INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

C^{2}A-SLU: Cross and Contrastive Attention for Improving ASR Robustness in Spoken Language Understanding

*Xuxin Cheng, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, Yuexian Zou**

ADSPLAB, School of ECE, Peking University, China

{chengxx, yaozy, zhihongzhu, ywl, lihongxiang}@stu.pku.edu.cn,
zouyx@pku.edu.cn

Abstract

Spoken language understanding (SLU) is a critical task in task-oriented dialogue systems. However, automatic speech recognition (ASR) errors often impair the understanding performance. Despite many previous models have obtained promising results for improving ASR robustness in SLU, most of them treat clean manual transcripts and ASR transcripts equally during the fine-tuning stage. To tackle this issue, in this paper, we propose a novel method termed C^{2}A-SLU. Specifically speaking, we add calculated cross attention to the original hidden states and apply contrastive attention to compare the input transcript with clean manual transcripts to distill the contrastive information, which can better capture distinctive features of ASR transcripts. Experiments on three datasets show that C^{2}A-SLU surpasses existing models and achieves a new state-of-the-art performance, with a relative improvement of 3.4% in terms of accuracy over the previous best model on SLURP dataset.

Index Terms: speech language understanding, ASR robustness, cross attention, contrastive attention

1. Introduction

Spoken language understanding (SLU) aims to extract semantic information from human speech inputs for typical subtasks [1, 2], mainly including intent detection and slot filling [3, 4, 5, 6]. There are two types of solutions for SLU, namely pipeline approaches and end-to-end approaches. The pipeline SLU methods combine automatic speech recognition (ASR) module and natural language understanding (NLU) module in a cascaded manner, allowing them to apply the external datasets and pre-trained models from natural language processing (NLP) community easily. However, pipeline approaches are susceptible to error propagation, where an imprecise ASR output can theoretically lead to errors in subtasks. As shown in Figure 1, "please turn up the speaker volume" is recognized as "please turn off the speaker volume" incorrectly, which leads to a wrong prediction.

As other tasks [7, 8, 9], enhancing learning representations to be resilient against errors has emerged as a highly effective approach to mitigate their negative impact, and it is gaining increasing attention. Following [10, 11], this paper aims to improve ASR robustness without using speech-related input features and we only concentrate on intent detection.

Although existing error-robust SLU models have made the promising progress, most of them treat clean manual transcripts and ASR transcripts as the same type without distinction in fine-tuning and simply combine them without discrimination, which limits the performance. In theory, clean manual transcripts and ASR transcripts can provide distinct types of knowledge, so the model fine-tuned on their combination is not capable of discriminating their specific contributions. The model trained on clean manual transcripts tends to achieve higher accuracy, whereas the model trained on ASR transcripts is usually more robust to ASR errors. Therefore, it is necessary to treat manual and ASR transcripts differently and leverage their respective characteristics to further improve the performance of the model.

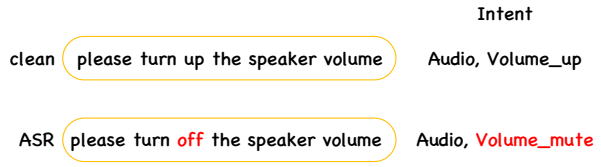

Figure 1: *Due to the ASR error, intent is predicted incorrectly.*

In this paper, we apply cross attention and contrastive attention to improve ASR robustness in SLU. Cross attention [12] is widely utilized to fuse multimodal features and has demonstrated its effectiveness in several pioneering works [13, 14, 15]. [14] adopts cross attention for open-vocabulary visual recognition. Inspired by them, we introduce cross attention to fuse features between clean manual transcripts and ASR transcripts to leverage their respective information. Contrastive attention is an effective method to detect abnormal information. [16] applies contrastive attention to selectively attend to relevant and irrelevant parts of the source sentence for abstractive sentence summarization. [17] uses contrastive attention to perform visual modeling and leverage relevant features. In our work, we utilize contrastive attention to capture unique features of ASR transcripts to make the model more robust to ASR errors.

Specifically, we concatenate the hidden states of clean manual transcripts and ASR transcripts and use the connected hidden states as the query (Q), key (K) and value (V) to compute the cross attention [12]. Then we add the calculated cross attention to the corresponding hidden states of the clean manual transcripts and the ASR transcripts as the processed hidden states. Contrastive attention consists of the aggregate attention and differentiate attention. In particular, the aggregate attention is responsible for identifying clean manual transcripts in the clean pool which are the most similar to the current input transcript. The differentiate attention is employed to extract the common information between the input transcript and the most similar transcripts and then subtract the common information from the input transcript to extract the distinctive properties between the input transcript and the clean manual transcripts.

Experiment results on three public benchmark SLU datasets SLURP, ATIS and TREC6 [10, 18, 19, 20] demonstrate that our C^{2}A-SLU significantly surpasses the previous best models and the model analysis further verifies the superiority of our method. The contributions of our work are three-fold:

- We propose a novel framework combining cross atten-

* Corresponding author.

10.21437/Interspeech.2023-93

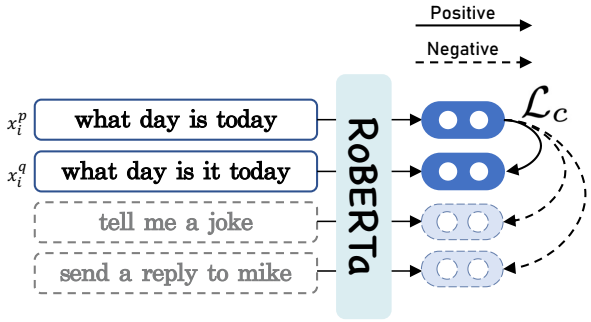

Figure 2: *The illustration of pre-training. We utilize self-supervised contrastive learning with the paired transcripts. A positive pair consists of clean data and the associated ASR transcript from the same audio.*

tion and contrastive attention to improve ASR robustness in the SLU task. To the best of our knowledge, we make the first attempt to apply contrastive attention to SLU, making it more robust to ASR errors.

- Experiment results demonstrate that C^{2}A-SLU is capable of dealing with the transcripts containing noise and achieves new state-of-the-art performance on three public benchmark SLU datasets.
- Model analysis shows that C^{2}A-SLU is able to align the ASR transcript and its associate clean manual transcript, which is helpful to further improve ASR robustness.

2. Method

Our proposed framework includes three elements: (1) Self-supervised contrastive learning (§2.1) in pre-training; (2) Cross attention (§2.2) in fine-tuning; (3) Contrastive attention (§2.3) in fine-tuning, consisting of aggregate attention (§2.3.1) and differentiate attention (§2.3.2).

2.1. Self-supvervised Contrastive Learning

Following [10], we utilize self-supervised contrastive learning in pre-training to learn sentence representations that are not susceptible to the misidentification. A pre-trained RoBERTa [21] is continually trained on spoken language corpus by utilizing the paired clean and noisy sentences.

Given a mini-batch of input data of N pairs of transcripts $B = \{(x_i^p, x_i^q)\}_{i=1..N}$, where x_i^p denotes a clean manual transcript and x_i^q denotes its associated ASR transcript. As shown in Figure 2, we first apply the pre-trained RoBERTa and utilize the last layer of `[CLS]` of RoBERTa to obtain the representation h_i^p for x_i^p and h_i^q for x_i^q:

$$h_i^p = \text{RoBERTa}(x_i^p) \tag{1}$$
$$h_i^q = \text{RoBERTa}(x_i^q) \tag{2}$$

Then we apply the proposed self-supervised contrastive loss $\mathcal{L}_c$ [22, 23] to adjust the sentence representations:

$$\begin{aligned}\mathcal{L}_c &= -\frac{1}{2N}\sum_{(h,h^+)\in P}\log\frac{e^{s(h,h^+)/\tau_c}}{\sum_{h'\neq h}^{B} e^{s(h,h')/\tau_c}} \\ &= -\mathbb{E}_P\Big[s(h,h^+)/\tau_c\Big] + \mathbb{E}\Big[\log\Big(\sum_{h'\neq h}^{B} e^{s(h,h')/\tau_c}\Big)\Big]\end{aligned} \tag{3}$$

where P is composed of $2N$ positive pairs of either (h_i^p, h_i^q) or (h_i^q, h_i^p), τ_c is the temperature hyper-parameter and $s(\cdot,\cdot)$ denotes the cosine similarity function. The first term of Equation 3 brings the clean manual transcript and its associated ASR transcript (positive example) near together and the second term pushes irrelevant ones (negative examples) far apart to promote uniformity in the representation space [24]. Note that for a transcript, its negative examples may be clean manual transcripts or ASR transcripts. For example, in Figure 2, "*tell me a joke*" is a clean manual transcript and "*send a reply to mike*" is an ASR transcript. They can both be the negative samples.

2.2. Cross Attention

To better leverage the respective characteristics of clean manual transcripts and ASR transcripts, we apply cross attention to process their hidden states. Suppose the hidden state size is d. We concatenate the hidden state $h_i^p \in \mathbb{R}^{1\times d}$ of the clean manual transcript x_i^p and the hidden state $h_i^q \in \mathbb{R}^{1\times d}$ of the associated ASR transcript x_i^q to obtain the connected hidden states h_i^c:

$$h_i^c = h_i^p \,||\, h_i^q \tag{4}$$

where $||$ denotes the concatenation operation.

We use $h_i^c \in \mathbb{R}^{2\times d}$ to obtain query matrix, key matrix and value matrix to compute cross attention O [12]:

$$O = \text{Softmax}\left(\frac{h_i^c W_1^Q W_1^{K\top} h_i^{c\top}}{\sqrt{d}}\right) h_i^c W_1^V \tag{5}$$

where $W_1^Q, W_1^K, W_1^V \in \mathbb{R}^{d\times d}$ are projection matrices. Then we add the cross attention O to the corresponding hidden states h_i^p and h_i^q to obtain the processed hidden states $h_i^{p,c}$ and $h_i^{q,c}$:

$$h_i^{p,c} = h_i^p + O_{[0:1],[0:d-1]} \tag{6}$$
$$h_i^{q,c} = h_i^q + O_{[1:2],[0:d-1]} \tag{7}$$

2.3. Contrastive Attention

As shown in Figure 3, we propose contrastive attention to enable our model to capture the differentiating properties between the input transcript and clean manual transcripts. Specifically, we first collect a clean pool $P = \{h_1^{p,c}, h_2^{p,c}, \ldots, h_{N_P}^{p,c}\}$ which consists of $N_P = 1,000$ clean manual transcripts randomly extracted from the training dataset. Contrastive attention is composed of aggregate attention (§2.3.1) and differentiate attention (§2.3.2), with the aim of obtaining the contrastive information between the input transcript and the clean pool P.

2.3.1. *Aggregate Attention*

As all transcripts in the pool P are clean, there is no inherent ranking or priority among them. As a result, it is natural to treat all the clean manual transcripts equally when capturing the contrastive information. However, some clean manual transcripts may have different grammatical structures from the input transcript, it is suboptimal to directly compare these clean manual transcripts with current input transcript. Therefore, we introduce the aggregate attention to assign higher weights to clean manual transcripts that are similar to the current input transcript, and lower weights to transcripts that are dissimilar to the input transcript. For the processed hidden state of input transcript h_c, the aggregate attention $\text{Agg}(h_c, P)$ of h_c and the pool P is:

$$\text{Agg}(h_c, P) = \text{Softmax}\left(\frac{h_c W_2^Q W_2^{K\top} P^\top}{\sqrt{d}}\right) P W_2^V \tag{8}$$

where $W_2^Q, W_2^K, W_2^V \in \mathbb{R}^{d\times d}$ are projection matrices. Moreover, inspired by [25], we repeat aggregate attention T times with different learnable weights to obtain the final output P' of aggregate attention to further improve the performance:

$$P' = [\text{Agg}(h_c, P)_1 \,||\, \text{Agg}(h_c, P)_2 \,||\, \ldots \,||\, \text{Agg}(h_c, P)_T] \tag{9}$$

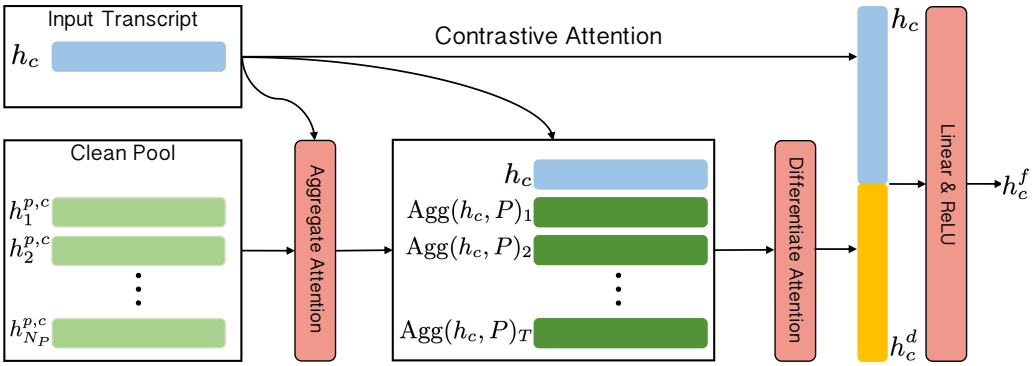

Figure 3: *The illustration of Contrastive Attention, which consists of the Aggregate Attention and Differentiate Attention. Aggregate attention is used to find the most similar clean manual transcripts to the current input transcript in the clean pool. Differential attention is implemented to summarize the common information between the input transcript and the most similar clean manual transcripts and subtracts this information from the input transcript to obtain the differential properties.*

where $\text{Agg}(h_c, P)_i$ denotes the output of i-th $\text{Agg}(h_c, P)$ and $||$ denotes the concatenation operation. By repeating T times, we can efficiently find the most similar clean manual transcripts P' to the input transcript h_c from T parts.

2.3.2. *Differentiate Attention*

To capture the contrastive information between the input transcript and the most similar clean manual transcripts, we introduce differentiate attention mechanism. We summarize the common information h'_c between the information of current input transcript h_c and the information of the most similar clean manual transcripts P' as follows:

$$h'_c = \text{Pooling}\left(\text{Softmax}\left(\frac{QW_3^Q W_3^{K\top} K^\top}{\sqrt{d}}\right) VW_3^V\right) \tag{10}$$

where Pooling denotes average pooling operation, Q, K, V are the concatenation of h_c and P', and $W_3^Q, W_3^K, W_3^V \in \mathbb{R}^{d\times d}$ are the projection matrices. Through such self-attention mechanism, we could utilize the similarity between h_c and P' to capture the significant common information.

Then, we subtract the common information h'_c from the input transcript h_c to obtain the contrastive information h_c^d:

$$h_c^d = h_c - h'_c \tag{11}$$

Finally, we utilize h_c and h_c^d to calculate the processed hidden state h_c^f and replace the original transcript h_c with h_c^f:

$$h_c^f = \text{ReLU}([h_c \,||\, h_c^d]W') \tag{12}$$

where $\text{ReLU}(\cdot)$ stands for the ReLU activation function [26] and W' is the matrix for linear transformation.

The final intent prediction p is calculated as follows:

$$p = \text{Softmax}(h_c^f W_I + b_I) \tag{13}$$

where W_I and b_I are weight matrix and bias vector.

2.4. Training Objective

2.4.1. *Pre-training*

It has been verified that adapting to unlabeled data can still improve the performance after domain-adaptive pre-training [27], so we continue training the masked language model (MLM) in pre-training. So the final proposed pre-training loss $\mathcal{L}_{pt}$ is the weighted sum of the self-supervised contrastive learning loss $\mathcal{L}_c$ and an MLM loss $\mathcal{L}_{mlm}$:

$$\mathcal{L}_{pt} = \lambda_{pt}\mathcal{L}_c + (1-\lambda_{pt})\cdot\mathcal{L}_{mlm} \tag{14}$$

where λ_{pt} is the coefficient balancing the two tasks.

2.4.2. *Fine-tuning*

Following [28, 29], the training objective in fine-tuning stage is the cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N_I} y_i \log p_i \tag{15}$$

where y_i denotes the golden intent label and N_I is the number of the intent labels.

3. Experiments

3.1. Datasets and Metrics

Following previous work [10], we conduct all the experiments on three publicly available benchmark datasets[1]: SLURP, ATIS and TREC6 [10, 18, 19, 20]. The statistics of the three datasets included are shown in Table 1.

Table 1: *The statistics of the three SLU datasets. The* `test` *set of SLURP dataset sub-sampled.*

Dataset	#Class	Avg. Length	Train	Test
SLURP	18×46	6.93	50,628	10,992
ATIS	22	11.14	4,978	893
TREC6	6	8.89	5,452	500

SLURP is a complex SLU dataset that contains various domains, speakers, and recording settings. The dataset consists of several (scenario, action) pairs and the prediction is considered correct only when both the scenario and action are predicted correctly. ATIS and TREC6 are two public SLU datasets for flight reservation and question classification, respectively. We utilize the synthesized text provided by Phoneme-BERT [30], where the data is first generated by a text-to-speech (TTS) model and then transcribed by an ASR model.

3.2. Experimental Setting

We compare our model with four baselines:

- `RoBERTa` [21]: a RoBERTa-base model directly fine-tuned on the target training data;

[1]SLURP is available at `https://github.com/MiuLab/SpokenCSE`, and ATIS and TREC6 are available at `https://github.com/Observeai-Research/Phoneme-BERT`.

Table 2: *Accuracy results on three SLU datasets. † denotes C²A-SLU obtains statistically significant improvements over baselines with $p < 0.01$.*

Model	SLURP	ATIS	TREC6
RoBERTa [21]	84.42	94.86	84.54
Phoneme-BERT [30]	84.16	95.14	86.48
SimCSE [23]	84.88	94.32	85.46
SpokenCSE [10]	85.64	95.58	86.82
C²A-SLU	**88.56**†	**96.84**†	**88.42**†

- `Phoneme-BERT` [30]: a RoBERTa-base model which is further pre-trained on an additional corpus with phoneme information. Phoneme sequences are generated from the ASR hypothesis via phonemizer[2];
- `SimCSE` [23]: a state-of-the-art sentence embedding method applying contrastive learning.
- `SpokenCSE` [10]: a strong baseline which applies self-supervised contrastive learning in pre-training and supervised contrastive learning in fine-tuning to improve ASR robustness in the SLU task.

We pre-train the model for 10K steps with a batch size 128 on each dataset, and finetune the whole model up to 10 epochs with a batch size 256 to avoid overfitting. The training process will early-stop if the loss on `dev` set does not decrease for 3 epochs. The mask ratio of MLM is set to 0.15, τ_c is set to 0.2, and λ_{pt} is set to 0.5. During both pre-training and fine-tuning, we use Adam optimizer [31] with $\beta_1 = 0.9, \beta_2 = 0.98$, and 4k warm-up updates to optimize parameters in our model. All experiments are conducted at an Nvidia Tesla-A100 GPU. The training process lasts a few hours.

3.3. Main Results

The performance comparison of C²A-SLU and baselines are shown in Table 2, from which we have following observations:

(1) Our C²A-SLU gains significant and consistent improvements on all tasks and datasets. Specifically, when using manual transcripts, it overpasses the previous state-of-the-art model SpokenCSE by 3.4%, 1.3%, and 1.8% on SLURP, ATIS and TREC6, respectively. This is because our model applies contrastive attention to extract the distinctive properties between the input transcript and the clean manual transcripts, which can further improve ASR robustness in SLU.

(2) In contrast, it is obvious that the improvement on SLURP dataset is more significant. A possible reason is that SLURP is a more challenging SLU dataset than ATIS and TREC6. An intent of SLURP is a (scenario, action) pair and the prediction is considered to be correct only if the scenario and action are both correctly predicted. Previous works treat the clean manual transcripts and ASR transcripts equally without discrimination during the fine-tuning stage, thus the ability to capture unique information of ASR transcripts is inadequate. As a result, due to inevitable ASR errors, it is common that one of the two components of an intent is incorrectly predicted. Our C²A-SLU aims to fully exploit the differences between ASR transcripts and clean manual transcripts, leading to the better performance of the model.

3.4. Analysis

3.4.1. Ablation Study

To verify the advantages of C²A-SLU from different perspectives, we conduct a set of ablation experiments. The experimental results are shown in Table 3. We can clearly observe that when contrastive learning in pre-training, cross attention and contrastive attention in fine-tuning are removed, the accuracy on three dataset all drops in varying degrees, which indicates that our method could indeed improve ASR robustness in SLU. In addition, it is obvious that when contrastive attention is removed, the degradation of accuracy is more significant. We believe the reason is that our method is able to understand the unique features of ASR transcripts more precisely via contrastive attention, and thus predict the intent more accurately.

[2] `https://github.com/bootphon/phonemizer`

Table 3: *Results of ablation experiments.*

Model	SLURP	ATIS	TREC6
C²A-SLU	**88.56**	**96.84**	**88.42**
w/o Contrastive Learning	87.92(↓0.64)	96.42(↓0.42)	87.88(↓0.54)
w/o Cross Attention	87.62(↓0.94)	96.26(↓0.58)	87.34(↓1.08)
w/o Contrastive Attention	86.81(↓1.75)	96.02(↓0.82)	87.16(↓1.26)

3.4.2. Visualization

To better understand how cross attention and contrastive attention affects and contributes to the final result, we show the visualization of an example in Figure 4. We visualize the representations by reducing the dimension with Principal Component Analysis (PCA) [32]. *"turn down volume twenty percent"* and *"i want to slow down my speaker"* are two manual transcripts with the same intent, and the representations of them and their associated ASR transcripts stay close to each other, which also demonstrates that our method can align the ASR transcript and its associate manual transcript with high accuracy.

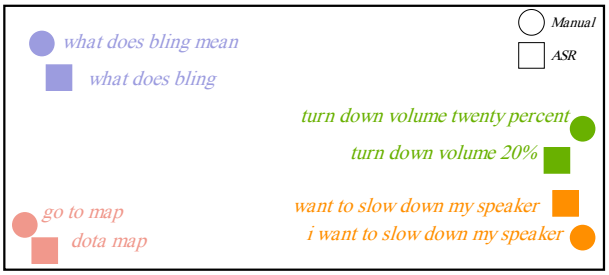

Figure 4: *Visualization of representations of clean manual transcripts and ASR transcripts. The circle and square in the same color means the corresponding clean manual transcript and ASR transcript are associated.*

4. Conclusions

In this paper, we propose C²A-SLU, a novel framework for improving ASR robustness in SLU. We apply cross attention to fuse features between clean manual transcripts and ASR transcripts and apply contrastive attention to capture unique features of ASR transcripts. Experiments and analysis on three public benchmark datasets demonstrate that C²A-SLU significantly outperforms the previous best models. In the future, we will explore how to further leverage the clean manual transcripts to improve ASR robustness in SLU.

5. Acknowledgements

We thank all the anonymous reviewers and program chairs for their valuable and insightful feedback. This paper was partially supported by Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001) and NSFC (No: 62176008) .

6. References

[1] Z. Zhu, X. Cheng, Z. Huang, D. Chen, and Y. Zou, "Towards unified spoken language understanding decoding via label-aware compact linguistics representations," in *Proc. of ACL Findings*, 2023.

[2] Z. Zhu, W. Xu, X. Cheng, T. Song, and Y. Zou, "A dynamic graph interactive framework with label-semantic injection for spoken language understanding," in *Proc. of ICASSP*, 2023.

[3] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[4] L. Qin, X. Xu, W. Che, and T. Liu, "Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1807–1816.

[5] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, "GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling," 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, 2021, pp. 178–188. [Online]. Available: https://aclanthology.org/2021.acl-long.15

[6] B. Xing and I. Tsang, "Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 159–169.

[7] X. Cheng, Q. Dong, F. Yue, T. Ko, M. Wang, and Y. Zou, "M 3 st: Mix at three levels for speech translation," in *Proc. of ICASSP*, 2023.

[8] Z. Zhu, X. Cheng, D. Chen, Z. Huang, H. Li, and Y. Zou, "Mix before Align: Towards Zero-shot Cross-lingual Sentiment Analysis via Soft-Mix and Multi-View Learning," in *Proc. of Interspeech*, 2023.

[9] X. Cheng, Z. Zhu, H. Li, Y. Li, and Y. Zou, "Ssvmr: Saliency-based self-training for video-music retrieval," in *Proc. of ICASSP*, 2023.

[10] Y.-H. Chang and Y.-N. Chen, "Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding," in *Proc. Interspeech 2022*, 2022, pp. 3458–3462.

[11] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, "Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," in *Proc. of ACL Findings*, 2023.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] O.-B. Mercea, L. Riesch, A. S. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 10 543–10 553.

[14] R. Qian, Y. Li, Z. Xu, M.-H. Yang, S. Belongie, and Y. Cui, "Multimodal open-vocabulary video classification via pre-trained vision and language models," *arXiv preprint arXiv:2207.07646*, 2022.

[15] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 269–280.

[16] X. Duan, H. Yu, M. Yin, M. Zhang, W. Luo, and Y. Zhang, "Contrastive attention mechanism for abstractive sentence summarization," 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, pp. 3044–3053.

[17] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2759–2767.

[18] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

[19] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[20] X. Li and D. Roth, "Learning question classifiers," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[21] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[23] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910.

[24] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.

[25] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *International Conference on Learning Representations*, 2017.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of ICML*, 2010.

[27] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.

[28] E. Haihong, P. Niu, Z. Chen, and M. Song, "A novel bi-directional interrelated model for joint intent detection and slot filling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5467–5471.

[29] D. Chen, Z. Huang, X. Wu, S. Ge, and Y. Zou, "Towards joint intent detection and slot filling via higher-order attention," in *IJCAI*, L. D. Raedt, Ed., 2022.

[30] M. N. Sundararaman, A. Kumar, and J. Vepa, "Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript," *arXiv preprint arXiv:2102.00804*, 2021.

[31] 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.

[32] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.