Interspeech 2022
18-22 September 2022, Incheon, Korea

Improving Spoken Language Understanding with Cross-Modal Contrastive Learning

Jingjing Dong[1*†] , *Jiayi Fu*[2*], *Peng Zhou*[2], *Hao Li*[2], *Xiaorui Wang*[2]

[1]Peking University, China
[2]Kuaishou Technology Co., Beijing, China

djj@stu.pku.edu.cn, {fujiayi, zhoupeng05, lihao25, wangxiaorui}@kuaishou.com

Abstract

Spoken language understanding(SLU) is conventionally based on pipeline architecture with error propagation issues. To mitigate this problem, end-to-end(E2E) models are proposed to directly map speech input to desired semantic outputs. Meanwhile, others try to leverage linguistic information in addition to acoustic information by adopting a multi-modal architecture. In this work, we propose a novel multi-modal SLU method, named CMCL, which utilizes cross-modal contrastive learning to learn better multi-modal representation. In particular, a two-stream multi-modal framework is designed, and a contrastive learning task is performed across speech and text representations. Moreover, CMCL employs a multi-modal shared classification task combined with a contrastive learning task to guide the learned representation to improve the performance on the intent classification task. We also investigate the efficacy of employing cross-modal contrastive learning during pretraining. CMCL achieves 99.69% and 92.50% accuracy on FSC and Smartlights datasets, respectively, outperforming state-of-the-art comparative methods. Also, performances only decrease by 0.32% and 2.8%, respectively, when trained on 10% and 1% of the FSC dataset, indicating its advantage under few-shot scenarios.

Index Terms: spoken language understanding, intent classification, multi-modal, contrastive learning, pretraining

1. Introduction

Spoken Language Understanding(SLU) which infers the semantic meaning of spoken utterances, is a critical component of spoken dialogue system [1–5]. The conventional SLU systems employ a cascaded approach consisting of two primary elements: an automatic speech recognition (ASR) module which decodes speech signals into transcripts; a natural language understanding (NLU) module that infers semantic information from ASR text output [6–9]. However, the cascaded approach has a few limitations. First, cascaded modules can lead to a propagation of error. Second, it loses rich prosodic information in speech signals such as speech rate, pitch and intonation.

To address the above limitations, end-to-end(E2E) SLU methods have been proposed, where the features extracted from the speech signal are directly mapped to intents or other SLU outputs. Serdyuk et al. [10] pass speech signals directly through a bi-directional GRU network for real-time intent classification. Lugosch et al. [11] integrate a pre-trained speech model into an end-to-end model on the supervised SLU task. Radfar et al. [12] introduce a pure transformer encoder-decoder model adaptable to predict slots and intentions with variable-length.

Another solution is to make use of the linguistic information besides the acoustic information from spoken utterances. Price et al. [13] concatenate the fixed embeddings output of an E2E SLU model and an NLU model to train a shared classifier. Chen et al. [14] propose a joint textual-phonetic pre-training approach for learning spoken language representations. Sharma et al. [15] employ a cross-modal attention layer to fuse both acoustic and linguistic embeddings for intent classification. Agrawal et al. [16] use pre-trained BERT model to learn speech and text embeddings in a latent space.

Some works explore possibilities of leveraging text information in E2E frameworks. Kim et al. [17] take phoneme posterior and sub-word-level text as input in the pretraining stage while finetuning solely on the phoneme representation. Jiang et al. [18] utilize text transcripts to distill the speech transformer and indicate an improvement for intent classification. As SLU tasks require datasets of large amounts and high quality, modality shortage and data scarcity are unignorable obstacles to better performances of various SLU models. To tackle this problem, Cha et al. [19] propose the flexible inputs model, which can take either speech signals or ASR transcripts as inputs.

Also, works have shown that aligning speech and text embedding with distance based approaches can lead to better performances [20, 21]. As for align methods, we seek for unsupervised contrastive learning methods with limited SLU data resources. Existing works concerning multi-modal contrastive learning are primarily focused on Vision-Language(VL) tasks. Harwath et al. [22] match visual objects and spoken words using triplet loss. Pielawski et al. [23] learn shared image representations with a contrastive loss based on noise-contrastive estimation (InfoNCE). Radford et al. [24] use batch construction loss to pretrain an image-text model, which improves the performance of zero-shot transfer learning.

In this work, we propose a novel multi-modal SLU method, named CMCL, which utilizes cross-modal contrastive learning approach to learn better multi-modal representation. CMCL employs a two-stream multi-modal framework and combines a multi-modal shared classification task with a cross-modal contrastive learning task to guide the learned representation to improve the performance on the intent classification task. Moreover, CMCL also adopts cross-modal contrastive learning objective during the pretraining stage to further improve the performance. Our contributions are summarized in three folds:

- We propose a novel multi-modal SLU method named CMCL, which utilizes cross-modal contrastive learning objectives to learn better multi-modal representations.
- We demonstrate the effectiveness of the cross-modal contrastive learning method under both full-data and few-shot settings and give evidence that combining multi-modal shared classification and pretraining can

[*]Equal contribution.
[†]This work was done when author was on an internship at Kuaishou.

10.21437/Interspeech.2022-658

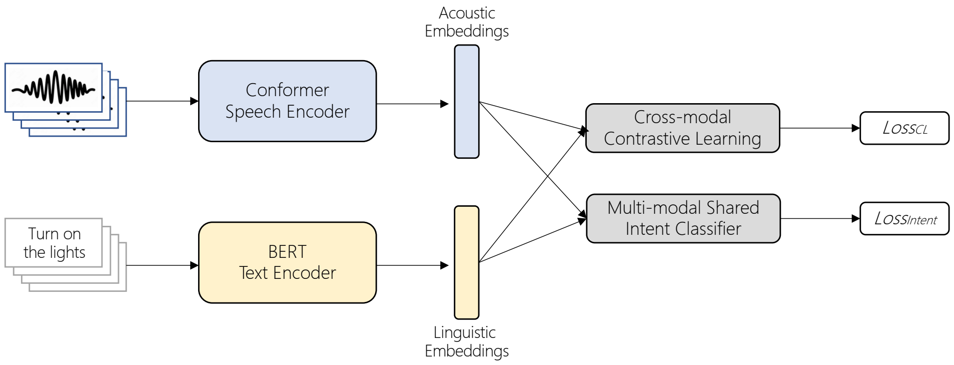

Figure 1: *The overview of proposed CMCL method.*

further improve its performance.

- The CMCL method outperforms state-of-the-art comparative methods on FSC and Smartlights datasets and its performance only slightly decreased under few-shot scenarios.

2. Proposed Method

In this section, we discuss the proposed CMCL method. Figure 1 shows the method's overview, and it mainly consists of four components, including speech encoder, text encoder, cross-modal contrastive learning and multi-modal shared intent classifier. Here, we use the encoder part of Conformer model [25] as the speech encoder, BERT model [26] as text encoder.

Initially, we extract the acoustic embeddings from the speech encoder and linguistic embeddings from the text encoder. Then the acoustic and linguistic embeddings are simultaneously aligned through cross-modal contrastive learning and fed into an intent classifier to predict the intent labels.

The model is optimized with two losses: contrastive learning loss from multi-modal embeddings and intent classification loss from the predictions and ground truths. Further details of each component of CMCL are as follows.

2.1. Conformer Speech Encoder

Conformer [25] is a convolution-augmented transformer model and has achieved outstanding performance in many speech-related deep learning tasks. In CMCL, the Conformer model consists of a convolution sub-sampling layer and several Conformer layers. Taking the computed speech features as input, the model max-pools its layer outputs across the time dimension and outputs a fixed-size vector as an acoustic embedding of the speech.

2.2. BERT Text Encoder

BERT [26] achieves state-of-the-art performance in sentence classification and other natural language understanding tasks. We follow the widely used setting in many SLU works [15, 16, 19], which use a pre-trained BERT[1] model as the text encoder. The text encoder takes either ground truth or ASR transcripts to output linguistic embeddings. As is common in BERT-based encoders, the last layer representation of the [CLS] token is used as the linguistic embedding of the utterance.

2.3. Cross-modal contrastive Learning

Cross-modal contrastive learning aims to learn better multi-modal representations from the input speech-text pairs. Inspired by CLIP [24], which uses an efficient unsupervised contrastive learning method using in-batch correct pairs mined through cosine similarity, we employ infoNCE [27] loss to align the acoustic and linguistic embeddings to uni-modal latent space.

Here, S and T denote the input batch of speech and text data, where the batch size is N. We first collect the acoustic and linguistic embeddings from the speech and text encoder, which are S_E and T_E respectively:

$$\begin{aligned} S_E &= SpeechEncoder(S), \\ T_E &= TextEncoder(T) \end{aligned} \tag{1}$$

, where $S_E \in \mathbb{R}^{N\times S_{dmodel}}$, $T_E \in \mathbb{R}^{N\times T_{dmodel}}$.

Next, a similarity matrix $A \in \mathbb{R}^{N\times N}$ as shown in Figure 2 is computed based on these embeddings and $A_{i,j}$ quantifies how similar are embeddings of speech sample S_i and text sample T_j:

$$A = Sim(S_E W_E, T_E) = \frac{S_E W_E \cdot T_E}{\|S_E W_E\| \cdot \|T_E\|} \tag{2}$$

, where $A \in \mathbb{R}^{N\times N}$. Here $W_E \in \mathbb{R}^{S_{dmodel}\times T_{dmodel}}$ is a learnable linear transformation, which converts acoustic embeddings into the same dimension as the linguistic embeddings.

Based on the similarity matrix A, we symmetrically calculate speech to text contrastive loss $Loss_{S2T}$ and text to speech contrastive loss $Loss_{T2S}$. Specifically:

$$\begin{aligned} Loss_{S2T} &= -\frac{1}{N}\left(\sum_{i=1}^{N} log\frac{e^{A_{i,i}}}{\sum_j e^{A_{i,j}}}\right), \\ Loss_{T2S} &= -\frac{1}{N}\left(\sum_{j=1}^{N} log\frac{e^{A_{j,j}}}{\sum_i e^{A_{i,j}}}\right), \\ Loss_{CL} &= \frac{1}{2}(Loss_{S2T} + Loss_{T2S}). \end{aligned} \tag{3}$$

The cross-modal contrastive loss($Loss_{CL}$) is the average cross-entropy loss on the rows and columns of the similarity matrix, where the diagonal entries are treated as correct classes. In contrast, other entries are treated as incorrect classes.

As $Loss_{CL}$ is minimized, the model learns multi-modal embedding in a latent space by jointly training a speech encoder and text encoder to maximize the cosine similarity of the speech

[1] https://huggingface.co/bert-base-uncased

and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings.

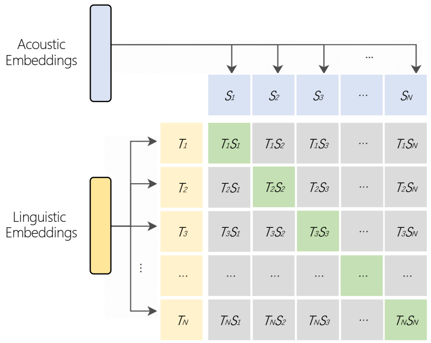
Acoustic Embeddings
Linguistic Embeddings
S1 S2 S3 ··· SN
T1 T1S1 T1S2 T1S3 ··· T1SN
T2 T2S1 T2S2 T2S3 ··· T2SN
T3 T3S1 T3S2 T3S3 ··· T3SN
··· ··· ··· ··· ··· ···
TN TNS1 TNS2 TNS3 ··· TNSN

Figure 2: *The cross-modal contrastive learning jointly trains the speech encoder and text encoder to predict the correct pairings of a batch of (speech, text) training examples.*

2.4. Multi-modal Shared Intent Classifier

We use a simple multi-layer feed-forward neural network for intent classification concerning the intent classification objective. To improve the generalization ability with unseen data, speech and text encoders share the weights of the intent classifier. In this way, the classifier is trained to perform predictions on either linguistic or acoustic embeddings. The output dimension varies according to the number of unique labels in different datasets.

The intent classification loss of two modalities is referred to as $Loss_{TI}$ and $Loss_{SI}$, which are both computed by cross-entropy calculation. The intent loss is calculated as follows:

$$\begin{aligned} Loss_{TI} &= CrossEntropy(Pred_{Text}, y), \\ Loss_{SI} &= CrossEntropy(Pred_{Speech}, y), \\ Loss_{Intent} &= Loss_{TI} + Loss_{SI} \end{aligned} \tag{4}$$

, where $Pred_{Text}$ and the $Pred_{Speech}$ are the predictions from the intent layer of each encoder, and y is the ground truth label.

The total loss ($Loss_{total}$) is further used to back-propagate the whole model:

$$Loss_{total} = Loss_{Intent} + Loss_{CL}. \tag{5}$$

3. Experiments

3.1. Datasets

We evaluate our CMCL on the two most commonly used SLU datasets; the details of used datasets are presented in Table 1.

FSC Dataset The utterances in the Fluent speech commands (FSC) dataset [10] includes speech commands of a virtual assistant. Each command consists of three slot values: action, object, and location. The combination of three slot values represents the intent label.

SmartLights Dataset The SmartLights [28] is a close-field subset of the Snips SLU dataset, which contains English spoken commands related to smart light appliances.

Data Processing We extract 80-dimensional F-Bank features for each wave file. Since the two datasets only contain ground truth transcripts, we use the ASR transcripts released by previous works, specifically with a Kaldi-trained ASR model as released in [19]. The WER of the first ASR transcripts for Smartlights and FSC data are 34.1% and 25.6%, respectively. To further examine the effects of different word error rates (WER) of ASR outputs, we use Google cloud system adopted by [29] to produce FSC ASR results whose WER is 6.5%.

Table 1: *Data statistics*

Datasets	Train	Valid	Test	Intents
FSC	23,132	3,118	3,793	31
SmartLights	1,328	166	166	7

3.2. Compared Methods

In order to demonstrate the effectiveness of the proposed method, we first develop three baseline methods. The first two methods only use a single modal encoder and classifier. The third baseline method uses multi-modal encoder and fed the concatenated acoustic and linguistic embeddings into the classifier. Specially:

- **Speech-Only** is of the same structure as our speech encoder, which is a 3-layer Conformer model, followed by a fully-connect layer for intent classification.
- **Text-Only** is a $BERT_{base-cased}$ model fine-tuned on intent classification task with only ASR results as input.
- **Multi-Modal** concatenates the two embeddings of both BERT text encoder and the Conformer speech encoder for inference stage.

The results are also compared with the previous representative SLU methods, including end-to-end and multi-modal methods. Multi-modal methods include CMLM (Agrawal et al., 2020 [16]), CMLMD(Cho et al. [30]) and (Cha et al., 2021 [19]). End-to-end methods include pre-trained E2E model(Lugosch et al., 2019 [11]), ST-BERT (Kim et al., 2021 [17]) and VQ-BERT+DS2(Kim et al., 2021 [31]). For a fair comparison, we exclude the model performance of a data augmentation setting.

3.3. Implementation

Training In the training stage, we train a 3-layer Conformer speech encoder and a 12-layer BERT text encoder under contrastive learning and classification objective. After acquiring a 512- and 768-dimension representation from the last-layer output of the speech and text encoder, we pass the 512-dimension output of the speech encoder through a fully-connected layer to match the 768-dimension embedding of the text encoder. Afterwards, we either concatenate or individually feed these embeddings into a shared classifier. The shared classification layer has an input size of 768 and the number of outputs depends on the intent number of the dataset. As for optimization techniques, we use Adam [32] optimizer with linear-decay learning rate schedule with a peak learning rate of 5e-5 for text-encoder, 1e-3 for speech-encoder and other parameters. We set the batch size to 16, the max sequences length to 100 and the number of the epoch to 200 for actual training.

Pre-training Since the contrastive learning task in the proposed method does not rely on task-specific supervised labels, we can use speech-text pair data to do model pretraining. We co-pretrain both the Conformer speech encoder and BERT text encoder on the train-clean-100 subset of LibriSpeech [33]. The other subsets in LibriSpeech are also tested to be added into pretraining data, but the experimental results did not show valid performance improvement. We adopt a single cross-modal contrastive learning objective for pre-training to evaluate its effect. The model is pretrained about 16K steps with the same batch size and maximum input sequence length as in the training stage.

Inference During inference, we use a prediction combination to leverage the outputs of the multi-modal intent classifier as shown in Figure 1. This combined prediction is an element-wise average of the intent classifier's outputs from the acoustic and linguistic embeddings.

Few-shot Scenario To examine the robustness of method performance to varying training data size, we test our model with small amounts of data. In particular, we randomly sample 10% subsets and 1% subsets of the FSC dataset as the train set and evaluate the model on the complete test set. To ensure the confidence of the results, we repeated the random sampling and model training five times. The reported experimental results are the averaged score.

Table 2: *Test accuracy(%) on the FSC and SmartLights dataset.*

	Model	FSC	SmartLights
Baselines	Speech Only	98.07	75.00
	Text Only	98.94	88.12
	Multi-modal	99.24	88.75
End2End	Lugosch et al. [11]	98.80	-
	Kim et al. [17].	99.50	84.65
	Kim et al. [31]	99.60	92.20
Multi-modal	Agrawal et al. [16]	97.65	74.10
	Cho et al. [30]	98.98	-
	Cha et al. [19]	99.18	89.76
	CMCL	**99.69**	**92.50**

3.4. Results

Main Results In Table 2, we present the intent classification accuracy on two datasets for the three baselines methods, multi-modal SLU methods, end-to-end methods and the proposed CMCL method.It can be found that Multi-modal baseline gives better results than either Text-Only or the Speech-Only baseline, indicating that using multi-modal information as inputs is better than using uni-modal ones. Also, it can be observed that compared with the multi-modal baseline, the CMCL method shows 0.45% and 3.75% improvement on FSC and Smartlights datasets, respectively, proving its effectiveness.

The fact that the CMCL model achieves 99.69% and 92.5%, which are state-of-the-art performances under both datasets compared to previous works, proves that it has produced better representations for the intent classification task.

Few-shot Scenario Table 3 reveals our model's performance on data shortage scenarios. Comparatively marginal performance degradation of 0.32% and 2.8% respectively can be observed on the 10% and 1% data shortage scenarios. Especially when only 1% of data are given, our model exceeds the end-to-end model proposed by Kim et al. [17] by 1.25%. This result demonstrates that our two-stream multi-modal framework with shared classification task is effective when extremely limited downstream task data are available, which is common in reality.

Ablation Studies Moreover, the results in Table 4 disclose that, in the ablation experiments, each of the three components, including cross-modal contrastive learning, multi-modal shared classification and pre-training, has more or less contributed to improving our multi-modal baseline and is indispensable to the overall performance of CMCL model. Results in column four and five further confirm above conclusions under few-shot scenarios. It supports our hypothesis that contrastive learning and the multi-modal intent classification objectives have assisted both modalities in gaining complementary information from each other.

Additionally, it can be concluded from the second and third column of Table 4 that, as the ASR WER of the FSC dataset decreased from 34.1% to 6.5%, the accuracy of our model at all settings increased accordingly from the range (98.31%-99.60%) to (99.24%-99.69%), supporting that our model can fair better accuracy using better ASR results.

Table 3: *Test accuracy(%) on the FSC dataset.10% and 1% denotes the data shortage scenarios where 10% and 1% of the training data are given for training.*

Model	Full	10%	1%
Lugosch et al. [11]	98.80	97.96	82.78
Cho et al. [30]	98.98	98.12	83.12
Kim et al. [17]	99.50	99.13	95.64
CMCL	**99.69**	**99.37**	**96.89**

Table 4: *Ablation accuracy(%) on Smartlights, FSC with 34.1%WER(FSC1), FSC with 6.5% WER(FSC2) and FSC2 under 1% and 10% few-shot scenarios. CL, MSC and PT denote contrastive learning, multi-modal shared classification and pre-training respectively.*

Model	Smart Lights	FSC1 Full	FSC2 Full	FSC2 10%	FSC2 1%
Multi-modal	88.75	98.31	99.24	96.89	94.46
+CL	91.25	98.93	99.27	97.39	94.79
+MSC	91.88	99.39	99.45	98.97	95.74
+PT	**92.50**	**99.60**	**99.69**	**99.37**	**96.89**

4. Conclusion

In this paper, we propose a two-stream framework with a cross-modal contrastive learning objective, aiming to acquire better multi-modal representations robust for downstream tasks. Results from randomised trials have proven our model's advantage on different datasets, each model component's validness, and our model's robustness on data-shortage occasions. We leave the extension of our method to other downstream tasks such as speech emotion recognition and spoken question answering as future work.

5. References

[1] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai, "A form-based dialogue manager for spoken language applications," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2. IEEE, 1996, pp. 701–704.

[2] D. Suendermann and R. Pieraccini, "Slu in commercial and research spoken dialogue systems," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 171–194, 2011.

[3] G. Skantze, C. Oertel, and A. Hjalmarsson, "User feedback in human-robot dialogue: Task progression and uncertainty," in *the HRI Workshop on Timing in Human-Robot Interaction, Bielefeld, Germany, March 3-6, 2014*, 2014.

[4] E. Iosif, I. Klasinas, G. Athanasopoulou, E. Palogiannidi, S. Georgiladakis, K. Louka, and A. Potamianos, "Speech understanding for spoken dialogue systems: From corpus harvesting to grammar rule induction," *Computer Speech & Language*, vol. 47, pp. 272–297, 2018.

[5] Y.-N. Chen, A. Celikyilmaz, and D. Hakkani-Tur, "Deep learning for dialogue systems," in *Proceedings of the 27th international conference on computational linguistics: Tutorial abstracts*, 2018, pp. 25–31.

[6] V. Goel, H.-K. Kuo, S. Deligne, and C. Wu, "Language model estimation for optimizing end-to-end performance of a natural language call routing system," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–565.

[7] S. Yaman, L. Deng, D. Yu, Y.-Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1207–1214, 2008.

[8] C.-J. Lee, S.-K. Jung, K.-D. Kim, D.-H. Lee, and G. G.-B. Lee, "Recent approaches to dialog management for spoken dialog systems," *Journal of Computing Science and Engineering*, vol. 4, no. 1, pp. 1–22, 2010.

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

[10] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5754–5758.

[11] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[12] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-end neural transformer based spoken language understanding," *arXiv preprint arXiv:2008.10984*, 2020.

[13] R. Price, M. Mehrabani, and S. Bangalore, "Improved end-to-end spoken utterance classification with a self-attention acoustic classifier," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8504–8508.

[14] Q. Chen, W. Wang, and Q. Zhang, "Pre-training for spoken language understanding with joint textual and phonetic representation learning," *arXiv preprint arXiv:2104.10357*, 2021.

[15] B. Sharma, M. Madhavi, and H. Li, "Leveraging acoustic and linguistic embeddings from pretrained speech and language models for intent classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7498–7502.

[16] B. Agrawal, M. Müller, M. Radfar, S. Choudhary, A. Mouchtaris, and S. Kunzmann, "Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding," *arXiv preprint arXiv:2011.09044*, 2020.

[17] M. Kim, G. Kim, S.-W. Lee, and J.-W. Ha, "St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7478–7482.

[18] Y. Jiang, B. Sharma, M. Madhavi, and H. Li, "Knowledge distillation from bert transformer to speech transformer for intent classification," *arXiv preprint arXiv:2108.02598*, 2021.

[19] S. Cha, W. Hou, H. Jung, M. Phung, M. Picheny, H.-K. Kuo, S. Thomas, and E. Morais, "Speak or chat with me: End-to-end spoken language understanding system with flexible inputs," *arXiv preprint arXiv:2104.05752*, 2021.

[20] P. Denisov and N. T. Vu, "Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning," *arXiv preprint arXiv:2007.01836*, 2020.

[21] Y.-A. Chung, C. Zhu, and M. Zeng, "Splat: Speech-language joint pre-training for spoken language understanding," *arXiv preprint arXiv:2010.02295*, 2020.

[22] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 649–665.

[23] N. Pielawski, E. Wetzer, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, and N. Sladoje, "Comir: Contrastive multimodal image representation for registration," *Advances in neural information processing systems*, vol. 33, pp. 18 433–18 444, 2020.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[28] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[29] S. Arora, A. Ostapenko, V. Viswanathan, S. Dalmia, F. Metze, S. Watanabe, and A. W. Black, "Rethinking end-to-end evaluation of decomposable tasks: A case study on spoken language understanding," *arXiv preprint arXiv:2106.15065*, 2021.

[30] W. I. Cho, D. Kwak, J. W. Yoon, and N. S. Kim, "Speech to text adaptation: Towards an efficient cross-modal distillation," *arXiv preprint arXiv:2005.08213*, 2020.

[31] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," *arXiv preprint arXiv:2010.13105*, 2020.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.