TRANSFER LEARNING WITH HUBERT AUDIO CLASSIFICATION FOR SARCASM AND AUDIO RECOGNITION

Vincent Lin[1,*]

[1]MIT

linv@mit.edu

ABSTRACT

Sarcasm recognition, for both humans and machines, is a subtle practice that involves many different signals across different modalities of input. One of the most important of these signals, however, is the particular patterns in pitch variation and prosody characteristic of delivering sarcastic statements; emotion recognition similarly shares these characteristic signals patterns. In this work, we present a transfer learning paradigm for audio classification with HuBERT that investigates whether the learned audio embeddings between sarcasm and emotion recognition are merely adjacent or truly interrelated and positively correlated in performance.

Index Terms— Sarcasm recognition, emotion recognition, transfer learning, audio classification

1. INTRODUCTION

Sarcasm recognition is a quite subtle art in human interaction. Whether in real conversation or in entertainment, in speech or in text, identifying sarcasm is an inherently multimodal task that involves many signals across multiple senses to fully encapsulate. While modeling multiple of these modalities computationally may be beneficial to sarcasm recognition in isolation, analysis of the characteristics of the speech audio alone as one of the largest indicators of sarcastic speech is in itself extremely valuable. Furthermore, we note that the task of emotion recognition is similarly closely related to analyzing speech characteristics such as pitch variation and the prosody in speech delivery. Given the similarities between the sarcasm and emotion recognition tasks, does effectively learning to perform well in one task influence or boost the performance in the other task? In the context of machine learning instead of human learning, can we demonstrate this inter-related latent representation in neural networks and audio encoders such as the Hidden-Unit BERT (HuBERT) [1] transformer audio encoder? The motivation behind this project is to discover the extent to which neural networks can detect sarcasm, and particularly whether emotion detection is not only an adjacent task, but one that can boost performance in the sarcasm detection task.

1.1. Related Works

Many previous studies in the field of sarcasm and emotion recognition make significant progress toward detecting each independently, but not toward integrating the two tasks. Much of initial sarcasm detection research begins by investigating the pitch variation, stress, and other prosodic aspects of speech as an improvement over text-based sarcasm recognition [2]. Much of emotion recognition follows a similar route in applying CNNs to analyzing pitch frequency in emotional speech [3]. More recently, work has been completed that investigates incorporating multiple modalities of data (e.g., audio + video); works in multimodal sarcasm recognition use transformer models to develop audiovisual encoders [4]; works in multimodal emotion recognition similarly use transformers and cross attention networks to incorporate information from multiple data modalities [5]. However, all of the aforementioned works have investigated the sarcasm and emotion tasks separately, and little recent work addresses the potential synergistic connection between the two when trained jointly.

2. METHOD

2.1. Data

MUStARD: The Multimodal Sarcasm Detection Dataset (MUStARD) [6] contains a collection of clips compiled from popular TV shows, each human-annotated with a binary sarcasm label. Each clip contains the audiovisual data from the show, as well as information about the main speaker and textual context surrounding the clip. In this paper, we only consider the raw audio data associated with each clip and its respective sarcasm label. Of the two datasets used in this work, MUStARD is the smaller of the two at 690 total samples, so we use all audio files for sarcasm speech modeling.

RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7] is a large multimodal database of emotional speech and song. The database contains audio clips from 24 professional actors (12 male, 12 female), ranging 8 human-annotated emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised), with 2

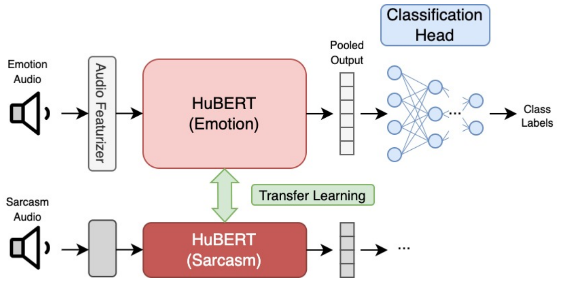

Fig. 1. Architecture for HuBERT audio classification fine-tuning and transfer learning. The architecture consists of an audio featurizer, the HuBERT audio encoder, and a linear classification head. Transfer learning is performed by freezing the weights of the fine-tuned HuBERT and fine-tuning it with a new classification head for the respective new task.

unique sentences uttered in a clean and isolated environment. We use a subset of RAVDESS that is both comparable in size to our sarcasm dataset and still equally representative of all genders, emotions, intensities, etc. of the original RAVDESS database; specifically, we use (24 actors * 8 emotions * 2 statements * 2 intensities - 48) for 720 total samples, where the subtracted quantity is due to the neutral emotion only having one intensity.

2.2. Modeling Approach

The general structure of our modeling approach is as follows: fine-tune HuBERT with a classification head in one task, then freeze the fine-tuned HuBERT weights and fine-tune a new classification head to test its performance on the other task. This is with the intention that after fine-tuning HuBERT on either sarcasm or emotion data, we will be able to evaluate the fine-tuned HuBERT's ability to complete both tasks and determine whether there truly is a relationship between the two tasks via the results of transfer learning.

2.3. Model Architecture

The full architecture of our HuBERT audio classification model is illustrated in Figure 1. For a single instance of the HuBERT classifier, the model consists of the following components:

Audio Featurizer: We input audio clips, at a sampling rate of 16 kHz, into the Wav2vec2 [8] audio featurizer to produce a series of input value sequences and corresponding transformer attention masks to feed into the HuBERT encoder.

HuBERT Encoder: The Hidden-Unit BERT (HuBERT) BERT-based transformer encoder [1] that uses a self-supervised speech representation learning to learn and generate tokenized embeddings from the input audio data. Here, we use HuBERT as a base for a strong foundational audio representation and a solid starting point for fine-tuning our audio representation for emotion and sarcasm recognition tasks. To generate embeddings for entire audio sequences, we calculate a pooled output by taking an average of the [CLS] BERT classification token embedding over the sequence dimension and use this averaged embedding as our audio sequence representation for downstream classification.

Classification Head: We implement a simple classification head that maps input embeddings to an array of class prediction logits via a series of fully-connected linear layers. The linear layers are of size [`input_embedding_dim`, 32, `num_classes`], which each layer followed by a ReLU activation and dropout layer. The same classification head architecture is used across all HuBERT classifier models, with the only different being the value of `num_classes` to fit its respectively fine-tuned classification task.

The full HuBERT classifier combines each of these components to map raw audio data to class label predictions of the corresponding tasks. We start by fine-tuning two instances of HuBERT and its respective classification head: one with the emotion dataset and one with the sarcasm dataset. Fine-tuning a HuBERT classifier for, for example, emotion recognition involves training the classifer model on the emotion audio samples and updating the weights of both the HuBERT encoder and the classification head to achieve the best prediction results/latent space representation for the emotion recognition task.

2.4. Transfer Learning

After fine-tuning a HuBERT classifier on a particular classification task, we then utilize transfer learning to evaluate the performance of the fine-tuned HuBERT on the other classification task. Here, we identify Task A as the *original* task on which the fine-tuned HuBERT classifer was trained, and Task B as the new evaluated task on which the transferred architecture will be trained. We achieve effective transfer learning via the following steps:

1. From a previous HuBERT classifier fine-tuned on Task A, freeze the model weights of the HuBERT encoder.

2. Transfer the frozen HuBERT model to a different classifier architecture suitable for Task B. This is necessary because, for example, a HuBERT classifier fine-tuned to predict 8 emotion classes requires a new classification head to instead predict a binary sarcasm label.

3. Fine-tune the second, transferred architecture on Task B. Since the HuBERT encoder in the transfered architecture is frozen, fine-tuning here will use the HuBERT weights fine-tuned on Task A to learn a new classification head suitable for evaluation on Task B.

After transfer learning, we will then have two classifier models using the same fine-tuned instance of HuBERT: 1) the HuBERT fine-tuned on Task A, predicting labels for Task A, and 2) the HuBERT fine-tuned on Task A, predicting labels for Task B. In other words, we can now evaluate same fine-tuned instance of Hubert on both the emotion task and sarcasm task separately.

3. EXPERIMENTS

3.1. Experimental Setup

Our experimental setup for model evaluation is as follows: first, we fine a HuBERT classifier on the emotion data to obtain an emotion fine-tuned instance of HuBERT (E_HuBERT) and its corresponding emotion classification head. We then utilize transfer learning to use the E_HuBERT weights to train an additional classification head that uses the emotion fine-tuned weights to predict sarcasm instead. We follow the steps above for a sarcasm fine-tuned instance of HuBERT (S_HuBERT) as well.

In total, we have implemented and trained four models: an emotion fine-tuned HuBERT with both an emotion recognition classification head (8 classes) and a sarcasm recognition classification head (binary class), and a sarcasm fine-tuned HuBERT (S_HuBERT) with both an emotion and sarcasm classification head. We expect each fine-tuned instance of HuBERT to perform best in their original data's benchmark (e.g., E_HuBERT should be best at classifying emotion), with the significance of our results coming from any performance gains found in the transfer learned models (e.g., E_HuBERT performing well/better than S_HuBERT in classifying sarcasm).

Each of the four HuBERT classifier models were fine-tuned on their respective data for 20 epochs, each with the following hyperparameters: learning rate of $5e^{-5}$, weight decay of0.01, dropout rate of 0.1. All classifier models had the same general architecture, and differed only in either the particular weights of the fine-tuned HuBERT and/or the weights/output dimension of the classification head.

3.2. Evaluation

We present our evaluation metrics in Table 1 for the emotion fine-tuned HuBERT (E_HuBERT) and the sarcasm fine-tuned HuBERT (S_HuBERT) in both the sarcasm classification task with MUStARD and the emotion classification task with RAVDESS. We also visualize the metrics pictorially in the bar charts in Figure 2.

Table 1. Evaluation metrics on the MUStARD sarcasm dataset (top) and the RAVDESS emotion dataset (bottom) for both the emotion fine-tuned HuBERT (E_HuBERT) and sarcasm fine-tuned HuBERT (S_HuBERT). Metrics provided are the class label accuracy and F1 classification score.

MUStARD metrics	Accuracy	F1
E_HuBERT	**0.7222**	**0.7370**
S_HuBERT	0.1407	0.1208

RAVDESS metrics	Accuracy	F1
E_HuBERT	**0.5652**	**0.5644**
S_HuBERT	0.5072	0.3365

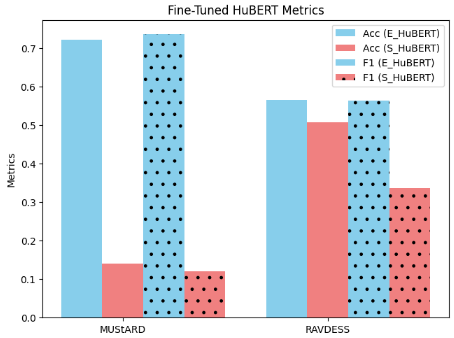

Fig. 2. Evaluation metrics for the emotion fine-tuned HuBERT and sarcasm fine-tuned HuBERT. For both HuBERTs, we compute the accuracy and F1 score on both the HuBERT's originally fine-tuned task and the transfer learned task.

We note a couple of general trends from the evaluation metrics. First is the general under-performance of S_HuBERT in all cases, including the sarcasm recognition task on which it was originally fine-tuned. The results for this model are not satisfactory, and this may be due to a variety of reasons. Since both E_HuBERT and S_HuBERT are derived from the same architecture, same training scheme, and same hyperparameters, the issue likely lies in the quality of the sarcasm training data itself: perhaps the clips pulled from TV shows aren't as isolated and focused as the emotion data; the sarcasm data could require more pre-processing than the emotion data to explicitly filter out any excess applause, laughter, or external speaking within the clip to further refine the audio; or, most profoundly, perhaps this is an indication that sarcasm isn't quite suited to be completely distinguishable via audio alone, which is a point that we discuss in more detail in our Discussion section.

We also note the relatively high performance of E_HuBERT in all cases. First, E_HuBERT was able to learn an effective representation of the emotion data and was able to correctly predict emotion with 72.22% accuracy – significantly better than S_HuBERT near-random classification. More interestingly, we see that E_HuBERT demonstrates performance gains in sarcasm recognition over S_HuBERT. In other words, the transfer learning of the emotion fine-tuned HuBERT demonstrated increased performance in the sarcasm recognition task compared to that of HuBERT not fine-tuned on emotion. This suggests that a strong representation of emotion in HuBERT can also lead to better representations of sarcasm in audio.

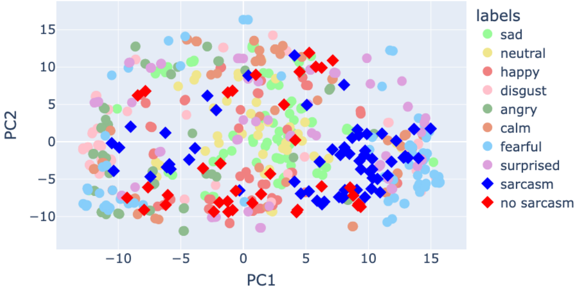

Fig. 3. PCA decomposition of the emotion HuBERT-generated audio embeddings from both the emotion and sarcasm datasets. We note the (loose) clustering of emotion classes and sarcasm classes amongst their respective dataset samples, as well as the relative distance between sarcasm samples and emotion samples across datsaets.

3.3. Analysis

To visualize the learned embeddings, we plot the principal component analysis (PCA) dimensionality reduction of samples from both the emotion dataset and sarcasm dataset, colored by class label. Here, we only consider the better-performing emotion HuBERT for embedding visualization, as analysis is not as clear for the under-performing sarcasm HuBERT.

Ideally, for a well-performing emotion HuBERT classifier, we would see distinct, isolated groupings of emotion labels in the PCA plot. Given that our trained emotion HuBERT did not achieve perfect accuracy, there is some confusion in distinguishing emotions such as ”neutral”, ”sad”, or ”calm”. However, we do still see some groupings toward the outer edges of the graph for strong emotions such as ”fearful” or ”surprised”, which seems intuitive for the more expression emotions to be classified more easily.

For the plotted sarcasm samples, we see a similar grouping between the positive (blue) sarcasm examples and the negative (red) ones. This is an indication that despite being fine-tuned for an emotion recognition class, the emotion fine-tuned HuBERT weights are still valuable and effective at recognizing sarcasm.

More profoundly, we note the distance between samples across datasets. Not only do we notice positive sarcasm examples clustered closely to other positive sarcasm examples, but we also see them generally clustering close to particular emotions such as ”fearful” and ”disgust”, which negative sarcasm examples cluster more closely to ”happy” and ”sad”. We calculate the closest emotion embedding vector to each sarcasm data sample by cosine vector distance and compile the results in Table 2 to show the top 3 closest emotions to positive and negative sarcasm examples. This evidence further reinforces the notion that strong representations of emotion can boost performance in sarcasm prediction, as the representations between both tasks are somewhat correlated. The results in Table 2 are somewhat intuitive for the positive sarcasm case, where the learned representations for the delivery and prosody of positive sarcasm are most closely related to the stronger of the 8 emotions; greater exaggerations in pitch frequency, pitch variation, etc. not only generally indicate stronger emotions but also positive sarcasm. The results for the negative sarcasm case are less intuitive and may be a result of both the imperfect emotion classification accuracy and shortcomings in sarcasm representation more generally.

Table 2. The nearest emotion embedding vectors for both positive sarcasm and negative sarcasm examples, calculated via cosine vector distance, summed across all samples.

Sarcasm?	Nearest Emotions	Freq. Count
Yes	*fearful*	61
	disgust	48
	angry	34
No	*happy*	49
	disgust	41
	sad	34

4. CONCLUSION

In summary, our contributions in this paper include 1) fine-tuned HuBERT audio classifier models, trained on both emotion and sarcasm recognition tasks, 2) a transfer learning methodology that allows us to evaluate our emotion fine-tuned HuBERT classifier on sarcasm data, and vice versa, and 3) discoveries in a positive correlation between effective emotional embeddings and increased sarcasm prediction performance.

One next step to take from this project is to address the under-performance of our sarcasm fine-tuned HuBERT, which is a significant limiting factor in our results and may be due to a number of reasons. Most notably, it could be an indication that speech audio alone is not nearly sufficient enough for fully determining sarcasm; other sarcasm cues are not only beneficial to but requires for learning an effective representation of sarcasm in audio embeddings, both in isolation and in conjunction with emotion recognition training.

Another natural next steps may be to utilize an improved method for calculating the HuBERT pooled output used for sequence embeddings, as using a method such as a weighted-sum mechanism to aggregate hidden layer representations [9] instead could be both beneficial to performance and computational efficiency.

Through our work in HuBERT transfer learning between sarcasm and emotion detection, we provide insight into the benefit of utilizing transfer learning to jointly learn embeddings of both emotion and sarcasm to boost the predictive performance in both tasks. Our findings also motivate additional work in multimodal sarcasm recognition that may ameliorate the accuracy shortcomings we find in this work; incorporating additional signals such as visual facial cues, textual context conditioning, etc. may not only greatly benefit sarcasm recognition alone, but also reinforce the relationships we discover between latent sarcasm embeddings and emotion embeddings in a fine-tuned audio latent space. We hope in future work to both pursue architectural improvements in sarcasm detection and further investigate whether the positive correlation between sarcasm and emotion embeddings retains in better performing models.

5. REFERENCES

[1] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[2] Ganesh Masud Prof. Sunantha Krishnan Prof. Vijaya Bharathi Jagan Ayush Jain, Prathamesh Patil, "Detection of sarcasm through tone analysis on video and audio files: A comparative study on ai models performance," *SSRG International Journal of Computer Science and Engineering*, vol. 8, pp. 1–5, 2021.

[3] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, Sept. 2023.

[4] Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha, "Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA, 2023, MM '23, p. 3926–3933, Association for Computing Machinery.

[5] Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung, "Multimodal speech emotion recognition using cross attention with aligned audio and text," in *Interspeech 2020*. Oct. 2020, interspeech$_2$020$, ISCA.$

[6] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria, "Towards multimodal sarcasm detection (an _Obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez, Eds., Florence, Italy, July 2019, pp. 4619–4629, Association for Computational Linguistics.

[7] Steven Livingstone and Frank Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," in *PLoS ONE*, 2018.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[9] Shu wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, Shinji Watanabe, and Hung yi Lee, "A large-scale evaluation of speech foundation models," 2024.