Cross-modal Transfer Learning via Multi-grained Alignment for End-to-End Spoken Language Understanding

Yi Zhu[1,2†], *Zexun Wang*[1†], *Hang Liu*[1], *Peiying Wang*[1], *Mingchao Feng*[1], *Meng Chen*[1], *Xiaodong He*[1]

[1]JD AI, Beijing, China, [2]LTL, University of Cambridge

yz568@cam.ac.uk, wangzexun3@jd.com, liuhang55@jd.com, wangpeiying3@jd.com, fengmingchao@jd.com, chenmeng20@jd.com, xiaodong.he@jd.com

Abstract

End-to-end spoken language understanding (E2E-SLU) has witnessed impressive improvements through cross-modal (text-to-audio) transfer learning. However, current methods mostly focus on coarse-grained sequence-level text-to-audio knowledge transfer with simple loss, and neglecting the fine-grained temporal alignment between the two modalities. In this work, we propose a novel multi-grained cross-modal transfer learning framework for E2E-SLU. Specifically, we devise a cross attention module to align the tokens of text with the frame features of speech, encouraging the model to target at the salient acoustic features attended to each token during transferring the semantic information. We also leverage contrastive learning to facilitate cross-modal representation learning in sentence level. Finally, we explore various data augmentation methods to mitigate the deficiency of large amount of labelled data for the training of E2E-SLU. Extensive experiments are conducted on both English and Chinese SLU datasets to verify the effectiveness of our proposed approach. Experimental results and detailed analyses demonstrate the superiority and competitiveness of our model.

Index Terms: spoken language understanding, cross-modal transfer learning, cross attention, contrastive learning

1. Introduction

Spoken language understanding (SLU) aims to comprehend user query given spoken utterance, so that dialogue systems can respond properly based on the query intent. Conventional SLU systems rely on the pipeline method consisting of two steps: 1) automatic speech recognition (ASR) model to convert audio signals from speech into text, and 2) natural language understanding (NLU) model to predict the intent from ASR output text. In comparison, end-to-end (E2E) SLU considers a single model that produces result directly from input audio without intermediate text output from ASR, which not only avoids error propagation from ASR to NLU, but also preserves useful information in speech signals such as prosody, pitch, and speech rate that are lost after ASR. Therefore, E2E-SLU has become increasingly popular and many efforts have been taken to catch up on the performance of the pipeline method [1–11].

Nevertheless, because of its high complexity with information of many aspects contained inside, it is generally difficult for raw audio signal to extract suitable linguistic features to solve language understanding tasks [1, 2]. In this respect, a lot of work has attempted to resort to text form, a more compressed meaning representation, by leveraging knowledge distillation [12, 13] to transfer textual knowledge of the teacher model, i.e. NLU model trained on text data, to the student E2E-SLU model [7–11]. Specifically, these approaches mostly focus on coarse-grained transfer with the audio-text pair: aligning the sequence-level audio representations produced by E2E-SLU model with the fixed text representations of NLU model mainly through minimising L1 [8,9] or L2 loss [7,10,11]. However, simple distance losses only refine single point pairs in the two representation spaces without harnessing the relationships of those pairs, thus incapable of aligning the two spaces in distributional perspective due to the inherent disparities between the two modalities. Furthermore, explicit alignment between audio frames and text tokens are neglected, which poses a great challenge for the text-to-audio transfer for SLU. Since the apprehension about user intent usually rests on certain keywords like entities, such fine-grained cross-modal alignment will be crucial in allowing E2E-SLU model to concentrate on more important audio frames for effective utterance interpretation.

In this work, we propose a novel Multi-grained Alignment Transfer Learning framework (**MATL**) for E2E-SLU. MATL is derived from the standard knowledge distillation approaches [7, 14], where the parameters of E2E-SLU model with speech input are updated according to an already trained NLU model when fed with the corresponding text. Concretely, we devise a cross attention (CA) module [15–17] for fine-grained alignment between the tokens of text and the frame features of speech, encouraging the E2E-SLU model to target at the salient acoustic features attended to each token during transferring the semantic information. To facilitate cross-modal representation learning in sequence level, we leverage contrastive learning (CL) [18–20] to pull together true (positive) text-audio pair representations among randomly sampled negative pairs. Additionally, we explore various data augmentation (DA) methods within MATL to mitigate the data scarcity issue of text-audio data. Extensive experiments are conducted on two SLU datasets with different languages: 1) the public English SLURP [21] and 2) our newly built in-house dataset JDTEL. Experimental results demonstrate that MATL outperforms the strong baselines in E2E-SLU (+2.29 on SLURP and +1.12 on JDTEL in accuracy). Detailed analyses show that both CA and CL are instrumental to the success of cross-modal transfer through text-audio alignment in both granularities. Moreover, the performance of MATL can be further boosted by different DA methods, closing the gap to the oracle NLU model, advancing towards serviceable E2E-SLU system in the real world.

2. Methodology

The overview architecture of our MATL framework is illustrated in Figure 1. MATL consists of a text-based NLU model (left green part), and an audio-based E2E-SLU model (right blue part). It follows the standard knowledge distillation procedure [12,13] which has been proven effective in both text and audio domains [7, 14]. The NLU model, a BERT-based Trans-

[†] Equal contribution. Work done during internship at JD AI.

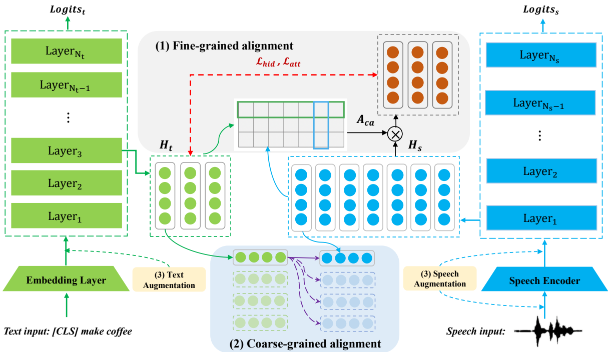


Figure 1: The architecture of our proposed model MATL.

former [22, 23] of layer number L_t, is finetuned first on the *labelled* text NLU data $\{\tilde{t}, y_t\}_1^{n_t}$ of n_t training instances, and fixed as the teacher model. It then transfers the textual knowledge to the student E2E-SLU model, a more shallow BERT with L_s layers on top of a speech encoder with waveform input, by instructing the student model to mimic the representations of various layers from the teacher model through *unlabelled* text-speech pairs $\{t, s\}_1^n$. Finally, the trained E2E-SLU model is evaluated on *labelled* speech SLU data $\{\tilde{s}, y_s\}_1^{n_s}$.

2.1. Fine-grained token-frame alignment with CA

For both BERTs in the NLU and E2E-SLU models, each layer i will output a hidden representation sequence $\mathbf{H}^i \in \mathbb{R}^{l \times d}$ of length l and dimensionality d, where each representation contains the contextual information for the specific step (token for NLU, frame for E2E-SLU), through multi-head self-attention with N_h attention heads, each being $\mathbf{A}^{i,n_h} \in \mathbb{R}^{l \times l}$ [22].To enable such fine-grained information transfer from text to audio, we adopt the previous approaches based on TinyBERT [7, 14] to fit the hidden representations and attentions of each layer in the student model to the certain layers of teacher model:

$$\begin{aligned} \mathcal{L}_{\text{hid}} &= \sum_{i}^{L_s} \mathcal{L}_h(\mathbf{H}_t^{g(i)}, \mathbf{H}_s^i) \\ \mathcal{L}_{\text{att}} &= \sum_{i}^{L_s} \sum_{n_h}^{N_h} \mathcal{L}_a(\mathbf{A}_t^{g(i),n_h}, \mathbf{A}_s^{i,n_h}) \end{aligned} \tag{1}$$

where $\mathbf{H}_t/\mathbf{H}_s$ and $\mathbf{A}_t/\mathbf{A}_s$ are the hidden and attention representations for text/speech, $\mathcal{L}_h/\mathcal{L}_a$ are the distillation losses for each hidden layer and attention head. $g(\cdot)$ is the same function in [14] to map each student layer to a particular teacher layer.

Due to the inherent disparities between both modalities, the sequence length of speech (l_s) is generally longer than that of text (l_t) in both $\mathbf{H}_t/\mathbf{H}_s$ and $\mathbf{A}_t/\mathbf{A}_s$, with $\mathbf{H}_t \in \mathbb{R}^{l_t \times d}$ and $\mathbf{A}_t \in \mathbb{R}^{N_h \times l_t \times l_t}$ (similarly for $\mathbf{H}_s$ and $\mathbf{A}_s$). Thus, it is not feasible to directly apply loss functions between two representations of different lengths, and simply padding text token representations to the same length of speech is not optimal either, as it will cause the frame representations in audio to be aligned to the meaningless padding representations in text.

To this end, we introduce the cross attention (CA) mechanism to capture the fine-grained interactions between text tokens and speech frames, which has been previously validated in token-phoneme scenario [16]. Specifically, for $\mathbf{H}_s$ of layer i and its corresponding teacher layer $\mathbf{H}_t$ (we omit the superscript i and $g(i)$ for simplicity), CA is formulated as the correlation map between the hidden representations of tokens and frames followed by a convolution kernel $\mathbf{K} \in \mathbb{R}^{l_s \times l_s}$ with normalisation over the speech sequence:

Dataset	#Intent	Train	Dev	Test	Avg_dur (s)
SLURP	828	11514	2033	2974	2.96
JDTEL	16	4314	1453	1433	3.13

Table 1: Dataset statistics. Avg_dur is average audio duration.

$$\mathbf{A}_{ca} = \text{Softmax}(\mathbf{H}_t \mathbf{H}_s^T \mathbf{K}), \quad \sum_{k}^{l_s} \mathbf{A}_{ca}(j,k) = 1 \tag{2}$$

where each row $A_{ca}(j, \cdot)$ is the attention vector for the token j over the whole speech sequence. Then the aligned hidden audio representation is obtained by weighting H_s with A_{ca}, and the student hidden layer distillation loss $\mathcal{L}_h$ can be calculated by minimising the mean squared error between the text and aligned speech hidden representations:

$$\mathcal{L}_h = \text{MSE}(\mathbf{H}_t, \mathbf{A}_{ca}\mathbf{H}_s) \tag{3}$$

To preform the distillation on each attention head, the audio attention $\mathbf{A}_s$ is transformed accordingly:

$$\mathcal{L}_a = \text{MSE}(\mathbf{A}_t, \mathbf{A}_{ca}\mathbf{A}_s\mathbf{A}_{ca}^T) \tag{4}$$

2.2. Coarse-grained sequence alignment via CL

The mainstream distillation methods on the sequence-level representations for BERT, i.e. the [CLS] token representations, are for student model to imitate teacher model with L1 [8, 9] or L2 loss [7, 10, 11] on a single text-audio pair at output layer. This does not provide strong teacher signal and cannot capture the difference in semantic relations among multiple pairs. Inspired by recent advance in contrastive learning (CL) for cross-modal transfer [24–26], we introduce a cross-modal contrastive loss within a mini-batch $\mathcal{B}$ of text-audio pairs to transfer the sequence-level semantic information from text to audio:

$$\mathcal{L}_{\text{cl}} = -\sum_{i}^{|\mathcal{B}|} \sum_{j}^{L_s} \log \frac{\exp\left(\text{sim}(\mathbf{t}_i^{g(j)}, \mathbf{a}_i^j)/\tau\right)}{\sum_{k \neq i}^{|\mathcal{B}|} \exp\left(\text{sim}(\mathbf{t}_i^{g(j)}, \mathbf{a}_k^j)/\tau\right)} \tag{5}$$

where $(\mathbf{t}_i^{g(j)}, \mathbf{a}_i^j)$ are the sequence-level representations for the i-th unlabelled text-audio pair within $\mathcal{B}$ generated by the $g(j)$-th layer of NLU BERT and the j-th layer of E2E-SLU BERT,[1] τ is the temperature hyperparameter, and sim$(\cdot,\cdot)$ is the cosine similarity. Given a true (positive) text-audio pair, the negative pairs are implicitly formed by iterating audio representations in the training batch. The benefit of this contrastive loss is to pull together the audio representations closer to their text representations, and push away from those of other texts.

2.3. DA within MATL

Due to the lack of high-quality labelled SLU data, as well as the nature of knowledge distillation and contrastive learning that they become more competitive with more data, data augmentation (DA) methods are essential for the success in cross-modal transfer [27]. We focus on distortion-based DA methods, which only disturb model input with preset noise, and explore various DA strategies within different parts of MATL (see Figure 1): 1) We apply **Cutoff** [28] by randomly erasing 15% of the tokens or feature dimensions from input sequence of both NLU and E2E-SLU BERTs to enable the model to focus on the whole sequence instead of a certain part of it; 2) For **Add Noise** [29], Gaussian

[1] We add a trainable [CLS] embedding in front of the E2E-SLU BERT input which is the output of speech encoder.

Models	Modality Information	Layers	SLURP		JDTEL	
			Acc	Macro-F1	Acc	Macro-F1
Oracle: NLU model	Transcription	12	86.28	85.92	74.60	68.03
B1: Wav2Vec2	Audio	24	74.71	69.55	59.94	50.21
B2: E2E-SLU model	Audio	24+4	75.15	71.16	61.13	50.25
STD [7]	Transcription→ Audio	24+4	75.82	71.86	63.01	53.98
SPLAT-seq [8]	Transcription→ Audio	24	76.43	72.72	59.67	46.36
XSTNet [3]	Transcription & Audio	24+12	75.29	71.15	60.92	50.75
MATL	Transcription → Audio	24+4	**78.72**	**73.97**	**64.13**	**54.69**
- DA	Transcription → Audio	24+4	77.47	73.12	63.94	54.30
- DA - CL	Transcription → Audio	24+4	76.03	71.67	63.71	54.24
- DA - CL - CA	Transcription → Audio	24+4	75.00	71.57	63.08	53.90

Table 2: Average test results in accuracy and Macro-F1 score of MATL and other models over three runs. The best results are in bold.

noise is added to the input embeddings of NLU BERT, E2E-SLU speech encoder, and E2E-SLU BERT to smooth the input embedding space; 3) **Dropout** [30] randomly sets elements in the embedding layer to zero by a specific probability and is advocated as a better data augmentation method for text [19]. 4) SpecAugment (**SpecAug**) [31] operates on the log-mel spectrogram of the input audio to focus less on the features of a particular frequency or time, and more on the entire spectrum. We apply a modified version to the raw waveform by masking its time steps and channels [32].

Finally, label information is distilled through the soft cross entropy loss from the logits of teacher to student models:

$$\mathcal{L}_{\text{pred}} = \text{CE}(logits_t, logits_s) \tag{6}$$

and the full training objective for MATL is the weighted sum of the distillation losses and contrastive loss:

$$\mathcal{L}_{\text{MATL}} = \alpha_1\mathcal{L}_{\text{hid}} + \alpha_2\mathcal{L}_{\text{att}} + \alpha_3\mathcal{L}_{\text{pred}} + \alpha_4\mathcal{L}_{\text{cl}} \tag{7}$$

where α_1, α_2, α_3, and α_4 are hyperparameter weights.

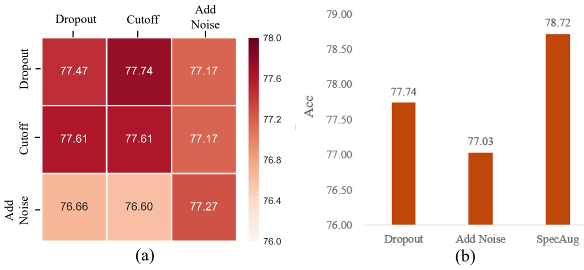
3. Experiments

3.1. Dataset

Experiments are conducted on two SLU datasets in different languages, where text-audio pairs along with their intent labels are available.[2] Their statistics are shown in Table 1: 1) **SLURP** [21] is a new challenging benchmark dataset for SLU. It contains 18 scenarios, such as music and cooking, with 46 defined actions, e.g. ticket and coffee. We concatenate them and use **scenario_action** as the intent. Since a text contains multiple audios, we choose the first audio for each text in the original file. 2) **JDTEL** is a Chinese SLU dataset for intent classification, which is collected from our online voice-enabled customer service bot. The audios are from real users after anonymisation, and the texts are manually transcribed.

3.2. Training details

To facilitate flexible experiments in multiple languages, we initialize our NLU text encoder with *bert-base-multilingual-cased* (mBERT) that has 12 Transformer layers and 768 hidden size. We use the Wav2Vec2 [32] multilingual variant *wav2vec2-xls-r-300m* speech encoder, which contains 24 layers and 1024 hidden units and is further updated during knowledge distillation.



(a) (b)

Figure 2: Performance visualisation for embedding layer DA of both modalities (a) and DA for audio input (b).

The E2E-SLU BERT has the same architecture as mBERT, except that the number of layers is 4. To connect Wav2Vec2 and E2E-SLU BERT, we first apply a CNN layer to compress the speech length, and then an MLP layer converts the representation dimensionality from 1024 to 768. The values α_1, α_2, α_3 and α_4 are set to 0.1, 0.1, 0.8 and 1.0. The temperature τ is 1.0. These hyperparameters are tuned on the SLURP development set. In DA, the mean and variance of the Gaussian noise is 0 and 0.01 in teacher and student BERT. In Wav2Vec2, the noise is $\sigma * n(t)$, where $\sigma \in [0.001, 0.015]$ is the amplitude factor and follows uniform distribution, and $n(t)$ is the Gaussian noise of mean 0 and variance 1 [29].

3.3. Main results and ablation study

We compare our proposed MATL framework with several baselines: 1) **STD** [7] can be seen as a simplified version of MATL, which is trained with $\mathcal{L}_{\text{hid}}$ and $\mathcal{L}_{\text{att}}$ by padding text and speech to the same sequence length along with $\mathcal{L}_{\text{pred}}$ using the true label information; 2) **SPLAT-seq** [8] aligns the sequence-level representations of both modalities by minimising L1 loss; 3) **XSTNet** [3] represents another paradigm of knowledge transfer by performing multi-task training for both modalities on a shared encoder, which has shown strong performance in E2E speech translation. Additionally, we also compare uni-modal baselines for the upper (text) and lower (speech) bound reference. The **Oracle** is our NLU teacher model finetuned and tested on transcribed text data, whereas Wav2Vec2 (**B1**) is our speech encoder and **B2** is our full E2E-SLU model. Both baselines only exploiting speech information without knowledge transfer.[3]

Table 2 illustrates the experimental results with *accuracy* and *macro-F1* score are reported as as evaluation metrics. It's observed that, the performance of B1 and B2 is far behind Oracle on both datasets, clearly manifesting the difficulty in lan-

[2] The well known Fluent Speech Commands (FSC) dataset [2] is not used due to its short text data with easiness leading to its test accuracy in most work over 99%.

[3] For fair comparison, all baselines use the same pretrained mBERT text encoder and/or Wav2Vec2 speech encoder as initialisation.

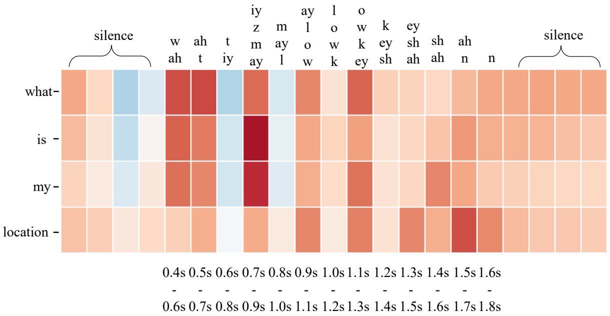


Figure 3: Visualisation of cross attention map. Row indicates speech time intervals and column for text tokens.

guage understanding task only with audio signals and emphasising the necessity and potential in cross-modal transfer for E2E-SLU. For knowledge distillation methods, all the three baselines can improve the performance over B1 and B2 (except for SPLAT-seq and XSTNet in Chinese), corroborating the usefulness and complementarity of text information in semantic transfer for E2E-SLU models. In contrast, our MATL produces the best results, consistently outperforming all the baselines by a significant margin on both datasets (**+2.29** on SLURP and **+1.12** on JDTEL in accuracy, $p < 0.05$ with student's t-test over three runs). We also conduct further ablation study by incrementally removing DA (- DA), CL (- DA - CL) and CA (- DA - CL - CA) and the results become worse, showing that they are all essential for MATL in cross-modal transfer. Besides, note that even without DA, MATL still beats all baselines, indicating our multi-grained alignment with CA and CL is more effective than single level alignment (e.g. STD and SPLAT-seq).

3.4. Discussion

The effect of DA. Here we study how various DA strategies mentioned in §2.3 influence the model performance. 1) We inspect the DA methods applied to the BERT embedding layers of both NLU and E2E-SLU models by enumerating the combinations of the three methods and displays their corresponding performance in Figure 2 (a), where the row indicates strategies for E2E-SLU BERT and the column for NLU BERT. In general, Dropout and (Token/Feature) Cutoff are the two most effective strategies, better than Add Noise for both modalities. The best performance is achieved by equipping NLU BERT with Dropout and E2E-SLU with Cutoff. This is possibly because the inherent representation disparities between the two modalities, where Cutoff employed on the longer speech representations helps to discard redundant features whereas Dropout is a more refined regularisation method for text representations to retain important information. 2) Atop the best strategy for BERT embeddings, we then consider DA approaches to the waveform input of Wav2Vec2. Figure 2 (b) presents that SpecAug, tailored for speech, indeed performs the best, generating the gain around 1 absolute point over Dropout.

Visualisation for CA. Figure 3 visualises CA map between the middle (second) layer in E2E-SLU BERT and its corresponding NLU BERT layer for more clear semantic relations [33, 34]. The phonemes representing speech appearing in the approximate time interval are labelled in row according to [16], together with text tokens in column. Evidently, the phonemes covering stressed syllables are well aligned with their matching tokens (w ah → what), while silence generally does not map to any token, indicating that MATL has enabled fine-grained semantic transfer through CA.

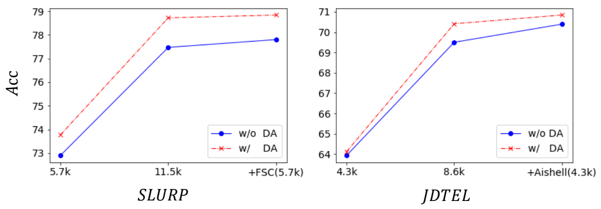


Figure 4: Results on increasing unlabelled text-audio pairs.

Train Language	Test Langeage	Text	Audio
English	English	86.28	78.72
Chinese	Chinese	74.60	64.13
English+Chinese	English	86.08	78.51
English+Chinese	Chinese	74.88	64.43

Table 3: Crosslingual and mono-lingual results with MATL.

Knowledge distillation with more data. To unlock the full potential of cross-modal transfer learning, we experiment on adding more text-audio pairs without intent labels. Both in-task and out-of-task data are explored, where in-task data comes from the same SLU task (e.g. we double the query size of JDTEL, and reduce the SLURP data by half as baseline), and out-of-task data comes from the public datasets (e.g. $5.7k$ FSC data [2] for English and $4.3k$ Aishell data [35] for Chinese). We conduct the experiments for each training size twice, with and without our DA methods, and present the results in Figure 4. The results demonstrate that more unlabelled pairs are beneficial and the gain brought by in-task data seems to be larger. Notably, by adding just $4.3k$ in-task data for JDTEL improves the results drastically, much closer to the text Oracle, which hints the huge potential of MATL in real applications. Besides, despite the usefulness with more data, DA methods can still enhance the model performance on top of that, which strengthens again the efficacy of our DA approaches.

Towards crosslingual E2E-SLU with MATL. The ultimate goal for a single crosslingual E2E-SLU model [11, 17] has now become more feasible with MATL thanks to the pretrained multilingual text and speech encoder. Here we perform crosslingual training by merging SLURP and JDTEL, and apply MATL to learn a single crosslingual E2E-SLU model. Table 3 shows a single crosslingual MATL can further improve on JDTEL (Chinese) with a slight drop in SLURP (English), indicating there is a synergy between the text and audio results and MATL's capability on crosslingual knowledge transfer.

4. Conclusions

In this paper, we presented MATL, a multi-grained alignment framework, to transfer knowledge from text to speech for E2E-SLU, containing three components. CA was devised for fine-grained token-frame alignment, while CL was leveraged for coarse-grained sequence-level alignment respectively. Besides, DA was introduced to bridge the gap between the two modalities. Experimental results demonstrated the superior performance of MATL and the importance of each component. Further analyses revealed that adding more unlabelled data, either in-task or out-of-task, improved the model performance. And a single E2E-SLU model could be obtained through MATL through crosslingual training to achieve competitive results against its mono-lingual counterpart. In the future, we will explore linguistic information from different levels and granularities in alignment, e.g. morphology and syntax, for more effective cross-modal transfer.

5. References

[1] 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.

[2] 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.

[3] R. Ye, M. Wang, and L. Li, "End-to-end speech translation via cross-modal progressive training," *arXiv preprint arXiv:2104.10380*, 2021.

[4] 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.

[5] X. Zhang and L. He, "End-to-end cross-lingual spoken language understanding model with multilingual pretraining," *Proc. Interspeech 2021*, pp. 4728–4732, 2021.

[6] Y. Huang, H.-K. Kuo, S. Thomas, Z. Kons, K. Audhkhasi, B. Kingsbury, R. Hoory, and M. Picheny, "Leveraging unpaired text data for training end-to-end speech-to-intent systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 7984–7988.

[7] 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.

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

[9] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 7463–7467.

[10] 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.

[11] B. Sharma, M. C. Madhavi, X. Zhou, and H. Li, "Exploring teacher-student learning approach for multi-lingual speech-to-intent classification," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021.* IEEE, 2021, pp. 419–426. [Online]. Available: https://doi.org/10.1109/ASRU51503.2021.9688041

[12] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[14] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4163–4174.

[15] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Z. Wang, Y. Le, Y. Zhu, Y. Zhao, M. Feng, M. Chen, and X. He, "Building robust spoken language understanding by cross attention between phoneme sequence and asr hypothesis," 2022. [Online]. Available: https://arxiv.org/abs/2203.12067

[17] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[19] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv*, 2021.

[20] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer," *arXiv preprint arXiv:2105.11741*, 2021.

[21] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[22] 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.

[23] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[26] L. Wang and A. van den Oord, "Multi-format contrastive learning of audio representations," *CoRR*, vol. abs/2103.06508, 2021. [Online]. Available: https://arxiv.org/abs/2103.06508

[27] K. M. Yoo, Y. Shin, and S.-g. Lee, "Data augmentation for spoken language understanding via joint variational generation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7402–7409.

[28] D. Shen, M. Zheng, Y. Shen, Y. Qu, and W. Chen, "A simple but tough-to-beat data augmentation approach for natural language understanding and generation," *arXiv preprint arXiv:2009.13818*, 2020.

[29] S. Wei, S. Zou, F. Liao *et al.*, "A comparison on data augmentation methods based on deep learning for audio classification," in *Journal of Physics: Conference Series*, vol. 1453, no. 1. IOP Publishing, 2020, p. 012085.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

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

[33] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[34] W. Liu, X. Fu, Y. Zhang, and W. Xiao, "Lexicon enhanced chinese sequence labeling using bert adapter," *arXiv preprint arXiv:2105.07148*, 2021.

[35] X. N. B. W. H. Z. Hui Bu, Jiayu Du, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017.