Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Tipo de artículo: Artículos originales
Temática: Reconocimiento de patrones
Recibido: 30/09/2022 | Aceptado: 17/02/2022 | Publicado: 04/03/2022

Identificación de idioma hablado en señales cortas aplicando transferencia de aprendizaje

Spoken language identification for short utterance with transfer learning

Ana Montalvo Bereau 0000-0002-2230-8079[1*]
Flavio Reyes Díaz 0000-0003-3358-3188[1]
Gabriel Hernández Sierra 0000-0002-2528-3487[1]
José Ramón Calvo de Lara 0000-0002-2186-8314[1]

[1]Centro de Aplicación de Tecnologías de Avanzada. 7ma A #21406 e/ 214 y 216, Playa. C.P. 12200. La Habana, Cuba. {amontalvo, freyes, gsierra, jcalvo}@cenatav.co.cu

[*]Autor para correspondencia: (amontalvo@cenatav.co.cu)

Resumen

En el presente trabajo se abordó el reconocimiento automático del idioma hablado en señales de corta duración, empleando una red neuronal convolucional pre-entrenada sobre un conjunto de imágenes. Partiendo del conocimiento transferido del dominio de imágenes reales a la clasificación de tareas sobre audio, se evaluó el impacto del aprendizaje multitarea tomando el reconocimiento de idioma como tarea principal y el reconocimiento del locutor como tarea auxiliar. Los experimentos se llevaron a cabo sobre un subconjunto del corpus Voxforge, y con una cantidad de señal significativamente menor a las empleadas por sistemas análogos de referencia. La evaluación se realizó sobre espectrogramas conformados con 3 segundos de señal. Los resultados arrojan que el reconocimiento del idioma hablado se beneficia del aprendizaje multitarea al usar como tarea auxiliar la identidad del locutor.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Palabras clave: Reconocimiento automático del idioma hablado; aprendizaje profundo; transferencia de aprendizaje; aprendizaje multitarea.

Abstract

In the present work, spoken language recognition in short utterances was addressed using a convolutional neural network pre-trained on a set of images. Starting from the knowledge transferred from the domain of real images to the audio classification tasks, we assess the impact of multitask learning, taking language recognition as the main task and speaker recognition as auxiliary task. The experiments were carried out on a subset of the Voxforge corpus, and with a significantly lower amount of signals than those used by analog reference systems. The evaluation was done over spectrograms conformed with 3 seconds signal. The results show that the spoken language recognition task benefits from multitasking learning by using the identity of the speaker as an auxiliary task.

Keywords: Spoken language recognition, deep learning, transfer learning, multitask learning

---

Introducción

El habla es la manifestación acústica del lenguaje y probablemente la principal forma de comunicación entre humanos. El desarrollo de las telecomunicaciones y del procesamiento digital de la información ha demandado esfuerzos por comprender los mecanismos de comunicación mediante habla.

En un mundo cada vez más globalizado, donde las fronteras entre países tienden a desaparecer y las necesidades de comunicación entre personas de distintas lenguas se multiplican, el reconocimiento automático del idioma hablado (RAIH) adquiere cada vez más protagonismo. Sistemas telefónicos multilingües que precisan de la identificación del idioma del hablante para su correcto reconocimiento, sistemas de indexación de contenidos multimedia por idioma, enrutamiento de llamadas, sistemas automáticos de traducción de idiomas, entre otros, son los grandes beneficiados de la investigación y avances realizados en este campo durante las últimas dos décadas.

Actualmente los sistemas de RAIH descansan en modelos de aprendizaje profundo, ya sea en la etapa de extracción de características aprendiendo representaciones (Padi et al., 2018), o en arquitecturas de extremo a

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

extremo (E2E de sus siglas en inglés) que agrupan y modelan conjuntamente la extracción de características y la clasificación del sistema (Jin et al., 2017). Hoy las redes neuronales profundas con arquitectura E2E lideran el RAIH, especialmente para señales de corta duración (3 segundos) (Shon et al., 2018).

No obstante los avances alcanzados, algunas de estas herramientas no generalizan[1] bien a nuevos dominios (Abdullah et al., 2020), hay también evidencias de que la eficacia se degrada cuando se introducen nuevos locutores (Montavon, 2009).

El presente estudio propone un método para el RAIH empleando algoritmos reconocidos por sus potencialidades para la generalización, lo cual representa un característica importante cuando se dispone de pocos datos para el entrenamiento y frente a señales de prueba de corta duración.

El método que se propone en este trabajo posee una arquitectura E2E que parte de una red neuronal convolucional (CNN) pre-entrenada sobre un conjunto de imágenes. La transferencia de conocimiento es la primera técnica de generalización que se aplica, ya que los parámetros de la red inicial son entrenados sobre imágenes de objetos reales muy distintas a los espectrogramas[2] que constituyen la entrada al método propuesto y la segunda técnica de generalización que se aplica es la conocida como aprendizaje multitarea, incorporando al RAIH la identidad del locutor como tarea auxiliar.

El reconocimiento automático del idioma hablado

Definimos la tarea del RAIH como un problema de clasificación discriminativa de una secuencia. Primero, una expresión oral de longitud variable es transformada en una secuencia de observaciones acústicas $\mathbf{X} = (x_1, ..., x_T)$, donde $x_t \in \mathbb{R}^k$ es un vector de rasgos acústicos en el instante de tiempo t. Dada una secuencia $\mathbf{X}$, el objetivo es predecir el idioma hablado $\hat{y}$.

Usando una red neuronal profunda como modelo de clasificación, el problema de LID se puede definir como:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{X};\boldsymbol{\theta}) \tag{1}$$

donde $\mathcal{Y}$ es un conjunto finito de idiomas, $\boldsymbol{\theta}$ son los parámetros del modelo y $P(y|\mathbf{X};\boldsymbol{\theta})$ representa la probabi-

[1]Con generalización nos referimos a la capacidad del modelo a mantener su desempeño sobre datos no vistos en el entrenamiento, considerando que ambos conjuntos de datos siguen la misma distribución.

[2]El espectrograma es una representación bidimensional en el dominio tiempo-frecuencia que puede considerarse como imagen.

lidad a posteriori de la etiqueta de idioma *y*.

El aprendizaje no es un proceso fácil, ni para los humanos ni para las máquinas. Es un proceso de trabajo pesado, que consume recursos y tiempo y, por lo tanto, era importante diseñar un método que evitara que un modelo olvidara la curva de aprendizaje que obtuvo de un conjunto de datos específico y también le permitiera aprender más de nuevos y diferentes conjuntos de datos.

La idea inicial detrás de la transferencia de aprendizaje es reutilizar la experiencia o el conocimiento ya obtenido, para mejorar el aprendizaje de cosas nuevas. La transferencia de conocimiento y al aprendizaje multitarea pueden considerarse como implementaciones particulares de la transferencia de aprendizaje aplicadas en diferentes condiciones o de distintas maneras (Bengio, 2012).

Transferencia de conocimiento en modelos convolucionales

Las CNNs han probado ser eficaces reduciendo las variaciones espectrales y modelando correlaciones en rasgos acústicos (Zhang et al., 2017) por lo que se han utilizado en una variedad de tareas de classificación de audio. Hay aproximaciones basadas en CNNs que utilizan como dato de entrada el audio en bruto o apenas preprocesado, valiéndose de convoluciones uni-dimensionales. Sin embargo, la mayoría de los resultados se han obtenido mediante el uso de CNNs en espectrogramas.

La transferencia de conocimiento en tareas de clasificación de audio se ha centrado principalmente en el pre-entrenamiento del modelo sobre corpus de audio significativamente grandes y diversos (Choi et al., 2017) que luego son ajustados a distintas tareas. Un enfoque innovador y diferente fue emplear una CNN entrenada sobre un conjunto masivo de imágenes y adaptarla con espectrogramas al dominio de clasificación en audios.

Trabajos aplicados al RAIH, que empleen modelos pre-entrenados sobre ImageNet (Deng et al., 2009) podrían citarse a Revay and Teschke (2019) y van der Merwe (2020). Estos trabajos presentan propuestas muy similares a la presente investigación: emplean arquitecturas E2E, parten de redes bien establecidas como ResNet y DenseNet y utilizan espectrogramas de señales cortas como entradas a las redes.

Otro ejemplo de modelo para la identificación del idioma desde una perspectiva de visión por computadora es Bartz et al. (2017), que aunque no parte de un modelo previamente entrenado sobre ImageNet, concibe el entrenamiento de una red híbrida CNN recurrente con centenares de miles de espectrogramas.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Aprendizaje multitarea

La mayoría de las técnicas del aprendizaje automático se centran en el aprendizaje de una única tarea aislada, y está claro que este enfoque desestima ciertos aspectos fundamentales del aprendizaje humano. Los seres humanos afrontan cada nueva tarea de aprendizaje equipados con los conocimientos adquiridos en las tareas de aprendizaje anteriores. Además, el aprendizaje humano con frecuencia implica abordar varias tareas de aprendizaje simultáneamente.

La aplicación del aprendizaje multitarea para el RAIH básicamente se ha centrado en relacionar la información fonética con el idioma, ya sea en enfoques E2E (Li et al., 2020) como en la etapa de representación (Zhao et al., 2019). Existen también trabajos donde el reconocimiento del idioma o dialecto es la tarea auxiliar y lo que se busca es relacionarla con los fonemas para mejorar la eficacia del reconocedor de habla (Mendes et al., 2019).

En el caso de tareas correlacionadas negativamente, como el idioma y las diferencia de dominio, se ha aplicado el aprendizaje multitarea adversarial para lograr un modelo que disminuya la dependencia con el dominio (Abdullah et al., 2020). Sobre la línea del aprendizaje adversarial es muy común el empleo de las redes generativas adversariales para el RAIH (Miao et al., 2018).

De acuerdo a Peng et al. (2019), emplear una representación altamente correlacionada con unidades fonéticas e independiente del locutor, favorece el RAIH empleando el enfoque i-vector. Sin embargo de los resultados experimentales del presente trabajo, con un enfoque de E2E y un conjunto cerrado de locutores, asociar el idioma y el locutor parece favorecer el RAIH.

Método propuesto

Arquitectura monotarea

Para inicializar la red fue empleada la red MobileNetV2 (Sandler et al., 2019) haciendo uso de la transferencia de conocimiento. MobileNetV2 fue desarrollada en Google y entrenada sobre el conjunto de datos ImageNet con 1,4 millones de imágenes y 1000 clases de imágenes web. A partir de este modelo pre-entrenado, los parámetros de las capas iniciales son fijados o “congelados” y se definen como entrenables solo las últimas 23 capas de la red.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

El modelo propuesto consiste entonces en la MobileNetV2 de 53 capas de profundidad (ver Fig.1a), de las cuales las 30 primeras permanecen congeladas. Este bloque será referido como pre-entrenado G_f, y puede ser considerado como un extractor de rasgos de alto nivel de abstracción, que transforma la secuencia de entrada $\mathbf{X}$ en un mapa de características $f = G_f(\mathbf{X}; \boldsymbol{\theta}_f)$.

Luego del bloque pre-entrenado siguen las 23 capas del bloque convolucional, que forman parte de la arquitectura de la MobineNetV2 pero que a diferencia de las primeras 30 capas, los parámetros de este bloque $\boldsymbol{\theta}_m$ son aprendidos sobre el conjunto de entrenamiento de la base de datos de idiomas $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_i^N$, siendo N la cantidad de muestras etiquetadas. Vale recalcar que como se puede apreciar en la ecuación 2, en la estimación de $\boldsymbol{\theta}_m$ incide de manera directa el conocimiento codificado en las capas congeladas que es transferido a G_m al recibir como entrada las transformaciones que se suceden en el bloque G_f.

G_m somete a su vez a f a una serie de transformaciones no lineales $\hat{y} = G_m(f; \boldsymbol{\theta}_m)$ que, seguido de una capa softmax, mapea $\hat{y}$ en una distribución de probabilidad sobre el espacio de los idiomas.

La función a minimizar durante el aprendizaje resulta entonces:

$$J(\boldsymbol{\theta}_m) = \sum_{(\mathbf{X}_i, y_i) \in \mathcal{D}} L_y\Big(G_m\Big(G_f(\mathbf{X}_i, \boldsymbol{\theta}_f); \boldsymbol{\theta}_m\Big); y_i\Big); \tag{2}$$

donde L_y es el error en la clasificación de idioma.

Arquitectura multitarea

En esta sección se aborda una segunda técnica de generalización, el aprendizaje multitarea. La aproximación propuesta busca mejorar la clasificación de idiomas, incorporando al aprendizaje la información no semántica contenida en la señal relativa a la identidad de los locutores.

Con este objetivo definimos un bloque de capas que serán compartidas, y que coincide con la red MobileNet $G_{Mobile} = G_m\Big(G_f(\mathbf{X}_i; \boldsymbol{\theta}_f); \boldsymbol{\theta}_m\Big)$. A este bloque le suceden dos ramas de capas completamente conectadas y sus respectivas capas softmax, una para cada tarea específica (ver Fig 1b).

Para el entrenamiento empleamos el mismo conjunto de datos $\mathcal{D} = \Big\{(\mathbf{X}_i, y_i, l_i)\Big\}_i^N$ pero esta vez cada muestra $\mathbf{X}_i$ cuenta con una etiqueta de idioma y del locutor. En esta configuración multitarea la función objetivo a

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

minimizar resulta ser:

$$J(\boldsymbol{\theta}_m, \boldsymbol{\theta}_y, \boldsymbol{\theta}_l) = \sum_{(\mathbf{X}_i, y_i, l_i) \in \mathcal{D}} \lambda_y * L_y\Big(G_y\Big(G_{Mobile}; \boldsymbol{\theta}_y\Big); y_i\Big) + \lambda_l * L_l\Big(G_l\Big(G_{Mobile}; \boldsymbol{\theta}_l\Big); l_i\Big). \quad (3)$$

Fig. 1 - Esquema de las configuraciones propuestas: (a) monotarea y (b) multitarea. Siendo $\hat{y}$ y $\hat{l}$ las variables que representan el idioma y el locutor predicho de cada muestra.

Siendo L_y y L_l los errores de clasificación de idioma y locutor respectivamente; y λ_y y λ_l pesos no negativos, que ponderan el impacto que la tarea en particular tendrá en la estimación de los parámetros del sistema. Un valor de λ cercano a cero significa que esa tarea no tendrá prácticamente influencia durante el entrenamiento del sistema y viceversa.

Configuración experimental

La base de voces empleada fue VoxForge (MacLean, 2009), un corpus de voces de código abierto que contiene muestras de más de 18 idiomas diferentes . Los datos consisten en archivos de audio de duración aproximada entre 5 y 10 segundos, con la transcripción del texto hablado, así como con etiquetas relativas al idioma, sexo e identidad del locutor. La calidad de las diferentes muestras varía según el equipo de grabación utilizado por el colaborador por lo que constituye un corpus con una gran variabilidad de sesión.

Para la experimentación, definimos un sub-conjunto de VoxForge formado por 5 idiomas: Alemán, Español, Francés, Inglés y Ruso. Aproximadamente 38 minutos por idioma permitieron conformar conjuntos de entrenamiento (60 %), validación (25 %) y prueba (15 %), cuidando que existiera un balance de locutores femeninos y masculinos.

Para la obtención de las imágenes (espectrogramas), primeramente se pre-procesaron los audios eliminando silencios iniciales y finales con Librosa (McFee et al., 2015). A continuación, se extrajeron los rasgos acústicos empleando Kaldi (Povey et al., 2011). Se aplicó un filtro de preénfasis a la señal y fue dividida en tramas de 20 milisegundos (con solape de 10ms). Sobre las tramas se aplicó una función de ventana y se computó la transformada de Fourier de corta duración para obtener el espectro de potencia o periodograma. El último paso consistió en aplicar a dicho espectro de potencia un banco de filtros triangulares (40 filtros) en escala Mel, obteniendo una representación por trama de vectores de 40 dimensiones conocidos como rasgos banco de filtros Mel (Mel-fbank).

Concatenando los 300 vectores iniciales de cada señal se conforma finalmente la matriz de 40×300 con la que se crea y salva la imagen en escala de grises, empleando OpenCV para python. De esta forma se utiliza de cada señal solo 3 segundos de audio asegurándose que no contenga silencios, solo pausas.

En la figura 1 se muestran las dos arquitecturas evaluadas. La arquitectura monotarea se presenta en la Fig.1a, formada por la MobileNetV2 y una capa softmax con 5 clases correspondientes a los idiomas a identificar.

La arquitectura multitarea se ilustra en la Fig.1b y está formada por la MobileNetV2 y dos ramas en paralelo encargadas de modelar de forma conjunta la identificación de idioma ($\hat{y}$) y de locutor ($\hat{l}$). Las dos ramas son iguales y compuestas por dos capas completamente conectadas que terminan en una capa softmax.

La incorporación del aprendizaje multitarea aumentó en 1millón la cantidad de parámetros entrenables de la red, resultando en aproximadamente 5.5M de parámetros a entrenar en la configuración más compleja, lo que repercutió en menos de un minuto más de entrenmiento en un core i7-4790 con 8 procesadores a 3.60GHz.

La plataforma de trabajo empleada para las técnicas de aprendizaje profundo fue TensorFlow 2.3.0, una de las más utilizadas a nivel global. TensorFlow puede utilizarse en conjunto con diferentes lenguajes de programación, sin embargo presenta mayor soporte para Python.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Resultados y discusión

Para la selección de algunos hiperparámetros de la red evaluamos su comportamiento sobre el set de entrenamiento y validación. Se hizo un análisis del comportamiento de la eficacia durante el entrenamiento y se observa como sobre el conjunto de validación se mantiene oscilando alrededor de un 80 % a partir de la época 20, valor en el que finalmente entrenamos los modelos. De forma empírica definimos la razón de aprendizaje en 10^{-4} y se aplicó el algoritmo de optimización de Adam.

En la Tabla 1 se presentan los resultados sobre los conjuntos de entrenamiento, validación y prueba de los modelos con ambas arquitecturas monotarea y multitarea (idioma-locutor).

Tabla 1 - Comparación de las arquitecturas monotarea y multitarea.

Conjunto	**Monotarea**		**Multitarea: Idioma y locutor**			
	$Perdida$	$Eficacia$	$Perdida_{idioma}$	$Eficacia_{idioma}$	$Perdida_{locutor}$	$Eficacia_{locutor}$
Entrenamiento	$4 * 10^{-4}$	1	0,0011	0.9970	0,0097	0.9978
Validación	0.0551	0.8090	0.0520	0.8286	0.5673	0.8377
Prueba	0.0677	0.7650	0.0510	0.8342	0.9174	0.7705

Obsérvese en la Tabla 1 que el valor de eficacia sobre el conjunto de prueba en la arquitectura multitarea es mayor que en la monotarea, por lo que se infiere que la incorporación de la información relativa al locutor en la arquitectura E2E que se propone, favoreció la capacidad discriminatoria del modelo en la tarea de RAIH. Obsérvese también cómo la diferencia en eficacia entre la validación y la prueba disminuye al pasar de la arquitectura monotarea a la multitarea, mostrando mayor capacidad de generalización en esta última.

En la Figura 2 se reportan los valores de precisión y especificidad de los modelos monotarea y multitarea (Idioma-Locutor) para cada uno de los 5 idiomas considerados.

La precisión y la especificidad son dos valores que nos indican la capacidad de nuestro estimador para discriminar los casos positivos de los negativos y son métricas que resultan relevantes, y complementarias a la eficacia, en problemas de clasificación con clases desbalanceadas. La precisión cuantifica la cantidad de predicciones positivas que realmente pertenecen a la clase positiva, mientras que la especificidad cuantifica la cantidad de predicciones positivas del total de ejemplos positivos en los datos.

La distribución de las clases en este caso es equilibrada, sin embargo el análisis permite ver para cada idioma las potencialidades y debilidades de los modelos. Por ejemplo de la configuración multitarea se observa que

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Fig. 2 - Precisión y especificidad de los modelos monotarea y multitarea (Idioma-Locutor) sobre el conjunto de prueba.

para el idioma inglés se tiene la mayor precisión, por tanto las muestras en ese idioma tienen gran probabilidad de ser detectadas. Por su parte el alemán es el idioma para el cual el modelo multitarea tiene menos falsos positivos.

De acuerdo a lo reportado en la Figura 2, la configuración multitarea exhibe los mejores valores tanto de precisión como de especificidad, para cuatro de cinco los idiomas.

En correspondencia con los valores de eficacia que aparecen en la Tabla 1, las matrices de confusión de la Figura 3 muestran como en la arquitectura multitarea, la confusión entre las etiquetas reales de los idiomas y las predicciones es menor que en la monotarea.

La Figura 3 permite comparar ambas arquitecturas mostrando relaciones entre los idiomas como por ejemplo la cercanía del francés y el ruso con el español. La tasa más alta de predicciones falsas negativas se produjo cuando las muestras españolas se clasificaron como francesas o rusas, esto se corresponde con el hecho de que en la arquitectura monotarea la más baja especificidad la tiene el idioma francés y en la multitarea el idioma ruso lo deplaza.

Si bien es cierto que para una comparación justa de algoritmos de reconocimiento de patrones lo ideal es que los experimentos se reproduzcan exactamente en las mismas bases de datos y bajo la mayor igualdad de condiciones posibles, en este caso se han encontrado referencias publicadas recientes, de entornos experimentales

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

(a) Monotarea (b) Multitarea

Fig. 3 - Matrices de confusión para el RAIH sobre el conjunto de prueba.

muy similares a los del presente artículo.

Los trabajos que se muestran en la Tabla 2 están organizados por fecha, siendo el último la propuesta que nos ocupa. Todos los trabajos abordan el RAIH con un enfoque E2E partiendo de espectrogramas y reconocen en conjuntos de 6 idiomas máximo.

Bartz et al. (2017) es el trabajo que más datos demanda, ya que no parte de una red pre-entrenada sobre ImageNet, sino que entrena desde cero la arquitectura que propone. Exhibe el mayor valor de eficacia pero lo hace sobre espectrogramas de 10 segundos, lo que representa una disparidad respecto al resto de los trabajos referenciados, sinembargo Revay and Teschke (2019) se le acerca mucho reconociendo sobre espectrogramas con menos de la mitad de la señal. Revay and Teschke (2019) parte de una Resnet50, que al igual que DenseNet121 y MobileNetV2 estan entrenadas sobre ImageNet. Un punto en común importante que tiene Revay and Teschke (2019) con este artículo es que ambos emplean VoxForge, aunque nuestra propuesta con aproximadamente 10 veces menos datos y espectrogramas sobre señales 1 segundo más cortas, obtiene una eficacia cercana. Por su parte van der Merwe (2020) tiene la complejidad de reconocer idiomas sudafricanos, que son acústicamente muy cercanos, lo que explica la baja eficacia respecto al resto de los métodos.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Tabla 2 - Comparación con trabajos similares reportados en la literatura

Referencia	Modelo	Entrada	Datos	Idiomas	Eficacia
Bartz et al. (2017)	E2E CNN-RNN	Espectrograma 10 seg	Youtube News EU (213k img)	6 idiomas	92 %
Revay and Teschke (2019)	E2E ResNet50 (ImageNet)	Espectrograma 4 seg	VoxForge (42k img)	6 idiomas	89 %
van der Merwe (2020)	E2E DenseNet121 (ImageNet)	Espectrograma 3 seg	NCHLT (Idiomas sudafricanos)	6 idiomas	81 %
Propuesta	E2E MobileNetV2 (ImageNet) multitarea	Espectrograma 3 seg	VoxForge (3.8k img)	5 idiomas	83 %

Conclusiones

Las CNNs pre-entrenadas para el reconocimiento de imágenes han sido empleadas anteriormente para transferir conocimiento al dominio de clasificación en audio, sin embargo la propuesta del presente artículo es original porque lo hace aplicando además el enfoque de aprendizaje multitarea para el RAIH, en señales de corta duración y con significativamente menos datos que los sistemas referenciales. Los autores no encontraron trabajos previos que propusieran esta combinación de técnicas de aprendizaje automático para el RAIH.

Los resultados del presente trabajo no solo muestran que la transferencia de aprendizaje del reconocimiento de imágenes al de idioma funciona. Muestran que la representación que obtiene dicha red luego de incorporar información relativa al locutor aporta información útil para el reconocimiento del idioma, o sea que la identidad del locutor incorporada como tarea auxiliar al RAIH en una arquitectura E2E, favorece el reconocimiento de idiomas alcanzando una eficacia del 83 % sobre muestras de 3 segundos de duración.

La aplicación del enfoque multitarea sobre una arquitectura como la MobileNetV2 destaca como una novedad y significó una mejora de la eficacia de aproximadamente un 7 %.

La existencia de trabajos recientes que aborden el problema del RAIH sobre muestras de corta duración empleando imágenes, habla de la pertinencia y actualidad de la investigación en esta área. En particular, alcanzar soluciones competitivas con un volumen de datos pequeño, induce a pensar que no será complejo adaptar el sistema a entornos nuevos y específicos, con poca disponibilidad de datos.

Revista Cubana de Ciencias Informáticas
Vol. 16, No. 1, Mes Enero-Marzo, 2022
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 77-91

Referencias

Badr M. Abdullah, Tania Avgustinova, Bernd Mobius, and Dietrich Klakow. Cross-domain adaptation of spoken language identification for related languages: The curious case of slavic languages, 2020.

Christian Bartz, Tom Herold, Haojin Yang, and Christoph Meinel. Language identification using deep convolutional recurrent neural networks, 2017.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.

Keunwoo Choi, Gyorgy Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Ma Jin, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai. End-to-end language identification using high-order utterance representation with bilinear pooling. 2017.

Zheng Li, Miao Zhao, J. Li, Yiming Zhi, Lin Li, and Q. Hong. The xmuspeech system for the ap19-olr challenge. In *INTERSPEECH*, 2020.

Ken MacLean. Voxforge. http://www.voxforge.org/, 2009. accessed 30/09/2021.

Brian McFee, Colin Raffel, Dawen Liang, D. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. 2015.

Carlos Mendes, Alberto Abad, João Paulo Neto, and Isabel Trancoso. Recognition of Latin American Spanish Using Multi-Task Learning. In *Proc. Interspeech 2019*, pages 2135–2139, 2019. doi: 10.21437/Interspeech.2019-2772.

Xiaoxiao Miao, I. Mcloughlin, Shengyu Yao, and Yonghong Yan. Improved conditional generative adversarial net classification for spoken language recognition. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 98–104, 2018.

Gregoire Montavon. Deep learning for spoken language identification. In *NIPS Workshop on deep learning for speech recognition and related applications*, pages 1–4. Citeseer, 2009.

Bharat Padi, Shreyas Ramoji, Vaishnavi Yeruva, Satish Kumar, and Sriram Ganapathy. The leap language recognition system for lre 2017 challenge - improvements and error analysis. In *Odyssey*, 2018.

Zhiyuan Peng, Siyuan Feng, and Tan Lee. Adversarial multi-task deep features and unsupervised back-end adaptation for language recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5961–5965, 2019. doi: 10.1109/ICASSP.2019.8682303.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society, 2011.

Shauna Revay and Matthew Teschke. Multiclass language identification using deep learning on spectral images of audio signals, 2019.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

Suwon Shon, Ahmed Ali, and James Glass. Convolutional neural networks and language embeddings for end-to-end dialect recognition, 2018.

Ruan van der Merwe. Triplet entropy loss: Improving the generalisation of short speech language identification systems, 2020.

Ying Zhang, Mohammad Pezeshki, Philemon Brakel, Saizheng Zhang, Cesar Laurent Yoshua Bengio, and Aaron Courville. Towards end-to-end speech recognition with deep convolutional neural networks, 2017.

Miao Zhao, Rongjin Li, Shijiang Yan, Zheng Li, Hao Lu, Shipeng Xia, Qingyang Hong, and Lin Li. Phone-aware multi-task learning and length expanding for short-duration language recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 433–437, 2019. doi: 10.1109/APSIPAASC47483.2019.9023014.

Conflicto de interés

Contribuciones de los autores

1. Conceptualización: Ana Rosa Montalvo Bereau
2. Curación de datos: José Ramón Calvo de Lara
3. Análisis formal: Ana Rosa Montalvo Bereau
4. Adquisición de fondos: Gabriel Hernández Sierra
5. Investigación: Ana Rosa Montalvo Bereau
6. Metodología: Ana Rosa Montalvo Bereau y José Ramón Calvo de Lara
7. Administración del proyecto: Ana Rosa Montalvo Bereau
8. Recursos: Ana Rosa Montalvo Bereau
9. Software: Ana Rosa Montalvo Bereau
10. Supervisión: José Ramón Calvo de Lara, Gabriel Hernández Sierra y Flavio Reyes Díaz
11. Validación: Ana Rosa Montalvo Bereau
12. Visualización: Ana Rosa Montalvo Bereau
13. Redacción - borrador original: Ana Rosa Montalvo Bereau
14. Redacción - revisión y edición: Ana Rosa Montalvo Bereau, Flavio Reyes Díaz, José Ramón Calvo de Lara y Gabriel Hernández Sierra

Financiación

La presente investigación es parte de un proyecto financiado por la Empresa DATYS.