INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

LanSER: Language-Model Supported Speech Emotion Recognition

Taesik Gong[1], *Josh Belanich*[2*], *Krishna Somandepalli*[2*], *Arsha Nagrani*[2], *Brian Eoff*[2], *Brendan Jou*[2]

[1]KAIST, Republic of Korea
[2]Google Research

taesik.gong@kaist.ac.kr, {joshbelanich,ksoman,anagrani,beoff,bjou}@google.com

Abstract

Speech emotion recognition (SER) models typically rely on costly human-labeled data for training, making scaling methods to large speech datasets and nuanced emotion taxonomies difficult. We present LanSER, a method that enables the use of unlabeled data by inferring weak emotion labels via pre-trained large language models through weakly-supervised learning. For inferring weak labels constrained to a taxonomy, we use a textual entailment approach that selects an emotion label with the highest entailment score for a speech transcript extracted via automatic speech recognition. Our experimental results show that models pre-trained on large datasets with this weak supervision outperform other baseline models on standard SER datasets when fine-tuned, and show improved label efficiency. Despite being pre-trained on labels derived only from text, we show that the resulting representations appear to model the prosodic content of speech.

Index Terms: speech emotion recognition, large language models, weakly-supervised learning

1. Introduction

In conversations, humans rely on both *what is said* (i.e., lexical content), and *how it is said* (i.e., prosody), to infer the emotion expressed by a speaker. State-of-the-art methods in speech emotion recognition (SER) leverage the interplay of these two components for modeling emotional expression in speech. However, such methods still show limitations on in-the-wild scenarios due to the variability in natural speech, and the reliance on human ratings using limited emotion taxonomies. Extending model training to large, natural speech datasets labeled by humans for nuanced emotion taxonomies is expensive and is further complicated by the subjective nature of emotion perception.

Despite both lexical content and prosody being complementary for emotion perception, the two components are *correlated*, and in many cases the content is predictive of the prosody. For example, when someone says, "I won the lottery" – an upbeat and lively prosody would sound congruent, and one might perceive the emotional expression as elation or triumphant. In this work, we investigate how we might leverage the emotions congruent with lexical content in large unlabeled speech datasets to serve as weak supervision for developing SER models.

We turn to Large Language Models (LLMs) to infer expressed emotion categories in textual content. Due to the knowledge they embed from pre-training on large text corpora [1, 2], LLMs have demonstrated capabilities in numerous downstream tasks [3], including a few subjective tasks such as social and emotion reasoning [4]. In domains such as computer vision, LLMs were explored to reduce the need for labeled data, e.g., for visual question answering [5]. However, to our knowledge, they have not been studied for emotion recognition tasks, particularly from natural speech.

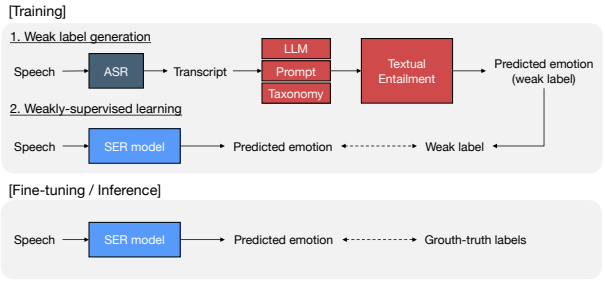

Figure 1: *The overview of LanSER. LLMs and textual entailment are used to infer weak emotion labels from speech content which are used to pre-train a SER model.*

We propose *LanSER*, that uses LLMs to infer emotion categories from speech content i.e., transcribed text, which serve as weak labels for SER (Figure 1). Overall, LanSER enables pre-training a SER model on large speech datasets without human labels by (1) extracting text transcripts from utterances using automatic speech recognition (ASR), (2) using pre-trained LLMs to infer weak emotion labels with an engineered prompt and predetermined taxonomy, and (3) pre-training the SER model with the weak labels. We demonstrate that LanSER improves SER performance and label efficiency by fine-tuning on benchmark datasets. Moreover, we show that despite the emotion labels being derived from speech content only, LanSER captures speech prosody information that is relevant to SER.

2. Related Work

SER with LLMs: Recently, LLMs were used to generate pseudo-labels for semi-supervised learning for speech sentiment analysis [6]. Here, LLMs were fine-tuned on a *labeled* sentiment dataset to explore narrow sentiment classes of negative, positive, and neutral. In contrast, our work avoids fine-tuning LLMs on task-specific datasets by inferring weak labels via textual entailment, enabling exploration with wider emotion taxonomies. In the context of multi-modal emotion recognition, MEmoBERT [7] used audio, visual, and text information with prompt learning for unsupervised emotion recognition. Herein, the visual model is pre-trained on a large labeled emotion dataset. In contrast, in our work, pre-training on large human-annotated emotion datasets is not necessary.

Self-supervised learning: Self-supervised learning has become a popular method using large amounts of unlabeled speech data

*Equal contribution.

10.21437/Interspeech.2023-1832

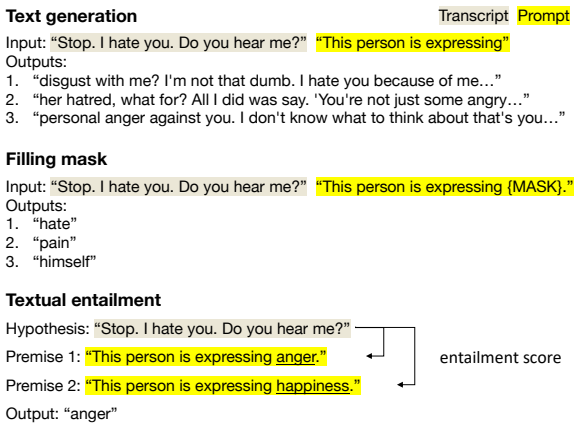

Figure 2: *Comparison of three weak label generation approaches: text generation, filling mask, and textual entailment.*

for pre-training [8, 9]. Recent studies found that large pre-trained models via self-supervised learning show effectiveness in various downstream speech tasks [10], including many paralinguistic tasks [9]. We view self-supervised learning and our weak supervision from LLMs as complementary, since the two methodologies can be combined for training SER models.

3. Methodology

An overview of the training and inference process of LanSER in shown in Figure 1. During pre-training, we use ASR to generate transcripts from speech utterances, which are fed into a LLM with appropriate prompt to extract weak emotion labels in predetermined taxonomy via textual entailment. These labels are used to pre-train a SER model via weakly-supervised learning. The pre-trained SER model can then either be used directly to output emotion predictions according to the emotion taxonomy used to extract weak labels, or can be adapted for a different taxonomy or dataset by fine-tuning.

We note that the emotions inferred using LLMs from speech content are proxies for the emotion being expressed, and may not capture the larger context or intent of the speaker. Thus, we treat them as "weak" emotion labels in our work.

3.1. Weak label generation via textual entailment

There are multiple ways to use LLMs for extracting weak emotion labels. Two dominant approaches in the literature are (i) text generation [2] and (ii) filling mask [11, 1, 7]. Figure 2 demonstrates the behaviors of text generation and filling mask for weak emotion label prediction. We used representative LLMs for each approach: GPT-2 for text generation and BERT [11] for filling mask. While these approaches show some success, the common limitation in a zero-shot setting is that they often output undesirable "noise", like irrelevant words (text generation), or non-emotional responses (e.g., "himself" in filling mask in the Fig. 2).

Thus, we want to constrain the LLM model to output only words relevant to emotion perception. To this end, we use textual entailment [12] to generate weak labels that also allows us to constrain the emotion taxonomy apriori. Figure 2 illustrates the entailment-based weak emotion label generation; at a high-level, this method calculates the entailment scores between an input transcript (called *hypothesis*) and prompts with candidate labels from the taxonomy (called *premise*), and then selects the item with the highest score as the weak label. Formally, let $x \in \mathcal{X}$ denote ASR transcripts from speech and $y \in \mathcal{Y}$ denote a candidate label in taxonomy $\mathcal{Y}$. A prompting function $g(\cdot)$ prepends a predefined prompt to the given input. $f(x, g(y))$ denotes the entailment score between a hypothesis x and a prompted label $g(y)$. The resulting weak emotion label $\hat{y}$ for a given transcript x is calculated as:

$$\hat{y} := \arg\max_{y \in \mathcal{Y}} f(x, g(y)). \quad (1)$$

The entailment scoring function f is a function typically parameterized by a neural network and fine-tuned on the entailment task. In our case, we use RoBERTa [13] fine-tuned on the Multi-genre Natural Language Inference (MNLI) [14] dataset. The MNLI dataset is composed of hypothesis and premise pairs for diverse genres, which is specialized for the textual entailment approach, and do not explicitly focus on emotion-related concepts.

3.2. Prompt engineering

Prompt engineering is a task-specific description embedded in inputs to LLMs (e.g., a question format) [15]. It is a critical component affecting zero-shot performance of LLMs on various downstream tasks [1, 16, 17]. In Section 4.2 we explore various prompts in order to understand the impact of prompt engineering for the entailment task. Ultimately, we found that the prompt "The emotion of the conversation is {}." performed best, and we use this prompt throughout our experiments.

3.3. Taxonomy

The choice of emotion taxonomy is critical in developing SER models as emotion perception and expression is nuanced. Common SER benchmarks typically use 4–6 emotion categories [18, 19], which do not capture the variability in emotion perception [20]. Thus we experiment with BRAVE-43, a finer-grained taxonomy [21]. We adopted and modified the BRAVE taxonomy which originally contains 42 self-reported emotions labels. We converted several two-word emotions to one-word emotions for simplicity and added "shock" to capture a negative version of "surprise", resulting in a total of 43 categories. Note this taxonomy is not speech-specific. We investigate the impact of taxonomy selection in Section 4.5. We expect fine-grained taxonomies to help learn effective representations by using the high degree of the expressiveness of LLMs.

4. Experiments

Our overarching hypothesis is that, given a sufficiently large amount of data, pre-training speech-only models on weak emotion labels derived from text improves performance on SER tasks. As such, throughout this paper, we focus on speech-only emotion recognition models. Additionally, our goal is not to obtain state-of-the-art results on downstream tasks but to assess, given a fixed model capacity, whether models pre-trained via LanSER achieve improved performance.

4.1. Data preparation

Pre-training data: We investigate two large-scale speech datasets for LanSER pre-training: People's Speech [22] and Condensed Movies [23]. People's Speech is currently the largest English speech recognition corpus, containing 30K hours of general speech. Condensed Movies is comprised

of 1,000 hours of video clips from 3,000 movies, where we use only the audio. We explore these two large-scale speech datasets to understand the impact of the amount of data and their distributions; while People's Speech has more samples from less emotional data sources (e.g., government, interview, health, etc.), Condensed Movies has fewer samples from a more emotional data source (movies). We use Whisper ASR [24] ("small" variant) to segment and generate transcripts for People's Speech and Condensed Movies datasets, resulting in 4,321,002 and 1,030,711 utterances, respectively.

Downstream tasks: We use two common SER benchmarks for downstream tasks: IEMOCAP [18] and CREMA-D [19]. IEMOCAP is an acted, multi-speaker database containing 5,531 audio clips from 12 hours of speech. We follow the commonly used four-class (anger, happiness, sadness, and neutral) setup [7, 25, 10, 9] and use speaker-independent train:val:test splits. CREMA-D has 7,441 audio clips collected from 91 actors. An important characteristic of CREMA-D is that it is linguistically constrained, having only 12 sentences each presented using six different emotions (anger, disgust, fear, happy, neutral, and sad). We use CREMA-D to validate that our models indeed learn prosodic representations, and do not just learn to use language to predict the emotional expression.

4.1.1. Baselines

We compare LanSER models fine-tuned on downstream datasets with the following four baselines:

Majority: Output the most prevalent class in the dataset [12].

GT Transcript + Word2Vec [26]: Each word in a ground-truth transcript is converted to a Word2Vec embedding. We compute the cosine similarity between the averaged transcript embedding and each class label, outputting the class with the highest similarity.

GT Transcript + LLM + Entailment [12]: Using the same methodology for producing weak labels, we process the ground-truth transcript with an LLM and entailment to output a classification according to the dataset's taxonomy.

Supervised: Supervised learning using the same model architecture as LanSER but without pre-training.

We include two language-based methods (Word2Vec and Entailment) to better understand how LanSER compares with models using lexical content alone. Note that the language baselines assume GT transcripts are available. In practice, these baselines would require an ASR pipeline to get transcripts, which may involve additional computational and developmental cost.

4.1.2. Implementation

We extracted mel-spectrogram features (frame length 32ms, frame steps 25ms, 50 bins from 60–3600Hz) from the audio waveforms as input to the model and used ResNet-50 [27] as the backbone network for training. For both pre-training and fine-tuning, we minimized the cross-entropy loss with the Adam [28] optimizer and implemented in TensorFlow [29].

For pre-training, we adopted a warm-up learning rate schedule where the rate warmed up for the initial 5% of updates to a peak of 5×10^{-4} and then linearly decayed to zero. We used a batch size of 256 and trained for 100K iterations.

For fine-tuning on the downstream tasks, we loaded the pre-trained weights and used a fixed learning rate of 10^{-4}. We set the batch size as 64 and trained for 10K iterations. We split the downstream datasets into a 6:2:2 (train:valid:test) ratio, and selected the best model on the validation set for testing.

Table 1: *Accuracy of extracted weak emotion labels with various prompts.* {} *indicates the masked position.*

Prompts	Acc.
This example is {}.	42.0%
I am {}. [7]	39.9%
I feel {}.	41.8%
I am feeling {}.	45.0%
This person is expressing {} emotion.	43.7%
A speech seems to express a feeling like {}. [16]	38.0%
A transcript seems to express a feeling like {}. [16]	38.9%
A conversation seems to express some feelings like {}. [16]	39.0%
The emotion of the conversation is {}.	**45.6%**
The emotion of the previous conversation is {}.	44.1%
The overall emotion of the conversation is {}.	45.1%

4.2. Prompt engineering

We investigated the impact of various prompts to infer weak emotion labels using IEMOCAP. We chose IEMOCAP because it has transcripts and human-rated labels with majority agreement referred here as "ground-truth". To evaluate the prompts, we compute accuracy by comparing the weak labels with the ground-truth. We also examined prompts used in previous emotion recognition studies [16, 7] and modified a few vision-specific prompts [16] for our study by replacing words such as "photo" or "image" with "speech".

Table 1 shows the accuracy for each prompt. The prompt ("I am {}.)" used in the related sentiment work [7] was not as effective at capturing emotional signals. Similarly, adapting vision-specific prompts [16] was ineffective. This suggests that it is worthwhile to tailor the prompt to the SER task. Among the prompts we explored, "The emotion of the conversation is {}." had the highest accuracy. We adopt this prompt to infer weak labels in all our experiments. We leave additional prompt tuning [30] as future work.

4.3. Fine-tuning

We fine-tune all models on the downstream tasks to evaluate their label efficiency and performance. To measure label efficiency, we varied the percentage of seen training data from 10% to 100% for each dataset. Table 2 shows the result. "LanSER (People's Speech)" means pre-training with Peoples Speech, while "LanSER (Condensed Movies)" refers to pre-training with Condensed Movies. In all cases, we used the BRAVE taxonomy (see Sec. 3.3) as the label space.

First, NLP baselines (Word2Vec and Entailment) fail on CREMA-D, as they only use lexical speech content. Interestingly, LanSER's results on CREMA-D suggest that the model can learn prosodic representations via weak supervision from LLMs. We attribute this result to pre-training with large-scale data, and it offers evidence to our hypothesis that speech and text emotions are correlated enough that SER models can learn to use prosodic features even with labels from text only given a sufficiently large amount of data.

Overall, LanSER outperforms the NLP and majority class baselines. Notably, LanSER pre-trained with the Condensed Movies showed improved accuracy than with the People's Speech. While People's Speech is comprised of fairly neutral speech data (e.g., government, interviews, etc.), Condensed Movies is comprised of movies having more expressive speech; from the emotion recognition perspective, Peoples Speech might introduce more noise than Condensed Movies.

To assess that performance improvements are being driven by the emotion labels inferred using LLMs, and not just the scale of the pre-training data, we compare the fine-tuning performance of LanSER to a model pre-trained on Condensed

Table 2: *Unweighted accuracy (%) of fine-tuning for downstream tasks with varying the percentage of fine-tuning data (10%, 30%, 50%, 70%, and 100%). Bold fonts indicate the highest accuracy.*

Downstream task	Method \ Fine-tuning data:	10%	30%	50%	70%	100%
IEMOCAP [18]	Majority	30.9%	30.9%	30.9%	30.9%	30.9%
	GT Transcript + Word2Vec [26]	34.9%	34.9%	34.9%	34.9%	34.9%
	GT Transcript + Entailment [12]	47.5%	47.5%	47.5%	47.5%	47.5%
	Supervised	38.5%	41.8%	45.5%	46.0%	47.6%
	LanSER (People's Speech)	42.0%	47.1%	47.2%	**50.0**%	50.6%
	LanSER (Condensed Movies)	**50.0**%	**51.7**%	**48.0**%	45.4%	**54.5**%
CREMA-D [19]	Majority	17.1%	17.1%	17.1%	17.1%	17.1%
	GT Transcript + Word2Vec [26]	19.1%	19.1%	19.1%	19.1%	19.1%
	GT Transcript + Entailment [12]	16.1%	16.1%	16.1%	16.1%	16.1%
	Supervised	37.8%	43.2%	48.2%	53.4%	57.2%
	LanSER (People's Speech)	35.5%	48.2%	51.5%	52.7%	55.8%
	LanSER (Condensed Movies)	**43.7**%	**49.9**%	**52.2**%	**53.6**%	**58.7**%

Table 3: *Unweighted accuracy (%) for fine-tuning on downstream tasks. LanSER (random labels) is pre-trained on Condensed Movies with BRAVE taxonomy labels assigned randomly.*

Downstream task	Method	Accuracy
IEMOCAP [18]	LanSER (random labels)	47.6%
	LanSER (weak labels)	**54.5**%
CREMA-D [19]	LanSER (random labels)	50.6%
	LanSER (weak labels)	**58.7**%

Table 4: *Zero-shot unweighted accuracy (%) of SER models.*

Downstream task	Method	Accuracy
IEMOCAP [18]	Scratch	22.9%
	LanSER (People's Speech)	30.9%
	LanSER (Condensed Movies)	**34.3**%
CREMA-D [19]	Scratch	16.3%
	LanSER (People's Speech)	15.9%
	LanSER (Condensed Movies)	**23.5**%

Movies using random uniformly sampled labels. As shown in Table 3, models pre-trained with weak labels outperform ones trained with random labels suggesting that the weak emotion labels inferred using LLMs are meaningful.

4.4. Zero-shot classification accuracy

A unique advantage of LanSER over self-supervised learning [8, 9] is that it enables SER models to support zero-shot classification. Table 4 shows the zero-shot classification accuracy: for LanSER, SER models were pre-trained with the taxonomy of the downstream dataset instead of BRAVE and evaluated in a zero-shot setting. We use models with randomly initialized weights and no training as a lower-bound of performance, referred to as "Scratch". Overall, LanSER shows higher accuracy than the baseline, although not as good as fine-tuning. These results suggest the potential of training large SER models that can perform well on various downstream tasks, without further fine-tuning. Improving zero-shot performance further using our proposed framework is part of our future work.

4.5. Impact of taxonomy

Figure 3 shows the impact of taxonomy selection. We compared the BRAVE taxonomy with downstream task's taxonomies. "PS" and "CM" refers to People's Speech and Condensed Movies, respectively. "IEMOCAP", "CREMA-D", and "BRAVE" means taxonomy used to generate weak labels. As shown, pre-training with the finer taxonomy (BRAVE) shows generally better accuracy when fine-tuned, with 4.2% accuracy improvement on average. This indicates that a fine-grained taxonomy is beneficial to learn effective representations by leveraging the high degree of the expressiveness of LLMs.

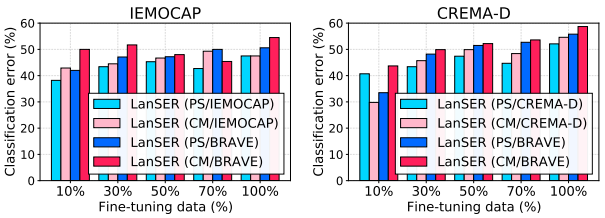

Figure 3: *Impact of taxonomy selection for pre-training.*

5. Caveats

Developing machine perception models of apparent emotional expression remains an open area of investigation. The models in this work do not aim to infer the internal emotional state of individuals, but rather model proxies from speech utterances. This is especially true when training on the output of LLMs, since LLMs may not take into account prosody, cultural background, situational or social context, personal history, and other cues relevant to human emotion perception. ASR transcription errors add another layer of noise.

The benchmark datasets we use in this work are relatively small and are labeled with limited emotion taxonomies. CREMA-D, while useful for its fixed lexical content, is an acted dataset where the emotional expression of its utterances may not well-represent natural speech.

6. Conclusion and Future Work

In this work, we proposed LanSER, a novel language-model supported speech emotion recognition method that leverages large unlabeled speech datasets by generating weak labels via textual entailment using LLMs. Our experimental results showed that LanSER can learn effective emotional representations including prosodic features.

We note several possible areas of future work. It may be possible to reduce the weak label noise via filtering mechanisms, or by modifying prompts to include more conversational context, like the previous and next utterances, or scene descriptions. Additionally, using LLMs to generate weak labels in an open-set taxonomy may better leverage their expressiveness. Finally, while in this work we used ResNet-50 as our backbone model, higher capacity models like Conformers [9] might better capture the complex relationship between speech and emotion on the pre-training datasets we explored. We believe that the initial investigation and findings of this work provide valuable insights for future SER research on large-scale unlabeled data.

7. References

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[2] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[3] P. Liang *et al.*, "Holistic evaluation of language models," *CoRR*, vol. abs/2211.09110, 2022.

[4] M. Sap, H. Rashkin, D. Chen, R. L. Bras, and Y. Choi, "Social IQa: Commonsense reasoning about social interactions," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 2019, pp. 4462–4472.

[5] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 1666–1677.

[6] S. Shon, P. Brusco, J. Pan, K. J. Han, and S. Watanabe, "Leveraging Pre-Trained Language Model for Speech Sentiment Analysis," in *Proc. Interspeech 2021*, 2021, pp. 3420–3424.

[7] J. Zhao, R. Li, Q. Jin, X. Wang, and H. Li, "Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4703–4707.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[9] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3169–3173.

[10] S. Yang *et al.*, "SUPERB: speech processing universal performance benchmark," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*. ISCA, 2021, pp. 1194–1198.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[12] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3914–3923.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[14] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *NAACL*. Association for Computational Linguistics, 2018, pp. 1112–1122.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 195:1–195:35, 2023.

[16] S. Deng, L. Wu, G. Shi, L. Xing, and M. Jian, "Learning to compose diversified prompts for image emotion classification," *CoRR*, vol. abs/2201.10963, 2022.

[17] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 3816–3830.

[18] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[19] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[20] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the national academy of sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[21] A. Cowen *et al.*, "How emotion is experienced and expressed in multiple cultures: a large-scale experiment," 2021.

[22] D. Galvez, G. Diamos, J. M. C. Torres, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[23] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Tech. Rep., Technical report, OpenAI, Tech. Rep., 2022.

[25] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 2608–2618.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 770–778.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[29] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[30] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of EMNLP*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059.