera la estimación dependerá del procesamiento interno de las redes neuronales. Para generalizar la ecuación (4) para el caso del reconocimiento de voz y obtener una distribución de incertidumbre, se necesita una métrica de distancia no euclidiana d que pueda calcularse entre secuencias de caracteres con longitudes potencialmente diferentes. Como propone [6], la variabilidad en las transcripciones se puede obtener utilizando la distancia de Levenshtein, que es una métrica de distancia simple derivada del número de operaciones de edición requeridas para transformar una cadena de caracteres en otra. Dado que no hay noción de promedio en un espacio no euclidiano, se utilizará el medoide de las T diferentes transcripciones de salida y_T como nuestra "media" $\overline{y}$ de transcripciones, que puede escribirse como:

$$\overline{y} = \underset{y \in \{y_1, \ldots y_T\}}{\operatorname{argmin}} \sum_{t} d(y, y_T)$$

Metodología

Se tomó una muestra de 10 archivos de audio del conjunto de datos CIEMPIESS-UNAM. La muestra fue contaminada con un 5% de ruido de fondo, mientras que el modelo Whisper se configuró con los siguientes parámetros: config.attention_dropout = 0.3, config.activation_ dropout = 0.5, config.dropout = 0.5. Estos parámetros activan el dropout en diferentes capas de la arquitectura. Para cada uno de los archivos de audio, se utilizó el sistema Whisper (versión pequeña) para generar transcripciones, considerando 10 iteraciones por audio. Para cada una de las 10 iteraciones correspondientes a un audio específico, se calculó el medoide de las transcripciones y, posteriormente, se computó la distancia de Levenshtein al

medoide encontrado. La función indicadora se utiliza para generar la probabilidad de obtener una distancia de Levenshtein específica. Esta probabilidad representa una forma de estimar numéricamente la incertidumbre en cuestión.

Para cada una de las transcripciones, se obtiene la tasa de error de palabras (Word Error Rate, WER) al comparar las transcripciones generadas con las transcripciones reales del audio. Esto permite vincular las incertidumbres obtenidas con una métrica comúnmente utilizada en sistemas de reconocimiento de voz.

Resultados

La Figura 1 muestra la dispersión de los datos. En el eje y se representa la distancia de Levenshtein con respecto al medoide (incertidumbre del puntaje) en función del WER; obtenido al comparar cada una de las transcripciones de audio con la verdad fundamental. Se observa un comportamiento lineal, como se esperaba, con un coeficiente de variación de R^2=0.62, lo que indica una correlación relativamente fuerte y positiva entre las dos variables representadas.

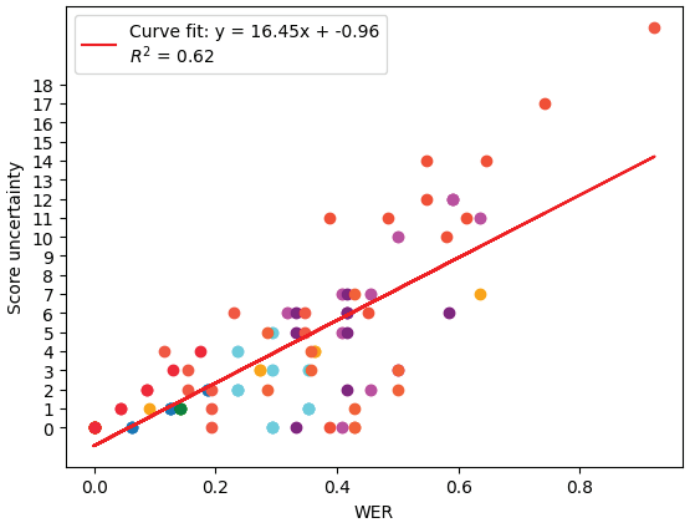

Figura 1. Puntaje de incertidumbre en función del WER.

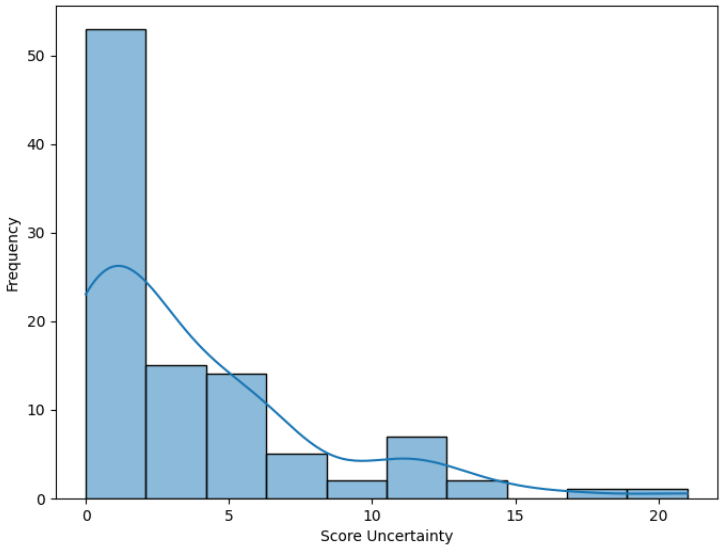

Figura 2. Distribución de frecuencia del puntaje de incertidumbre.

La Figura 2 muestra la distribución de frecuencia de la distancia de Levenshtein con respecto a los medoides de cada muestra de audio de prueba. Se puede observar que el pico del histograma está a la izquierda, lo que indica un bajo número de operaciones requeridas para convertir la cadena de texto de las transcripciones en el medoide correspondiente. La Figura 3 muestra las probabilidades de obtener una incertidumbre de puntuación específica dentro del conjunto obtenido. Se puede observar que la probabilidad más alta es 0.25, lo que corresponde a obtener un valor de distancia cero, lo que significa transcripciones sin variaciones en todas las iteraciones. Debido al comportamiento lineal observado previamente, si las transcripciones no varían en cada una de las iteraciones, se esperaría un WER cercano a cero, es decir, que la baja variabilidad en las predicciones resultaría en las transcripciones coincidiendo con la verdad fundamental.

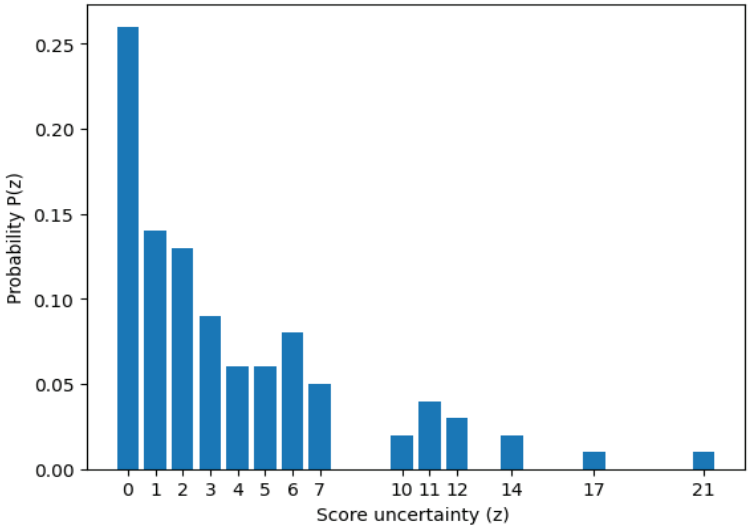

Figura 3. Distribución de frecuencia normalizada.

Conclusiones

Se propuso una técnica para estimar la incertidumbre de un modelo de reconocimiento automático de voz (ASR) basado en la arquitectura de un transformer. Para trabajos futuros, se recomienda aumentar el tamaño de la muestra, introducir variaciones en la configuración de dropout del modelo y utilizar muestras contaminadas con diferentes niveles de ruido, comparándolas con muestras no contaminadas, con el fin de generalizar los resultados mediante pruebas estadísticas e identificar cuantitativamente cómo el ruido afecta el rendimiento del modelo. Además, se sugiere obtener las probabilidades proporcionadas por el modelo para cada token seleccionado aplicando una función softmax a los logits de cada predicción de salida; con el fin de calcular una fiabilidad promedio del modelo y compararla con el WER generado en cada transcripción; si el modelo se comporta correctamente, se esperaría observar una pendiente negativa, lo que indica que un menor WER corresponde a una mayor fiabilidad en la salida generada por el modelo.

Referencias

[1] Díaz, C., Calderon-Ramirez, S., y Aguilar, L. D. M. (2022). Data quality metrics for unlabelled datasets. En 2022 ieee 4th international conference on bioinspired.

[2] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., y Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 .

[3] Mena, J., Pujol, O., y Vitria, J. (2021). A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. ACM Computing Surveys.

[4] Loquercio, A., Segu, M., y Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. IEEE Robotics and Automation Letters, 5 (2), 3153–3160.

[5] Gal, Y., y Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. En international conference on machine learning (pp. 1050–1059)

[6] Jayashankar, T., Roux, J. L., y Moulin, P. (2020). Detecting audio attacks on asr systems with dropout uncertainty. arXiv preprint arXiv:2006.01906.