Desarrollo de un sistema de ecualización automática a partir de un modelo de incrustación de palabras

Juan David Rengifo Mera

11 de septiembre de 2024

Resumen

En este trabajo se desarrolla una herramienta de ecualización automática basada en descripciones del sonido mediante el uso de modelos de aprendizaje profundo y procesamiento del lenguaje natural. El objetivo es hacer accesible para cualquiera ecualizaciones de audio que respondan a sus deseos creativos, utilizando modelos de incrustación de palabras como GloVe, Tok2Vec, GPT y BERT, y combinándolos en un modelo de ensamble con una capa de atención.

Palabras clave: Ecualización Automática, Aprendizaje Profundo, Procesamiento del Lenguaje Natural, Modelos de Incrustación de Palabras, Capa de Atención.

1. Introducción

La ecualización de audio es un proceso crítico en la producción musical que requiere un balance adecuado de las frecuencias de audio. Este proceso puede ser complejo y demandar conocimientos técnicos avanzados. La formulación del problema radica en cómo implementar una herramienta de ecualización automática que utilice descripciones en lenguaje natural para ajustar parámetros de audio de forma precisa y eficiente [9].

Revisamos distintos enfoques en la literatura sobre ecualización de audio, destacando las técnicas tradicionales y modernas. Sin embargo, pocos estudios han explorado el uso de modelos avanzados de procesamiento de lenguaje natural como GPT y BERT en este contexto.

El objetivo de este estudio es desarrollar una herramienta de ecualización automática basada en descripciones del sonido utilizando modelos de aprendizaje profundo y procesamiento del lenguaje natural [11].

2. Fundamentación Teórica

2.1. Sonido

El sonido es un tipo de energía que se propaga en forma de ondas a través de medios como el aire, el agua o los sólidos. La frecuencia determina el tono y la amplitud de un sonido [14].

2.2. Ganancia de audio

La ganancia de audio se refiere a la relación entre la amplitud de la señal de entrada y la amplitud de la señal de salida en un sistema de amplificación de audio [21].

2.3. Filtro de audio

Un filtro de audio modifica las características de frecuencia de una señal de sonido. Los tipos de filtros incluyen: paso bajo (permite frecuencias bajas), paso alto (permite frecuencias altas), de banda (selecciona un rango de frecuencias), de rechazo de banda (atenua una gama estrecha de frecuencias), y de estante (aumenta o reduce frecuencias por encima o por debajo de una frecuencia de corte) [26].

2.4. Ecualización

La ecualización de audio ajusta la respuesta de frecuencia mediante el aumento o reducción selectiva de ciertas frecuencias [26].

2.5. Ecualización paramétrica

La ecualización paramétrica permite ajustes precisos y flexibles de la respuesta en frecuencia de una señal de audio [13].

2.6. Modelos de Incrustación de Palabras

Los modelos de incrustación de palabras en NLP representan las palabras como vectores densos y de baja dimensionalidad en un espacio vectorial continuo. Ejemplos de estos modelos incluyen Word2Vec [40], GloVe [39], Tok2Vec [41], y BERT [37]. Estos modelos capturan relaciones semánticas y sintácticas entre palabras, facilitando tareas de NLP como análisis de sentimiento, traducción de idiomas y clasificación de texto [11].

2.7. Aprendizaje de Máquina en Ecualización Automática

Las Redes Neuronales Profundas (DNNs) están compuestas por múltiples capas de nodos interconectados que modelan relaciones complejas en los datos [5]. El Algoritmo de Descenso de Gradiente se utiliza para minimizar una función de pérdida iterativamente [38]. La Regresión Logística predice resultados binarios en un conjunto de datos, usada comúnmente en tareas de clasificación [7].

3. Metodología

3.1. Dataset: Social EQ

El conjunto de datos SocialEQ agrupa descriptores semánticos asociados a configuraciones de ecualización en el ámbito del audio. Cada registro abarca un descriptor semántico, el idioma del mismo, identificador de audio, una puntuación de coherencia y valores correspondientes a 40 parámetros de ecualización. Los participantes seleccionaban y calificaban 40 variaciones de un archivo de audio basado en términos descriptivos. Se restringió la investigación a descriptores en inglés, reduciendo el número a 918 ejemplos, representando 388 descriptores únicos [9].

3.2. División de Datos para Entrenamiento y Pruebas

Se empleó una validación cruzada de cuatro pliegues para asegurar que las palabras en el conjunto de prueba fueran completamente desconocidas durante el entrenamiento. Se seleccionaron palabras de Alta Calidad (HQ) de la literatura de mezcla de audio y palabras Altamente Calificadas (HR) con alta coherencia en el conjunto de datos SocialEQ. Cada pliegue de validación contenía una mezcla de palabras HQ y HR, distribuidas para evaluar al menos una vez cada palabra [9].

Además, se implementó la función `train_test_split` de `scikit-learn` para dividir los datos (80 % entrenamiento, 20 % pruebas), maximizando el uso de los datos disponibles [11].

3.3. Modelos de Incrustación de Palabras

Se integraron cuatro modelos de incrustación de palabras: GloVe [39], Tok2Vec [41], GPT [35] y BERT [37]. Estos modelos transforman las palabras en vectores semánticos, estandarizando la entrada para la red neuronal [11].

3.4. Arquitectura de la Red Neuronal

La red neuronal modela incrustaciones de palabras en predicciones de parámetros de ecualización. La arquitectura incluye capas densas con activación ReLU y regularización con dropout. La capa de salida tiene 40 neuronas con activación sigmoidea para predecir los parámetros de ecualización. Se utilizó una función de pérdida de error absoluto medio y un optimizador SGD con una tasa de aprendizaje inicial de 0.1, que se ajustó cada 10,000 actualizaciones [11].

Se incluyó una capa de atención para cada tipo de incrustación de palabras, mejorando el modelo al centrarse en partes relevantes de las incrustaciones al realizar predicciones [35].

3.5. Normalización de los Parámetros de Ecualización

Los valores de ecualización se normalizaron linealmente en un rango de 0 a 1 (dB en [-4, +4]), para evitar problemas asociados con valores atípicos. Esta normalización estandariza la entrada para la red [11].

4. Resultados

4.1. Modelo Tok2Vec

El modelo Tok2Vec implementa una red neuronal que utiliza embeddings tokenizados específicos para el análisis de texto. La Figura 1 muestra el error en el entrenamiento y la validación del modelo Tok2Vec a lo largo de las épocas.

La gráfica de pérdida (Figura 1) sugiere que el error de entrenamiento disminuye significativamente al principio, indicando una fase de aprendizaje eficiente. Sin embargo, el error de validación comienza a estabilizarse y luego aumenta ligeramente después de cierto punto, lo que podría indicar que el modelo está experimentando sobreajuste [3].

4.2. Modelo GloVe

En este experimento, utilizamos únicamente los embeddings de GloVe para entrenar el modelo. La Figura 1 muestra el error en el entrenamiento y la validación del modelo GloVe a lo largo de las épocas.

La gráfica de pérdida (Figura 1) ilustra que el error de entrenamiento disminuye significativamente al inicio, lo que indica que el modelo está aprendiendo efectivamente. Sin embargo, el error de validación se estabiliza y comienza a exhibir fluctuaciones, lo cual puede indicar la presencia de sobreajuste [39].

4.3. Modelo BERT

En este experimento, utilizamos los embeddings proporcionados por BERT para entrenar el modelo. La Figura 1 muestra el error en el entrenamiento y la validación del modelo BERT a lo largo de las épocas.

La gráfica de pérdida (Figura 1) indica que el error de entrenamiento disminuye considerablemente al inicio, lo que sugiere una etapa de aprendizaje efectiva. Sin embargo, el error de validación se estabiliza y presenta fluctuaciones, lo que puede indicar sobreajuste [37].

4.4. Modelo GPT-4

En este experimento, utilizamos los embeddings proporcionados por GPT-4 para entrenar el modelo. La Figura 1 muestra el error en el entrenamiento y la validación del modelo GPT-4 a lo largo de las épocas.

La gráfica de pérdida (Figura 1) indica que el error de entrenamiento disminuye considerablemente al inicio, lo que sugiere una etapa de aprendizaje eficiente. Sin embargo, el error de validación muestra un incremento significativo después de cierto punto, lo cual es una clara indicación de sobreajuste [35].

4.5. Modelo de Ensamble

En este experimento, utilizamos un modelo de ensamble que combina varios embeddings para mejorar las representaciones y, en última instancia, el rendimiento del modelo. La Figura 1 muestra el error en el entrenamiento y la validación del modelo de ensamble a lo largo de las épocas.

La gráfica de pérdida (Figura 1) sugiere que el error de entrenamiento disminuye significativamente al inicio, indicando que el modelo está aprendiendo efectivamente. Sin embargo, el error de validación se estabiliza y fluctúa, lo cual podría indicar la presencia de sobreajuste [35].

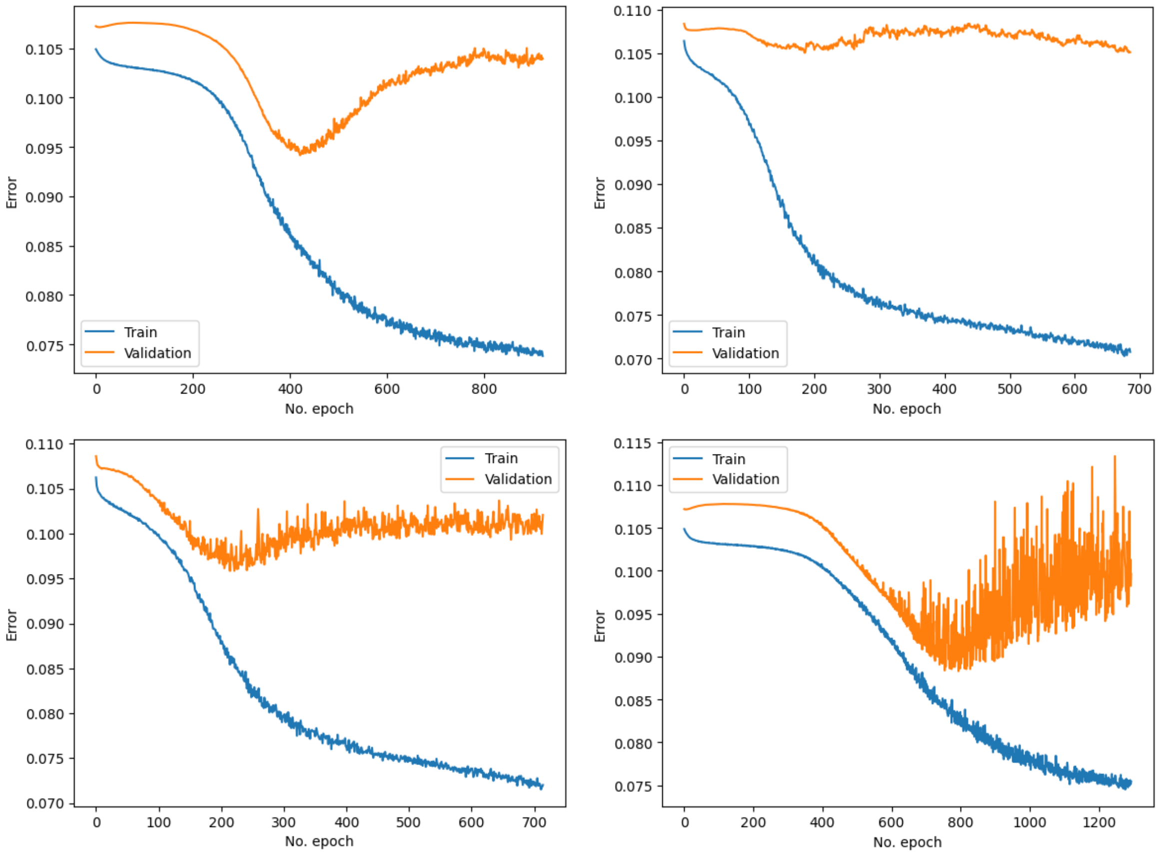

Figura 1: Desempeño de error de entrenamiento y validación para los modelos Tok2Vec, GloVe, BERT y GPT-4 nombrados de izquierda a derecha de arriba hacia abajo respectivamente.

Modelo	MAE Entren.	MAE Valid.	MAPE Valid.
Ensamble	0.0684	0.1013	23.9113 %
GPT-4	0.0918	0.0963	21.8592 %
Tok2Vec	0.0768	0.1012	21.5715 %
GloVe	0.0722	0.1066	24.3336 %
BERT	0.0739	0.1009	21.2371 %
Capa de Atención	0.0693	0.0976	21.6324 %

Cuadro 1: Comparación de métricas clave entre diferentes modelos evaluados.

5. Feedback de Uso

Para evaluar el desempeño y la experiencia del usuario con el sistema de ecualización automática, se realizó una prueba con un productor musical de la ciudad de Cali, Xavier Martínez. Xavier destacó la utilidad creativa del sistema, pero señaló la necesidad de una interfaz gráfica y permitir descripciones más ricas para mejorar la precisión.

6. Referencias

1 . Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: https://doi.org/10.1109/tpami.2013.50.

2 . Z. Xie and Y. Li, “Large-scale support vector regression with budgeted stochastic gradient descent,” *In-*

ternational Journal of Machine Learning and Cybernetics*, vol. 10, no. 6, pp. 1529–1541, Jun. 2018, doi: https://doi.org/10.1007/s13042-018-0832-7.

3 . G. Hinton et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: https://doi.org/10.1109/msp.2012.2205597.

4 . A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2012, doi: https://doi.org/10.1145/3065386.

5 . Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: https://doi.org/10.1038/nature14539.

6 . D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: https://doi.org/10.1038/323533a0.

7 . D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.

8 . T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

9 . M. B. Cartwright, B. Pardo, “Social-EQ: Crowdsourcing an Equalization Descriptor Map,” presented at the 14th International Society for Music Information Retrieval (ISMIR) Conference, pp. 395–400, Nov. 2013.

10 . Don and C. Davis, “Sound system equalization,” *Audio Engineering Explained-for professional audio recording*, pp. 525–551, 2010.

11 . S. Venkatesh, D. Moffat, and E. R. Miranda, “Word Embeddings for Automatic Equalization in Audio Mixing,” *Journal of the Audio Engineering Society*, vol. 70, no. 9, pp. 753–763, Nov. 2022, doi: https://doi.org/10.1774/3/jaes.2022.0047.

12 . M. A. Martínez Ramírez, D. Stoller, and D. Moffat, “A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net,” *Journal of the Audio Engineering Society*, vol. 69, no. 3, pp. 142–151, Mar. 2021, doi: https://doi.org/10.17743/jaes.2020.0031.

13 . B. De Man, and J. O. D. Reiss, “A Knowledge-Engineered Autonomous Mixing System,” presented at the 135th Audio Engineering Society Convention, paper 8961, Oct. 2013.

14 . T. D. Rossing, *Science of Percussion Instruments*, World Scientific Publishing Co. Pte. Ltd., 2002.

15 . F. A. Everest and K. C. Pohlmann, *The Master Handbook of Acoustics*, McGraw Hill Professional, 2009.

16 . P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A K Peters, Ltd., 2002.

17 . D. M. Howard and J. Angus, *Acoustics and Psychoacoustics*, Focal Press, 2009.

18 . J. Eargle, *The Microphone Book*, Focal Press, 2004.

19 . G. Ballou, *Handbook for Sound Engineers*, Focal Press, 2008.

20 . F. Rumsey and T. McCormick, *Sound and Recording*, Focal Press, 2014.

21 . G. Eberle, *Audio Engineering Explained*, Focal Press, 2011.

22 . A. M. Noll, *Introduction to Telecommunications Electronics*, Artech House, 2003.

23 . R. F. Lyon, *Human and Machine Hearing: Extracting Meaning from Sound*, Cambridge University Press, 2017.

24 . F. Rumsey, *Desktop Audio Technology: Digital Audio and MIDI Principles*, Focal Press, 2014.

25 . J. Watkinson, *The Art of Sound Reproduction*, Focal Press, 1998.

26 . B. Katz, *Mastering Audio: The Art and the Science*, Focal Press, 2007.

27 . R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*, Focal Press, 2008.

28 . F. E. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*, Focal Press, 2008.

29 . A. Nisbett, *The Sound Studio: Audio Techniques for Radio, Television, Film and Recording*, Focal Press, 2013.

30 . B. Benson, *Audio Engineering Handbook*, McGraw-Hill, 2006.

31 . K. C. Pohlmann, *Principles of Digital Audio*, McGraw-Hill Education, 2015.

32 . J. Dunn, *The Art of Digital Audio*, Focal Press, 2000.

33 . D. M. Huber, R. E. Runstein, *Modern Recording Techniques*, Focal Press, 2005.

34 . I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

35 . A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

36 . A. Joulin et al., "FastText.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2017.

37 . J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

38 . S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

39 . J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," 2014.
Available: https://nlp.stanford.edu/pubs/glove.pdf

40 . T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.
Available: https://arxiv.org/pdf/1301.3781

41 . M. Honnibal and I. Montani, "spaCy: Industrial-strength Natural Language Processing in Python." Available: https://spacy.io/