TECNIACÚSTICA
CUENCA 2023

MEJORAS EN LA DETECCIÓN DE EVENTOS ACÚSTICOS MEDIANTE LA PONDERACIÓN DE DESEQUILIBRIO DE ACTIVIDAD Y PREDICCIONES DÉBILES BASADAS EN EL MÁXIMO

Carlos M. Castorena[1*]
Maximo Cobos[1]
Francesc J. Ferri[1]

[1]Universitat de Valencia. Dept. Informatica 46100 Burjassot (Valencia) , España

RESUMEN

La detección de eventos de sonido es uno de los temas más relevantes en el procesamiento de audio con redes neuronales artificiales. En este trabajo, se presentan dos estrategias muy simples para mejorar la detección de eventos de audio en el sistema de base propuesto en el reto internacional DCASE (Detection and Classification of Acoustic Scenes and Events) 2022. La primera estrategia propuesta se enfoca en mejorar la generación de predicciones débiles a partir de las predicciones de actividad del sistema. Esto resulta en un cálculo de pérdida más sólido y coherente para ejemplos con etiquetas débiles, lo que a su vez mejora la capacidad del sistema para identificar eventos de audio con mayor precisión.

La segunda estrategia se basa en una pérdida ponderada de entropía cruzada binaria, que tiene en cuenta el desequilibrio resultante de la actividad de eventos en cada lote de entrenamiento. Esto garantiza que el sistema se entrene de manera más equilibrada y pueda capturar de manera efectiva los eventos de audio en diferentes proporciones. La combinación de ambas estrategias ha demostrado resultados interesantes, ya que se lograron mejoras en el rendimiento del sistema sin requerir un esfuerzo considerable en el diseño y entrenamiento del mismo.

ABSTRACT

Sound event detection is one of the most relevant topics in audio processing with artificial neural networks. In this work, two very simple strategies are presented to enhance audio event detection in the baseline system proposed in the 2022 DCASE (Detection and Classification of Acoustic Scenes and Events) international challenge. The first proposed strategy focuses on improving the generation of weak predictions from the system's activity predictions. This results in a more robust and coherent loss computation for examples with weak labels, which in turn enhances the system's ability to identify audio events more accurately.

The second strategy is based on a weighted binary cross-entropy loss, which takes into account the imbalance resulting from event activity in each training batch. This ensures that the system is trained in a more balanced manner and can effectively capture audio events in different proportions. The combination of both strategies has demonstrated interesting results, as improvements in system performance were achieved without requiring considerable effort in its design and training.

Palabras Clave— detección de eventos de sonido, desbalance, error ponderado de entropía cruzada.

1. INTRODUCCIÓN

La tarea de detectar eventos de sonido en entornos domésticos del desafío "Detection and Classification of Acoustic Scenes and Events 2022 (DCASE2022)" consiste en diseñar sistemas que no solo predigan la presencia o ausencia de los 10 eventos domésticos considerados, sino que también proporcionen sus límites temporales. Para el entrenamiento se cuenta con 10,000 audios sintéticos con una duración de 10 segundos cada uno, que tienen su respectiva etiqueta con marcas de tiempo (etiquetado fuerte), 1,578 audios reales que solo tienen su etiqueta débil (solo se identifica la presencia del sonido) y, además, se cuenta con 14,412 muestras de audio sin etiquetar [1].

La metodología base para el desafío, consta en una red neuronal convolucional recurrente (CRNN, por sus siglas en ingles), donde varias capas convolucionales funcionan como extractores de características y las capas recurrentes analizan los mapas de características de salida de manera secuencial, adicionalmente una serie de capas densas realizan las funciones de clasificación. El modelo toma

* ***Autor de contacto****: carlos.castorena@uv.es*

como entrada el espectrograma de Mel normalizado de segmentos de audios de 10 segundos, durante el entrenamiento, 12 muestras con etiquetado fuerte, 12 muestras con etiquetado débil y 24 muestras sin etiquetar conforman un lote de entrenamiento. El error total a optimizar se compone por tres tipos de errores: el error fuerte y débil, que se calcula con la entropía cruzada binaria de las muestras del lote que cuenten con su respectivo etiquetado ya sea fuerte o débil, y el error de consistencia que se calcula con un error cuadrático medio de las muestras sin etiquetas. La salida de este modelo son las predicciones fuertes, una matriz que contiene la información de las activaciones en distintos fotogramas y las predicciones débiles, que se representan en forma de vector e indica la presencia o ausencia de un evento [2].

Es común que haya una diferencia significativa de muestras de distintas clases, esto se le conoce como desbalance, en términos del problema de detección de eventos de audio podemos encontrar dos tipos, el primero a nivel evento, referente a la diferencia que hay de muestras entre los distintos eventos, por ejemplo, en los datos hay mas audios que presentan el evento “Perro” que el evento “Aspiradora”. En segundo lugar a nivel fotograma, cuando el número de fotogramas con activación y aquellos que no tienen ningún evento es considerablemente diferente, esto pasa cuando los eventos son demasiado cortos respecto a los 10 segundos de la duración total [3]. Este problema de desequilibrio, al igual que en otros contextos, suele perjudicar el rendimiento de la red neuronal durante su entrenamiento [3]. El uso de funciones de pérdida que contrarresten los efectos negativos de los problemas de desequilibrio de actividad suele llevar a resultados generales mejores [4].

2. METODOLOGÍA

En este trabajo, se proponen dos modificaciones principales al modelo base. La primera se refiere a la forma en que se calculan las predicciones débiles, mientras que la segunda tiene como objetivo incluir mecanismos de pérdida ponderada para reducir los efectos del desequilibrio entre fotogramas activos y no activos. Estas modificaciones se describen a continuación.

2.1. Predicción débil

La salida del modelo base corresponde a la secuencia temporal de probabilidades de actividad de clase, donde para cada fotograma de tiempo se proporciona una probabilidad de actividad para cada clase, a lo que nos referiremos como predicciones fuertes. Sin embargo, es importante notar que el modelo utiliza ejemplos con etiquetas tanto fuertes como débiles para actualizar los pesos

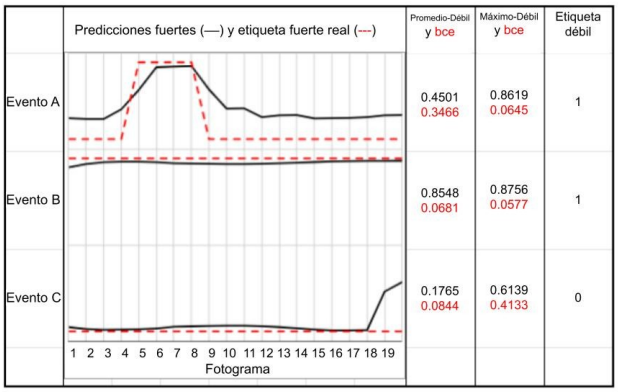

Figura 1. Representación de las predicciones débiles

durante el entrenamiento. Para ejemplos de audio con etiquetas fuertes, está disponible una secuencia de actividad de clase fotograma a fotograma, que se puede comparar directamente con la salida del modelo. Para ejemplos con etiquetas débiles, solo está disponible una etiqueta binaria que indica la presencia de un evento de clase a lo largo de toda la duración del ejemplo. Para tales ejemplos, se genera una predicción débil a partir de la salida del modelo fotograma por fotograma, tomando el promedio de todos los fotogramas. Es importante destacar que una etiqueta débil también se puede obtener a partir de ejemplos con etiquetas fuertes generando una etiqueta de "1" siempre que haya algún fotograma con actividad para cualquiera de las clases. Por último, los ejemplos sin etiquetar se tratan siguiendo el enfoque semisupervisado estudiante-profesor del modelo base.

Una dirección de mejora se refiere a la forma en que se tratan las predicciones débiles en el modelo base. Actualmente, cuando no se utiliza la capa de atención, la predicción débil se calcula a partir del promedio de las predicciones fuertes (*Promedio-Débil*), de modo que la función de costo de entropía cruzada binaria (*bce*) empujará las predicciones de todos los fotogramas del mismo evento hacia "1", contradiciendo la salida natural del modelo fotograma a fotograma. Por esta razón, se propone usar el valor máximo de las predicciones fuertes como predicción débil (*Máximo-Débil*), reemplazando el promedio. De esta manera, se espera que las predicciones débiles correspondientes sean "1" cuando haya al menos un fotograma activo para una clase de evento determinada. La Figura 1 muestra la diferencia entre ambos cálculos. El evento A está activo solo en un pequeño número de fotogramas, lo que lleva a una predicción *Promedio-Débil* pequeña y una pérdida *bce* alta, mientras que la predicción *Máximo-Débil* refleja mejor el hecho de que el evento está presente, como se deduce de la predicción fuerte. Cuando un evento está activo en todos los fotogramas (Evento B), casi

no hay diferencia entre *Promedio-Débil* y *Máximo-Débil*. Finalmente, el Evento C no está activo, pero la predicción fuerte indica incorrectamente actividad de clase, lo que se traduce adecuadamente en un valor significativo de pérdida cuando se usa *Máximo-Débil*, pero no tanto para *Promedio-Débil*.

2.2. Entropía cruzada binaria ponderada

Para balancear los datos, se utiliza una entropía cruzada binaria ponderada (*bce* por sus siglas en inglés), aplicando un factor *w* que pondera el error de manera diferente y proporcional según la etiqueta. La función de pérdida se define como:

$$wbce(x,y)=\frac{1}{s}\sum_{s=1}^{S}(1-w)\cdot y_s\cdot\log(x_s) + w\cdot(1-y_s)\cdot\log(1-x_s), \quad (1)$$

donde *S* es el número de muestras en el lote (batch), y_s y x_s son las etiquetas y las predicciones correspondientes para la muestra *s*, respectivamente. Es importante destacar que tanto x_s como y_s son vectores en R^T cuando se usan etiquetas débiles, o matrices en R^{TxK} cuando se tratan etiquetas fuertes, donde *K* es el número de clases de eventos y *T* es el número de fotogramas temporales considerados. El peso *w* se elige como la proporción de eventos activos en el lote con respecto al número total de eventos en ese lote. Cuando y_s es una etiqueta débil, *w* se calcula de la siguiente manera:

$$w=\frac{\sum_{s=1}^{S}\left(\frac{\sum_{k=1}^{K}y_s^k}{K}\right)+1}{S+1}. \quad (2)$$

Es importante notar que *w* representa la densidad de eventos en el lote: cuando *w* está cerca de *0*, esto significa que hay pocas muestras en el lote con eventos activados, mientras que cuando está cerca de *1*, casi cada muestra tiene todos los eventos activos simultáneamente. Por otro lado, al calcular la pérdida para una etiqueta fuerte, utilizamos:

$$w=\frac{\sum_{s=1}^{S}\left(\frac{\sum_{k=1}^{K}\sum_{t=1}^{T}y_s^{(t,k)}}{T\cdot K}\right)+1}{S+1}, \quad (3)$$

donde en este caso *w* representa la proporción de fotogramas activados en el lote. Es relevante notar que tanto para etiquetas fuertes como débiles, un y_s completamente compuesto de ceros resulta en *w=0*, lo que produce una pérdida nula independientemente de la predicción x_s. Para abordar esta situación, se suma uno al numerador y al denominador.

2.3. Configuraciones del sistema

La codificacion de los sistemas se componen por las estrategias aplicadas durante el entrenamiento, en primer lugar, cuando se aplica *wbce* se incluye la palabra "Balanceado" unido con un gion al tipo de etiqueta (Fuerte o Débil) indicando que se aplica el error ponderado, por ejemplo, si se ha aplicado *wbce* al error fuerte, el nombre incluira *"Fuerte-Balanceado"*, de lo contrario, solo aparecera *"Fuerte"* que indica que no hay modificacion a este tipo de error. En segundo lugar, hay dos posibilidades para la prediccion debil, *"Promedio-Débil"* y *"Maximo-Débil"*, la primera correspondiente al baseline y la segunda como una propuesta de mejora, un ejemplo de esta codificacion es: *"Máximo-Débil-Balanceado"* que indica que es utilizada la estrategia basada en el máximo para el calculo de la etiqueta debil y al mismo tiempo el calculo del error debil es ponderado. Finalmente en aquellas estrategias en las que es aplicada una capa de atencion se indica con la palabra "Atencion". Se pruebas diferentes combinaciones generando un total de 10 experimentos aplicando las estrategias propuestas, mas dos sistemas de referencia.

3. RESULTADOS

La Figura 2 muestra los resultados obtenidos para los diferentes sistemas evaluados. En el eje *x* se presenta la puntuación de Detección de Eventos Sonoros Polifónicos [5] (*PSDS*, por sus siglas en inglés) para el escenario 1, que se enfoca principalmente en cuán rápido se detecta una activación. En el eje *y*, se muestra la *PSDS* para el escenario 2, que mide la confusión entre eventos. Entre los métodos propuestos, los mejores resultados en cuanto a la métrica PSDS1 se obtienen para los sistemas que utilizan *Máximo-Débil*, ocupando las primeras 5 posiciones. Sin embargo, en cuanto a la PSDS2, su rendimiento disminuye considerablemente, excepto cuando se equilibran tanto las predicciones débiles como las fuertes ("Fuerte+*Máximo-Débil-Balanceado"*).

Los sistemas presentados en la Competencia por obtener los mejores resultados se indican en azul. Los 3 métodos se basan en el balanceo de actividad propuesto ya sea en el error fuerte, débil o una combinación de ambos. El sistema "Fuerte+Promedio-Débil-Balanceado" presenta un ligero aumento en la métrica PSDS-escenario2 con respecto al modelo base. El sistema "Fuerte+Máximo-Débil-Balanceado" tiene un buen rendimiento en comparación con el modelo base con capa de atención en la métrica PSDS-

escenario1 y, finalmente, "Fuerte-Balanceado+Máximo-Débil-Balanceado" muestra un resultado competitivo para ambas métricas.

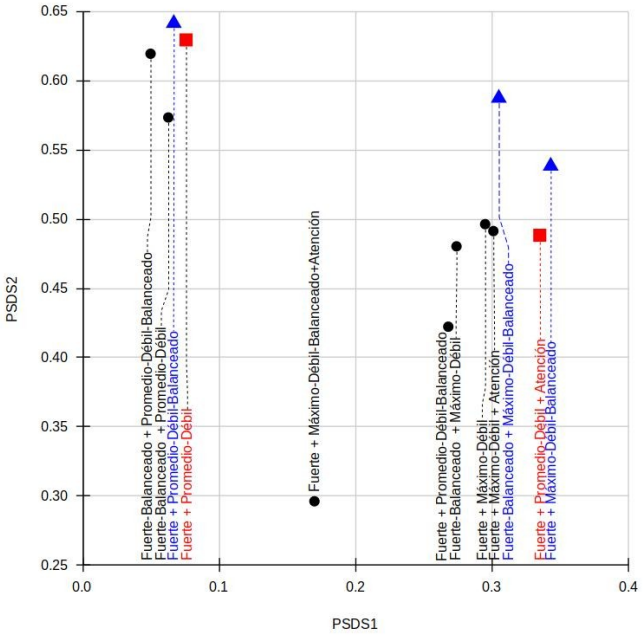

Figura 2. Resultados de PSDS para las estrategias propuestas y metodologías base.

Cuando se utiliza el ponderado por desequilibrio, los resultados son mejores solo cuando no se incluye la capa de atención. Si se incluye la atención, las estrategias propuestas no son realmente efectivas, como se puede observar en los sistemas "Fuerte+Máximo-Débil-Balanceado+Atención" o "Fuerte+Promedio-Débil-Balanceado+Atención"

4. CONCLUSIONES

Se han propuesto dos cambios motivados por la actividad para el modelo base en la Tarea 4 de DCASE2022. Las estrategias propuestas, a pesar de ser muy simples, condujeron a una ligera mejora en el rendimiento del sistema base. Los resultados muestran que las estrategias de balanceo pueden llevar a ciertas ganancias de rendimiento con poco esfuerzo dentro de la etapa de entrenamiento del modelo.

5. REFERENCIAS

[1] R. Serizel, N. Turpault, A. Shah, and J. Salamon, “Sound event detection in synthetic domestic environments,” in ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain, 2020.

[2] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in Workshop on Detection and Classification of Acoustic Scenes and Events, New York City, United States, October 2019.

[3] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, “Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance,” Applied Acoustics, vol. 196, p. 108882, 2022.

[4] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, “Addressing imbalance in multi-label classification using weighted cross entropy loss function,” in 2020 27th National and 5th International Iranian Conference on Biomedical Engi- neering (ICBME), 2020, pp. 333–338.

[5] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, p. 162, 2016