PERSONALIZED SENTIMENT ANALYSIS USING MULTIMODAL TRANSFORMER (11-751/18-781)

Zhenyu Lu[1], *Angela Chen*[1*], *Ajith Potluri*[1*]

[1] Carnegie Mellon University
{zhenyulu, xinyuc2, ajithp}@andrew.cmu.edu

Abstract

In recent years, there has been an increasing trend in the field of emotion recognition, owing to its application in various fields. Human emotion involves a combination of acoustic and facial features. This is known as multi-modal emotion recognition. Most of the research done in this field has been on a broad and generalized data, which lacks user-specific features. Our proposed approach uses a Multi-modal transformer model (audio, video and language) and a user adapter to create personalized models with few-shot transfer learning. We use the model to evaluate on the CMU-MOSI dataset.

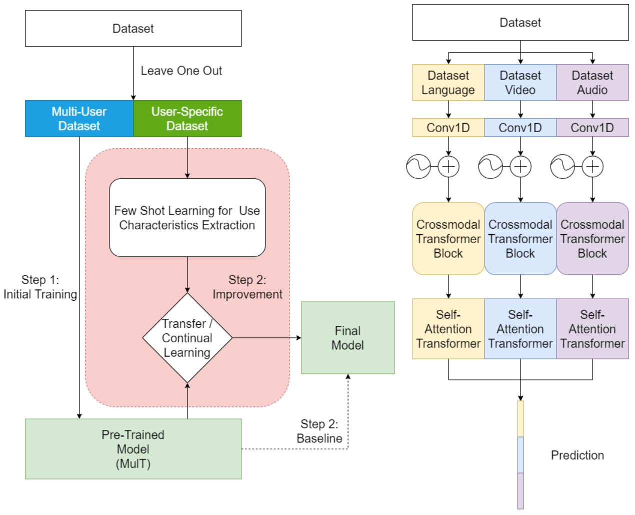
Index Terms— transfer learning, personalized sentiment analysis, multi-modal transformer

1. INTRODUCTION

The importance of emotion recognition or sentiment analysis has seen an escalating trend and has wide application in the real world in recent years. Smart-devices, such as digital assistance, are increasingly being developed, so understanding human emotion can potentially extensively improve the performance of such devices that have intensive interaction with humans. For instance, device could match music corresponding to the mood of audience. Benefited from the incline of public's mental awareness, wearable sensors with a focus on mental health sensing have been widely employed [1] and emotion can be an important indicator evaluating users' mental status. Emotion recognition models have been widely studied in both unimodal and multimodal form: DialogueRNN proposes an attentive RNN model for emotion detection in conversation [2]; GLobal-Aware Multi-scale (GLAM) is CNN based emotion detection model with different convolutional kernel to learn multi-scale feature representation [3]. Despite speech is one of the most widely used modalities for emotion recognition, emotion detection models can also be trained with data from different modalities to improve performance and existing models are discussed in the related study.

One of the main challenges for sentiment analysis is that personal emotion expression can significantly vary with respect to the user's personality, vocal attributes, etc. Therefore, personalized model can be a solution to further improve the recognition performance [4]. However, in the real-world, it is almost impossible to collect a large amount of data from a specific user to create a model and there are also concerns around privacy-preservation.

Transfer learning is widely adopted in the computer vision and speech recognition domain [5, 6]. To address issue of data scarcity, we propose a transfer learning based pipeline – we first train a generalization model with large public datasets and fine tune the model with information extracted from very few small user-specific dataset in a few-shot manner. Our baseline model is adopted from MulT [7] that can process inputs from three modalities (language, audio, facial expression). In addition to the base model, we apply User Adapter [8] to extract user-specific features. We evaluate our proposed model on CMU-MOSI [4], a sentiment analysis dataset. We select 10 users as our candidates for user-specific dataset and run experiments on each selected user in a leave-one-out way. There are some performance improvement with our personalized model (e.g.: Ac_2 improved by 4%), but not all evaluation metrics are improved.


Dataset
Leave One Out
Multi-User Dataset
User-Specific Dataset
Few Shot Learning for Use Characteristics Extraction
Step 1: Initial Training
Step 2: Improvement
Transfer / Continual Learning
Final Model
Step 2: Baseline
Pre-Trained Model (MulT)
Dataset
Dataset Language
Dataset Video
Dataset Audio
Conv1D
Conv1D
Conv1D
Crossmodal Transformer Block
Crossmodal Transformer Block
Crossmodal Transformer Block
Self-Attention Transformer
Self-Attention Transformer
Self-Attention Transformer
Prediction

(a) Pipeline **(b) Baseline Model (MulT)**

Fig. 1. Pipeline

2. RELATED STUDIES

2.1. Multi-modal Emotion Recognition

Most multi-modal models constructed for emotion detection used a mixture of natural language, facial gestures, and acoustic behaviors as their input. One such transformer based network, called multi-modal end-to-end transformer (ME2ET) uses a two-pass strategy to model the interaction between tri-modal features [9]. Contextualized Graph Neural Network based Multi-modal Emotion recognitioN (COGMEN) leverages local information (i.e., inter/intra dependency between speakers) and global information (context) [10] to

[*] Equal contributions

model the complex dependencies (local and global information) in a conversation. Multi-modal Transformer (MulT) is emotion recognition model specifically designed to process unaligned sequences as multi-modal language sequences often exhibit “unaligned” nature and require inferring long term dependencies across modalities [7]. At the heart of the model is the cross-modal attention module, which latently adapts streams from one modality to another (e.g., vision → language) by repeated reinforcing one modality’s features with those from the other modalities, regardless of the need for alignment. However, the work so far has been performed on a large generalized dataset which lacks personalization, our approach is to use the Multi-modal Transformer for user-specific sentiment analysis.

2.2. Transfer Learning

Transfer learning focuses on adapting knowledge from available auxiliary resources to transfer this learning to a target domain, where a very few or even no labelled data is available. It involves reusing knowledge from one field to another using neural networks. One Transfer learning approach utilizes Deep Belief Networks (DBN) to improve the performance of speech emotion recognition systems in cross-language and cross-corpus scenarios [11]. Our approach is highly inspired by the framework proposed by Nikolaos et al. [1]. They first trained a CNN with a multi-user dataset for generalization and then is fine-tuned with a small speaker-specific dataset by freezing transferred layers and adding new layers. Our approach uses transfer learning to make use of the generalized data and make it more personalized.

2.3. Prompting

Prompting is the technique of adding instructions or conditions to the input sequence in NLP. GPT-3 [12] uses prompts to adapt its generalized model to different tasks. Prompting technique has been thoroughly studied in prior works, especially for models like BERT and RoBERTa [13–15]. Prefix tuning [16], as an extension of prompting, is a prompt-like token used for light weight fine tuning for different language generation tasks, which freezes its base model but learns trainable task-specific embedding. In the computer vision domain, vision transformer (ViT) [17] also introduces such kind of virtual token as the class token and only use the token for predictions,forcing the learnable token to interact with the data sequence. User Adapter [8], which has the same mechanism, introduces a token for each individual user instead of specific context or condition. We utilize a similar method as User Adapter to produce our personalized transformer model.

2.4. Knowledge Distillation

Knowledge Distillation (KD), first introduced by Hinton et al [18], is the process of transferring knowledge learnt from a model (teacher model) to another model (student model). While learning the new model, the training paradigm not only considers hard labels (maximum scores of the student model) but also soft labels (output from the teacher model). Knowledge Distillation can be used for multiple purposes. Firstly, it can function as compressing a large model into a lighter model. In addition, it can also be used for continual learning. In this context, the model has to learn new knowledge (more tasks) continuously, causing significant trouble of catastrophic forgetting [19], which refers to the phenomenon that the neural network model drastically forget learned knowledge previously while learning new knowledge. Knowledge Distillation can serve as a regularization method to deal with the trade-off between reattaining old knowledge and learning new knowledge. Our paper adopt Knowledge Distillation for this purpose. Distillation can also be token based, as introduced in DeiT by Touvron et al [20]. DeiT adds another distillation token similar to the class token in ViT and uses such distillation token to compute the soft label.

3. PROBLEM FORMULATION

Our goal is to design a user-specific model based on the multi-modal transformer model (MulT) [7] with the consideration of three modalities: language *L*, video *V*, and audio *A*. We express our input data as $\{X^{(i)}_{\{L,V,A\}}\}^N_{i=1}$, where N is the number of users.

$$X^{(i)}_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$$

where T is the sequence length and d is the feature dimension. For a specific user i, we train the MulT model with input data $\{X^{(j)}_{\{L,V,A\}}\}^N_{j=1,j\neq i}$ and save the hyperparameters after fine-tuning on the user-specific validation dataset as θ_{MulT}. Our baseline model will generate predictions

$$\hat{Y}^{(i)} = \text{MulT}(X^{(i)}_{\{L,V,A\},\text{test}} | \theta_{\text{MulT}})$$

directly on the user-specific test dataset. Our improvement experiment is discussed in Section 4, where we apply transfer learning on user-specific features and pre-trained model outputs. The evaluation metrics are discussed in Section 5, and our goal is to maintain the accuracy scores and improve the correlation score.

4. METHOD

As shown in Fig. 1(a), we first partition the dataset in the cross-subject setting, and conduct leave-one-out analysis for 10 subjects by splitting the dataset into user-specific dataset $X^{(i)}_{\{L,V,A\}}$ (containing features from a single subject) and multi-user dataset $\{X^{(j)}_{\{L,V,A\}}\}^N_{j=1,j\neq i}$. We use the multi-user dataset to train the MulT model (step 1). We skip the step 2 for the baseline model and then use the train set of the user-specific dataset to fine-tune the model and evaluate on the user-specific validation dataset (step 3). In addition, we apply knowledge distillation to combine user-specific characteristics with the knowledge we learned from the pre-trained model.

4.1. Model

Our model is composed of two parts – base model and user adapter. MulT is transformer based model that fuses multi-modal features by multiple cross-modal transformers as shown in Fig. 1(b), with each cross modal transformer interleaved the target modality with attention learning across other two domains. Detailed structure of MulT can be found in [7]. We use MulT as our base model, but more complicated transformer based models can be easily adapted to the framework as well.

To retrieve user-specific characteristic from limited samples, we adopt the User Adapter [8] in addition to our base model. Therefore, the base model can retain the meta knowledge learnt from the pre-trained stage and the User Adapter can focus on learning features that are unique to the user. The core of User Adapter is its learnable embedding u appending to the end of input sequence as shown in

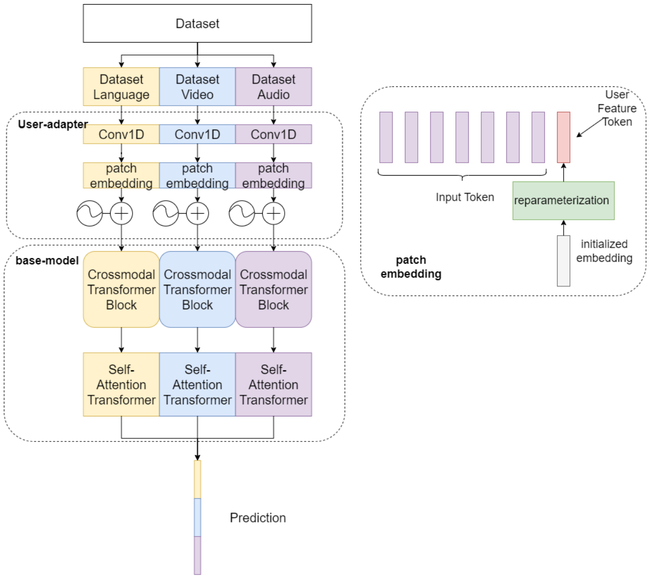


Fig. 2. User Adapter Architecture

figure 2. For each modality, we have an individual learnable token adding to the input sequence.

The base model contains a 1D temporal convolution layer to reinforce the locality of each element of the input sequences and project the features from different modalities onto the same dimension:

$$\hat{X}_{\{L,V,A\}} = Conv1D_{temporal}(X_{\{L,V,A\}}, k_{\{L,V,A\}}),$$

where $k_{\{L,V,A\}}$ are the sizes of the convolution kernels for modalities {L,V,A}, and d is the shared projected dimension. We move it from the base model to the user-adapter, such that the user-specific embedding can be append to the input sequence after passing through the convolution layer:

$$h = [\hat{X}, u],$$

where u is the final user-specific embedding token.

To avoid unstable optimization of the user-specific embedding, we also incorporate parametrization with an MLP layer MLP_θ,

$$u = MLP_\theta(u'),$$

where u' denotes embedding before parametrization. This parametrization can also be replaced with more complex methods.

4.2. Training

At the pre-trained stage, unlike many other existing literature, we also train the User Adapter together with the base model, since there are not enough samples to optimize the User Adapter at few-shot training stage.

During the training stage, we adopt separate learning rates for the base model and the adapter. We also implemented Knowledge Distillation by adding an extra loss term computing the difference between the output from pre-trained model and the loss from current model:

$$Loss = (1 - \lambda) * L_{kd}(\hat{y}, y_{old}) + \lambda * L_{raw}(\hat{y}, y),$$

Table 1. Baseline Results without Fine Tuning

Metric	MAE^l	Corr^h	Ac$_7^h$	F1^h	Ac$_2^h$
1	0.95	0.68	0.33	0.74	0.76
2	1.37	0.54	0.40	0.60	0.67
3	1.01	0.63	0.33	0.70	0.71
4	0.96	0.48	0.40	0.93	0.93
5	0.68	0.80	0.53	0.84	0.83
6	0.91	0.63	0.27	0.73	0.73
7	0.67	0.83	0.40	0.96	0.96
8	1.08	0.65	0.23	0.76	0.79
9	0.89	0.78	0.43	0.84	0.85
10	0.90	0.81	0.30	0.81	0.82
Avg	0.94	0.68	0.36	0.79	0.81

Table 2. User Adapter with Few-Shot Learning

Metric	MAE^l	Corr^h	Ac$_7^h$	F1^h	Ac$_2^h$
1	0.73	0.72	0.50	0.79	0.80
2	1.09	0.63	0.27	0.80	0.83
3	1.80	0.66	0.07	0.46	0.39
4	0.82	0.50	0.37	0.95	0.90
5	0.91	0.82	0.30	0.84	0.83
6	0.77	0.67	0.33	0.74	0.73
7	0.71	0.84	0.50	0.90	0.89
8	0.93	0.46	0.27	0.93	0.86
9	0.76	0.73	0.37	0.82	0.81
10	1.13	0.73	0.23	0.85	0.86
Avg	0.97	0.68	0.32	0.81	0.79

where $\hat{y}$ is the network prediction, y_{old} is the prediction generated from pre-trained model, y is the ground truth, and λ is the hyper-parameter.

$$L(\hat{y}, y) = |y - \hat{y}|$$

We experimented multiple distillation functions, including cross entropy, KL divergence, etc., and found using the same L1 loss function as the raw loss achieves the best performance,

5. EXPERIMENTS

5.1. Datasets

The multimodal features are extracted as follows. Audio features are extracted using COVAREP [21], video features are taken from Facet (iMotions), and text features are extracted using GloVe word embeddings [22]. We split the full dataset into multi-user dataset and user-specific dataset. 70% of multi-user dataset is used for training and 30% for validating the pre-trained model. User-specific dataset is used to fine tune the pre-trained model.

CMU-MOSI [4] is a dataset annotated with sentiment (including annotated video and audio features) based on 2,199 clips from 93 videos. Prior works use this dataset for multimodal testing. The annotated sentiment score ranges from -3 (strongly negative) to 3 (strongly positive). The sampling rate of video and audio features is 12.5 at 15 Hz, and the textual language data is expressed as discrete word embeddings based on the segmented words. 10 subjects with the highest sample size were selected as the user-specific dataset. Our evaluation metrics include five scores: 7-class accuracy (Ac$_7$:

sentiment score classification scoring from -3 to 3), binary accuracy (Ac_2: positive/negative sentiments), F1 score, Mean Absolute Error (MAE), and the correlation between model's prediction with human (Corr).

5.2. Baseline

To generate our baseline results, we directly apply the pre-trained model and evaluate on the user-specific test set for each selected subject. For each evaluation metric, the mean is calculated among all leave-out-out splits, and is shown on Table 1.

5.3. Experimental Setups

During the pre-trained stage, we set up a 5-layer MulT model with 10-head transformer. For our User Adapter, the prompt length is set to 1 such that we have only one token appending to the input. The output dimension is 30 and is consistent with our input feature dimension. We re-parameterize the user-specific embedding with MLP and the hidden dimension is set to 512. We adapt Adam optimizer and use 1e-3 learning rate to train 40 epochs on data.

During the fine-tuning stage, we adopt 5-shot learning setup and the five samples are randomly selected from the user-specific training dataset. We utilize a separate learning rate strategy: the learning rate for base model and User Adapter is 1e-4 and 1e-3 respectively. The temperature for Knowledge Distillation loss is set to 1 and λ is 0.6.

For each selected user, we run one experiment by first splitting up the whole dataset into multi-user dataset and the user-specific dataset and then run through the full pre-train and fine-tune process.

6. RESULTS AND DISCUSSION

For our baseline without fine tuning, we obtained Ac_7 accuracy of 36 percent, Ac_2 accuracy of 81 percent, F1 score of 79, Mean Absolute Error of 0.94 and correlation of 068. These results are comparable with our references.

As an intermediate model setup, after implementing the User Adapter, we fine tune the model with 5 samples from user-specific dataset without any regularization, and the result is shown on Table 2. However, comparing our baseline and the improved model with fine tuning, we found the overall performance declines. Exploring the performance per case, we found there are some cases where performance improved, but some getting worse significantly. For instance, when conducting experiments on user 3, the binary accuracy drops from 0.71 to 0.39. One possibility accounting for the downgrade is that we ran experiments under few shot setting and we also randomly picked those samples. Those biased samples can also bias the model significantly especially under the few-shot setting.

To deal with this problem, we incorporate the concept of knowledge distillation into our final design. We found the traditional knowledge distillation works better, compared with the distillation loss token design proposed in DeiT, which is another trainable token appending to the input sequence.

For our final setup, we found improvements in performance of some evaluation metrics as shown in Table 3, but not all. Baseline is evaluated without learning, FS-UA stands for Few-Shot with User Adapter, FS-UA-KD denotes Few0Shot with User Adapter and Knowledge Distillation. The tests are evaluated on CMU-MOSI dataset. The average binary accuracy increase by 4% when comparing our final result with the baseline. However, for more detailed level based, such as Ac_7, the lower average accuracy is not expected. For some experiments, the Ac_7 does improve: Ac_7 improves from 0.33 to 0.5 when experimenting with user. Trained with user-specific samples, ideally would make the model have a greater capability to classify the users' sentiment states both in correctness (binary) and intensity (7-class). The primary reason would still be biased samples since we have only 5 samples for training. Dealing with the biased samples will be one possible future directions.

Table 3. Subject Specific Multimodal Sentiment Analysis Results

Metric	MAE^l	$Corr^h$	Ac_7^h	$F1^h$	Ac_2^h
Baseline	0.94	0.68	0.36	0.79	0.81
FS-UA	0.97	0.68	0.32	0.81	0.79
FS-UA-KD	0.89	0.69	0.35	0.85	0.85

7. CONCLUSION

In this project, we constructed a personalized sentiment analysis model and ran experiments on CMU-MOSI dataset. As mentioned in the discussion, the results are not fully as expected, so new ideas have to be introduced to make the model more robust to the biased samples. The model needs to extract user-specific features from the samples while not allowing the unrelated information in the samples to bias the model. Emotion detection is another similar domain that is especially in need of personalized model. Therefore, we will adapt our model for emotion detection in the future.

Personalized model is especially important for HCI related tasks, since individual difference makes the trained model difficult to generalize to different users. Similarly, changes that a person may make over time to his or her mindset might exert extra vulnerability for predictive model especially. Therefore, a personalized model that can continuously update itself based on user input throughout the time is going to be the future trend of digital world. Especially in the medical or health-related domain, electronic health records are always user dependent and subject to change over time. Further, users value their privacy and might not be willing to share their personal information for training a model, a personalized model specially designed for each individual is a perfect solution for the issue. And light weight models are also preferred for personalizing of models since those models are more likely to be deployed on edge devices. And keeping the model small and easy to fine tune is needed and will be another future direction to work on.

8. REFERENCES

[1] Nikolaos Vryzas, Lazaros Vrysis, Rigas Kotsakis, and Charalampos Dimoulas, "A web crowdsourcing framework for transfer learning and personalized speech emotion recognition," *Machine Learning with Applications*, vol. 6, pp. 100132, 2021.

[2] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," 2018.

[3] Wenjing Zhu and Xiang Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," 2022.

[4] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016.

[5] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," 2014.

[6] Dong Wang and Thomas Fang Zheng, "Transfer learning for speech and language processing," 2015.

[7] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019.

[8] Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan, "Useradapter: Few-shot user learning in sentiment analysis," in *FINDINGS*, 2021.

[9] Yang Wu, Pai Peng, Zhenyu Zhang, Yanyan Zhao, and Bing Qin, "An efficient end-to-end transformer with progressive tri-modal attention for multi-modal emotion recognition," 2022.

[10] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi, "Cogmen: Contextualized gnn based multimodal emotion recognition," 2022.

[11] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps, "Transfer learning for improving speech emotion classification accuracy," 2018.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[13] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig, "How can we know what language models know?," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] Timo Schick and Hinrich Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[16] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network (2015)," *arXiv preprint arXiv:1503.02531*, vol. 2, 2015.

[19] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10347–10357.

[21] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep — a collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543, Association for Computational Linguistics.