SEMI-SUPERVISED SPEECH-LANGUAGE JOINT PRE-TRAINING FOR SPOKEN LANGUAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken language understanding (SLU) requires a model to analyze input acoustic signals to understand its linguistic content and make predictions. To boost the models' performance, various pre-training methods have been proposed to utilize large-scale unlabeled text and speech data. However, the inherent disparities between the two modalities necessitate a mutual analysis. In this paper, we propose a novel semi-supervised learning method, AlignNet, to jointly pre-train the speech and language modules. Besides a self-supervised masked language modeling of the two individual modules, AlignNet aligns representations from paired speech and transcripts in a shared latent semantic space. Thus, during fine-tuning, the speech module alone can produce representations carrying both acoustic information and contextual semantic knowledge. Experimental results verify the effectiveness of our approach on various SLU tasks. For example, AlignNet improves the previous state-of-the-art accuracy on the Spoken SQuAD dataset by 6.2%.

1 INTRODUCTION

Spoken language understanding (SLU) tackles the problem of comprehending audio signals and making predictions related to the content. SLU has been widely employed in various areas such as intent understanding (Tur & De Mori, 2011; Bhargava et al., 2013; Ravuri & Stolcke, 2015; Lugosch et al., 2019), question answering (Lee et al., 2018; Chuang et al., 2020) and sentiment analysis (Zadeh et al., 2018). Early approaches leverage a two-step pipeline: use automatic speech recognition (ASR) to transcribe input audio into text, and then employ language understanding models to produce results. However, this cascaded architecture has several drawbacks. First, the transcription produced by the ASR module often contains errors, which adversely affects the prediction accuracy. Second, even if the transcription is perfect, the rich prosodic information (e.g., tempo, pitch, intonation) is inevitably lost after ASR. In comparison, humans often leverage these information to better understand and disambiguate the content. Therefore, there has been a rising trend of end-to-end approaches to retain information from audio signals to carry out the understanding task (Serdyuk et al., 2018; Chen et al., 2018; Haghani et al., 2018).

While end-to-end SLU methods are effective, they often suffer from a shortage of labeled training data, especially when the target task is in a novel domain. One solution is to leverage self-supervised training as is done in pre-trained language models. BERT (Devlin et al., 2019), GPT (Radford et al., 2018) and RoBERTa (Liu et al., 2019) are first pre-trained on large-scale unannotated text in a self-supervised fashion to learn a high-quality representation before being fine-tuned on downstream tasks with a modest amount of labeled data. Borrowing this idea, several pre-training methods have been proposed for acoustic input, e.g., wav2vec (Schneider et al., 2019; Baevski et al., 2020), contrastive predictive coding (Oord et al., 2018; Rivière et al., 2020), and autoregressive predictive coding (Chung et al., 2019; 2020; Ling et al., 2020), to capture contextual representation from unlabeled speech data. Nevertheless, these methods only focus on acoustic data during pre-training. As a result, the produced embeddings may not be optimal for the language understanding task.

To solve these problems, we propose a novel speech-languaget joint pre-training framework, AlignNet. AlignNet contains a speech module and a language module for multi-modal understanding. The speech module is a transformer architecture trained from scratch and the language module is initialized from BERT. Both modules leverage large-scale unannotated data for pre-training via masked language modeling. In the speech module, each frame is seen as a token and is replaced with zero

vector with a certain probability. For each masked frame, we minimize the L1-distance between the generated representation and the original frame.

Then, to make the speech module aware of the contextual information extracted from the language module, we design an alignment loss to align the representations from both modules in a shared latent semantic space. In detail, we propose two alignment methods, i.e. sequence-level and token-level, to leverage a small amount of paired speech and transcripts and minimize the disparity between the audio representation from the speech module and the text representation from the language module. In this way, the speech representation will carry not only the acoustic information but also the contextual knowledge from the text. After this alignment, when text input is absent during fine-tuning, the speech module alone can produce representations that bridge the speech input and the language understanding output.

We conduct extensive evaluations on several downstream SLU tasks, including Fluent Speech Commands for intent detection, Switchboard for dialog act classification, CMU-MOSEI for spoken sentiment analysis, and Spoken SQuAD for spoken question answering. AlignNet achieves superior results in all datasets. For example, AlignNet improves the previous state-of-the-art accuracy on the Spoken SQuAD dataset by 6.2%. Furthermore, we show that AlignNet can perform well even given a tiny portion of the labeled training data in downstream tasks.

2 Related Work

Spoken language understanding In recent years, due to its flexibility and effectiveness, end-to-end spoken language understanding (SLU) has been proposed and applied to various tasks (Qian et al., 2017; Serdyuk et al., 2018; Lugosch et al., 2019). For instance, Qian et al. (2017) uses an auto-encoder to initialize the SLU model. Lugosch et al. (2019) pre-trains the model to recognize words and phonemes, and then fine-tunes it on downstream tasks. Chen et al. (2018) pre-trains the model to categorize graphemes, and the logits are fed into the classifier. In most of these approaches, the model pre-training requires labeled speech data, e.g., word or phonemes corresponding to audio signals. As a result, the massive unlabeled speech data cannot be utilized by these models.

Self-supervised pre-training Pre-trained models have achieved great success in both language and speech domains. In language, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), UniLM (Dong et al., 2019) and BART (Lewis et al., 2020) have been successfully applied in natural language inference (Zhang et al., 2020b), question answering (Zhu et al., 2018) and summarization (Zhu et al., 2019). These pre-trained models leverage self-supervised learning tasks such as masked language modeling (MLM), next sentence prediction, and de-noising autoencoder.

In speech, wav2vec (Schneider et al., 2019) leverages contrastive learning to produce contextual embeddings for audio input; vq-wav2vec (Baevski et al., 2020) further learns to discretize audio input to enables more efficient MLM training with Transformer (Vaswani et al., 2017). Pre-trained speech models have been applied in automatic speech recognition (ASR) (Ling et al., 2020; Chung & Glass, 2020) and phoneme recognition (Song et al., 2020).

Nevertheless, an SLU model must incorporate both acoustic and language understanding capabilities to project speech signals to semantic outputs. Thus, a pre-trained model for SLU needs to address tasks beyond a single modality.

Speech and text joint pre-training Recently, SLU applications have prompted joint pre-training on both speech and text data. SpeechBERT (Chuang et al., 2020) applies MLM to pairs of audio input and transcripts. However, compared to our work, there are several crucial differences. First, SpeechBERT contains a separate phonetic-semantic embedding module to produce paired audio data, which requires pre-training and may generate errors. Second, both the pre-training and fine-tuning phases of SpeechBERT adopt input of both speech and text signals. However, many SLU datasets only contain speech input, which does not align with SpeechBERT. In comparison, our model aligns the speech module with the language understanding module during pre-training, and only needs speech input for downstream tasks.

Denisov & Vu (2020) proposes to align speech embeddings with language embeddings, in a method similar to ours. However, there are several key differences. Firstly, Denisov & Vu (2020) employs

the encoder of a pre-trained ASR model, which already requires plentiful of annotated speech to obtain. Our model conducts self-supervised learning for the speech module based on unlabeled audio inputs. Secondly, in addition to sentence embedding alignment, we propose a token-level alignment, which is suitable for token-level downstream tasks. Thirdly, our model uses a much smaller speech corpus for alignment (10 hours) than Denisov & Vu (2020) (1,453 hours), yet still outperforms it by 4.6% in the downstream dialog act classification task.

3 METHOD

In this section we introduce AlignNet, a model for learning joint contextual representation of speech and language. The model consists of a speech module and a language module that share a similar architecture learning algorithm. Both modules are first pre-trained with unannotated speech or text data. Then, we leverage an alignment task with a small amount of paired speech and transcript data to align the representations from both modules in a shared latent semantic space so that the information learned by the language module is transferred to the speech module. After pre-training, the language module is discarded and only the speech module is used in downstream tasks. Below we formally describe the two modules (§3.1 and § 3.2), the alignment loss (§3.3), and the training procedure (§3.4). Figure 1 provides an overview of AlignNet.

3.1 SPEECH MODULE

The goal of this module is to learn a representation that contains useful acoustic information about speech utterances such as their phonetic content and speaker characteristics.

Formally, the input to the speech module consists of n audio features based on 80-dimensional log Mel spectrograms, $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. The input is fed into the Transformer architecture to produce output embeddings $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$.

To boost its capacity for contextual understanding, we use the idea of masked language modeling (MLM) (Devlin et al., 2019; Liu et al., 2020; Wang et al., 2020). Specifically, each audio feature $\boldsymbol{x}_i$ is replaced with a zero vector with a probability of 15%. The corresponding output $\boldsymbol{s}_i$ is trained to be close to the original feature $\boldsymbol{x}_i$.

Furthermore, according to SpecAugment (Park et al., 2019), the input features $\{\boldsymbol{x}_i\}$ can be seen as comprising two dimensions: time, i.e., the subscript i, and channel, i.e., the elements in each $\boldsymbol{x}_i$. While conventional MLM masks certain timepoints, the input signals can also be masked in the channel dimension. In other words, for each $1 \leq j \leq d$, where $\boldsymbol{x}_i \in R^d$, $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n,j}$ are masked and set to zero vectors with a probability of 15%. This channel masking is combined with temporal masking to reinforce the model's capability to utilize contextual information from both time and channel, and reduce the impact of co-adaptation between features. So the loss function for the speech module is:

$$\mathcal{L}_{sp} = \sum_{\boldsymbol{x}_i=1,2,...,n} \|\boldsymbol{x}_i - \boldsymbol{s}_i\|_1 \tag{1}$$

3.2 LANGUAGE MODULE

The language module aims to offer contextual understanding for text input. Given token embeddings $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_m\}$, the module produces contextual representations $\{\boldsymbol{t}_1, ..., \boldsymbol{t}_m\}$.

We employ the BERT-base model (Devlin et al., 2019) to initialize the language module. In order to adapt the model to speech domain, we adapt the language model using the MLM task on transcripts from LibriSpeech. The corresponding cross-entropy loss function is denoted by $\mathcal{L}_{text}$.

3.3 ALIGNING SPEECH AND LANGUAGE REPRESENTATIONS

The input to most SLU tasks consists of only audio signals, but the model is required to conduct semantic understanding, which can be best handled when textual information is present. Therefore, we propose to align the pre-trained speech and language representations in a shared semantic latent space.

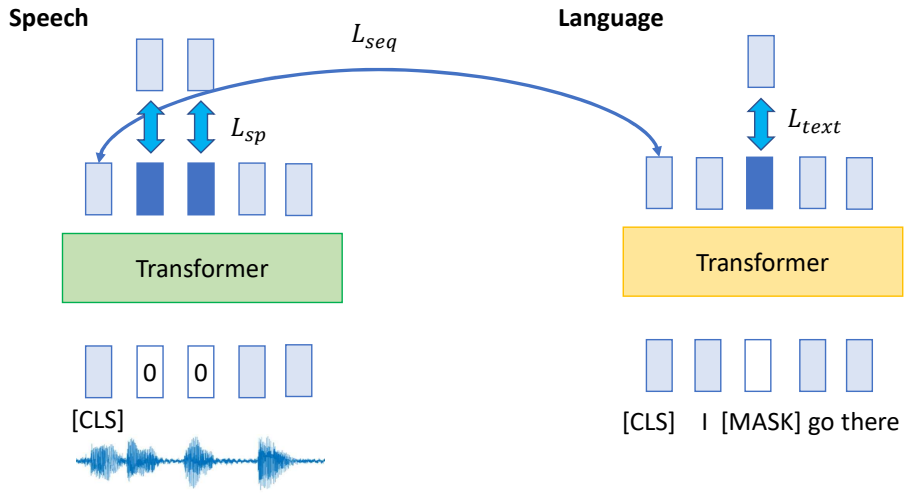

Figure 1: Architecture of AlignNet. The speech module and language module conduct self-supervised learning via masked language modeling. The two modules are then aligned using paired audio and text data with sequence-level alignment. After pre-training, only the speech module is used in downstream tasks.

Suppose an audio input $\{\boldsymbol{x}_i\}_{i=1}^n$ comes with its transcript $\{\boldsymbol{y}_j\}_{j=1}^m$. We first prepend both inputs with a special token [CLS], which has learnable embeddings represented by $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ respectively. The speech and language modules separately produce the output representations $\{\boldsymbol{s}_0, ..., \boldsymbol{s}_n\}$ and $\{\boldsymbol{t}_0, ..., \boldsymbol{t}_m\}$. We then propose two methods to align the embeddings from the modules: sequence-level and token-level alignment.

Sequence-level alignment. As the special [CLS] token can be seen as a representation of the whole input sequence, we minimize the distance between the representation of [CLS] in speech input and that of [CLS] in text input:

$$\mathcal{L}_{seq} = \|\boldsymbol{s}_0 - \boldsymbol{t}_0\|_1 \tag{2}$$

After pre-training, the output embedding of [CLS] in the speech module will be close to its corresponding text embedding in the language module, even when the transcript is absent in downstream tasks. Thus, this alignment can help with sequence-level SLU tasks to predict the property of the whole audio input, e.g., intent classification.

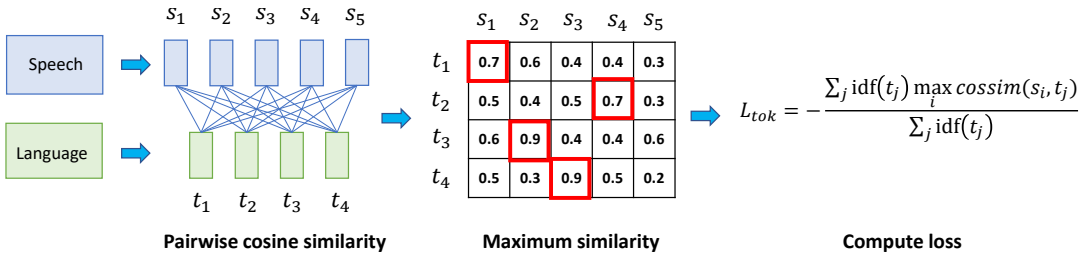

Figure 2: Token-level alignment between speech and language modules.

Token-level alignment. To achieve a finer level of alignment, each audio feature should be compared with its corresponding text token. Although forced alignment (Gorman et al., 2011) can establish this correspondence between audio signals and individual words, it is very laborious to acquire such labeling. Therefore, we propose to automatically align audio features to textual tokens.

Inspired by BERTScore (Zhang et al., 2020a), for each output text embedding $\boldsymbol{t}_j$, we compute its cosine similarity with each output audio embedding $\boldsymbol{s}_i$, and select the audio feature with the highest similarity.

The alignment then maximizes the sum of these maximum similarities over all tokens, weighted by each token's inverse document frequency (idf) to reduce the impact of common words:

$$\mathcal{L}_{tok} = -\frac{\sum_{j=1}^{m} \text{idf}(\boldsymbol{t}_j) \max_i \text{cossim}(\boldsymbol{s}_i, \boldsymbol{t}_j)}{\sum_{j=1}^{m} \text{idf}(\boldsymbol{t}_j)} \tag{3}$$

The token-level alignment is illustrated in Figure 2. It can help with token-level SLU tasks to predict the category of various segments of audio input, e.g., extractive spoken question answering.

3.4 Training procedure

We use the `train-clean-360` subset of the LibriSpeech corpus (Panayotov et al., 2015) to pre-train the speech module, i.e., minimizing $\mathcal{L}_{sp}$. This subset contains 360 hours of read speech produced by 921 speakers. We use 80-dimensional log Mel spectrograms and normalize them to zero mean and unit variance per speaker as input acoustic features, i.e., $\boldsymbol{x}_t \in \mathbb{R}^{80}$.

We then randomly sample 10 hours of transcripts to pre-train the language module, i.e., minimizing $\mathcal{L}_{text}$. Then, the corresponding 10 hours of audio is paired with these transcripts for the alignment task, i.e., minimizing $\mathcal{L}_{seq}$ or $\mathcal{L}_{tok}$.

During fine-tuning, only the speech module is used in downstream SLU tasks.

4 Experimental Setup

4.1 Baselines

We include a number of strong baselines from recent literature for each downstream task (Lugosch et al., 2019; Duran & Battle, 2018; Ghosal et al., 2018; Chuang et al., 2020). We also compare with another speech-language joint pre-training framework (Denisov & Vu, 2020).

To verify the effectiveness of our proposed framework, we experiment with the following variants of our model, including whether to pre-train the model, whether to use the language module and which alignment task to apply.

1. **AlignNet-Scratch**: AlignNet without pre-training, i.e., the speech module is trained from scratch on downstream tasks.
2. **AlignNet-Speech**: AlignNet pre-trained without the language module.
3. **AlignNet-Seq**: AlignNet with sequence-level alignment, but language module is not updated with MLM.
4. **AlignNet-Seq-MLM**: AlignNet with sequence-level alignment, and language module is updated with MLM.
5. **AlignNet-Tok**: AlignNet with token-level alignment, but language module is not updated with MLM.
6. **AlignNet-Tok-MLM**: AlignNet with token-level alignment, and language module is updated with MLM.

The speech module of AlignNet is a 3-layer Transformer encoder where each layer has a hidden size of 768 and 12 self-attention heads. The language module has the same configuration as $\text{BERT}_{\text{BASE}}$ and its parameters are initialized from the pre-trained $\text{BERT}_{\text{BASE}}$ parameters released by Devlin et al. (2019).

4.2 Downstream SLU Tasks

We evaluate our model on four different types of SLU applications: intent detection, dialog act classification, spoken sentiment analysis, and spoken question answering. The first three belong to multi-class classification tasks, and the last one is a span prediction problem, which will be described in more details below. Table 1 summarizes the used dataset for each application. For all datasets, we use 80-dimensional log Mel spectrograms as input acoustic features as in the pre-training stage.

Table 1: Summary of SLU datasets. For the rows of Train, Validation, and Test, the numbers indicate the number of utterances in the split.

Task	Intent detection	Dialog act classification	Spoken sentiment analysis	Spoken question answering
Dataset	FSC	SwBD	CMU-MOSEI	Spoken SQuAD
Num. of classes	31	42	7	-
Train	23,132	97,756	16,216	35,111
Validation	3,118	8,591	1,835	2,000
Test	3,793	2,507	4,625	5,351

Intent detection We use the Fluent Speech Commands corpus (FSC) (Lugosch et al., 2019) for intent detection, where the goal is to correctly predict the intent of an input utterance. In this dataset, each utterance is annotated with three slots: action, object, and location, where each slot can take one of multiple values. The combination of slot values is defined as the intent of the utterance, and there are 31 unique intents in total. In this work we follow the original paper to formulate intent detection as a simple 31-class classification task.

Dialog act classification We use the NTX-format Switchboard corpus (SwDA) (Calhoun et al., 2010), a dialog corpus of 2-speaker conversations. The goal is to correctly classify an input utterance into one of the 42 dialog acts.

Spoken sentiment analysis We use the CMU-MOSEI dataset (Zadeh et al., 2018), where each utterance is annotated for a sentiment score on a $[-3, 3]$ Likert scale: [-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive]. We treat the task as a 7-class classification problem. And we only use audio signals in the input data.

For the above three applications, during fine-tuning, an MLP network with one hidden layer of 512 units is appended on top of the speech module. It converts the output representation of [CLS] for class prediction. Both the pre-trained speech module and the randomly initialized MLP are fine-tuned on the training set for 10 epochs with a batch size of 64 and a fixed learning rate of 3e-4. We compute classification accuracy after each training epoch and pick the best-performing one to report results on the test set.

Spoken question answering We use the Spoken SQuAD dataset (Li et al., 2018), which is augmented[1] from SQuAD (Rajpurkar et al., 2016) for spoken question answering. The model is given an article in the form of speech and a question in the form of text. The goal is to predict a time span in the spoken article that answers the question. In other words, the model outputs an audio segment extracted from spoken article as the answer. The model is evaluated by Audio Overlapping Score (AOS) (Li et al., 2018): the more overlap between the predicted span and the ground-truth answer span, the higher the score will be.

During fine-tuning, given a spoken article and a question in the text form, the pre-trained speech module extracts audio representations of the article and pass them to a randomly initialized 3-layer Transformer encoder along with the tokenized textual question as input. The Transformer then uses the self-attention mechanism to implicitly align elements of the input audio and textual features. For each time step of the audio input, the Transformer is trained to predict whether this is the start of the span with a simple logistic regression classifier. A separate classifier is used for predicting the end of the span.

5 Results and Analyses

5.1 Main results

Table 2 shows the performance of models on all four downstream tasks. Each number from our model is an average over 3 runs. Based on the results, we make the following observations.

[1]Li et al. (2018) used Google text-to-speech to generate the spoken version of the articles in SQuAD.

Table 2: Results on all downstream datasets. All numbers of our models are average of 3 runs. The metric is classification accuracy for FSC, SwBD and CMU-MOSEI. The metric for Spoken SQuAD is Audio Overlapping Score (AOS).

Model	FSC	SwBD	CMU-MOSEI	Spoken SQuAD
AlignNet-Scratch	97.6	65.8	68.8	30.4
AlignNet-Speech	99.5	67.5	69.0	57.7
AlignNet-Seq	99.5	74.6	72.5	62.7
AlignNet-Seq-MLM	99.5	**76.3**	74.7	**65.9**
AlignNet-Tok	99.2	71.2	70.4	63.8
AlignNet-Tok-MLM	99.2	72.7	71.2	58.0
AlignNet-Seq-MLM 1-hour	99.5	75.8	65.3	65.3
Lugosch et al. (2019)	98.8	-	-	-
Duran & Battle (2018)	-	75.5	-	-
Ghosal et al. (2018)	-	-	**75.9**	-
Chuang et al. (2020)	-	-	-	59.7
Denisov & Vu (2020)	**100.0**	71.7	-	-

Firstly, compared with AlignNet-Scratch, all pre-trained models achieve superior results, especially more than 30% gain on Spoken SQuAD, proving the effectiveness of pre-training.

Secondly, the inclusion of language module and the alignment task during pre-training is very beneficial. For instance, on CMU-MOSEI dataset, AlignNet-Seq-MLM outperforms AlignNet-Speech by 5.7%, and it outperforms several baseline systems from recent literature. We argue that as SLU tasks require the model to interpret acoustic signals and their semantic meaning, the language module and alignment task will guide the speech module towards a mutual understanding of both modalities.

Thirdly, comparing AlignNet-Seq against AlignNet-Tok, we find that the sequence-level alignment outperforms token-level alignment on the first three sequence classification tasks, while token-level alignment achieves higher accuracy on Spoken SQuAD, a token classification task. It shows that a closer resemblance between the pre-training goal and the property of downstream task will lead to better performance.

Fourthly, updating the language module using MLM during pre-training is helpful. Although the language module has been initialized with BERT, adaptation to the speech domain can help with semantic understanding in the downstream task.

Finally, we experimented with a version of AlignNet which uses only 1 hour of transcribed speech sampled from Librispeech, AlignNet-Seq-MLM 1-hour. It achieves comparable results with the best variant AlignNet-Seq-MLM: same accuracy on FSC, 0.5% less on SwBD and 0.6% less on Spoken SQuAD. This shows that with a small amount of labeled speech data, our pre-training framework can achieve good results on downstream tasks.

5.2 Robustness to size of training data

As human labeling process is time-consuming and labor-intensive, the amount of labeled training data for downstream tasks is often small and insufficient. In this section, we show that with effective pre-training, the model will be less dependent on the amount of downstream labeled data.

We randomly sample 50%, 10%, 5%, and 1% of the training data in the downstream tasks, and evaluate the performance of different variants of AlignNet when fine-tuned on the sampled data.

Table 3 shows the performance on all four downstream tasks with varying training data sizes. We observe that among the variants, AlignNet-Seq-MLM is least sensitive to training data sizes. For instance, in FSC, with only 10% of the training data, its accuracy only drops 0.4 points. In comparison, both AlignNet-Scratch and AlignNet-Speech drops about 10 points. And the gaps are in general larger when the size of training data further shrinks. Therefore, our proposed joint pre-training of speech and language modules can help the model quickly adapt to downstream tasks given a modest amount of training data.

Table 3: Performance on downstream tasks with varying training data sizes.

FSC					
Model	100%	50%	10%	5%	1%
AlignNet-Scratch	97.6	95.1	86.8	25.4	2.7
AlignNet-Speech	99.5	99.0	90.4	52.7	34.0
AlignNet-Seq-MLM	99.5	99.5	99.1	63.8	45.7
SwBD					
AlignNet-Scratch	65.8	65.3	59.8	45.9	31.6
AlignNet-Speech	67.5	66.5	63.6	46.7	33.7
AlignNet-Seq-MLM	76.3	70.7	66.9	53.2	38.6
MOSEI					
AlignNet-Scratch	68.8	66.3	50.7	32.6	15.4
AlignNet-Speech	69.0	67.5	56.6	37.3	20.5
AlignNet-Seq-MLM	74.7	72.9	64.1	44.3	23.0
Spoken SQuAD					
AlignNet-Scratch	30.4	27.9	22.3	15.9	9.8
AlignNet-Speech	57.7	55.0	51.2	46.5	32.4
AlignNet-Seq-MLM	65.9	63.8	60.1	50.3	37.9

6 CONCLUSIONS

Spoken language understanding (SLU) tasks require an understanding of the input audio signal and its underlying semantics. In this paper, we presented a novel semi-supervised joint pre-training framework, AlignNet, to carry out both speech and language understanding tasks during pre-training. Besides a self-supervised training on the speech and language modules, we propose two methods to align the semantic representations from both modules using a modest amount of labeled speech data. The speech module can quickly adapt to downstream tasks and achieve superior results on various SLU datasets including intent detection, dialog act classification, spoken sentiment analysis and spoken question answering. This joint pre-training also makes the model less sensitive to the amount of labeled training data in downstream domains.

For future work, we plan to integrate automatic speech recognition (ASR) and natural language generation (NLG) into our framework to achieve good results on spoken language generation tasks.

REFERENCES

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *ICLR*, 2020.

Aditya Bhargava, Asli Celikyilmaz, Dilek Hakkani-Tür, and Ruhi Sarikaya. Easy contextual intent prediction and slot detection. In *ICASSP*, 2013.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4): 387–419, 2010.

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. Spoken language understanding without speech recognition. In *ICASSP*, 2018.

Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. In *Interspeech*, 2020.

Yu-An Chung and James Glass. Generative pre-training for speech with autoregressive predictive coding. In *ICASSP*, 2020.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. In *Interspeech*, 2019.

Yu-An Chung, Hao Tang, and James Glass. Vector-quantized autoregressive predictive coding. In *Interspeech*, 2020.

Pavel Denisov and Ngoc Thang Vu. Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning. In *Interspeech*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *NAACL-HLT*, 2019.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

Nathan Duran and Steve Battle. Probabilistic word association for dialogue act classification with recurrent neural networks. In *EANN*, 2018.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *EMNLP*, 2018.

Kyle Gorman, Jonathan Howell, and Michael Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. From audio to semantics: Approaches to end-to-end spoken language understanding. In *SLT*, 2018.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. ODSQA: Open-domain spoken question answering dataset. In *SLT*, 2018.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-Yi Lee. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*, 2018.

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP*, 2020.

Andy Liu, Shu-Wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP*, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pre-training approach. *arXiv preprint arXiv:1907.11692*, 2019.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *ICASSP*, 2015.

Daniel Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Cubuk, and Quoc Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, 2019.

Yao Qian, Rutuja Ubale, Vikram Ramanaryanan, Patrick Lange, David Suendermann-Oeft, Keelan Evanini, and Eugene Tsuprun. Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system. In *ASRU*, 2017.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Suman Ravuri and Andreas Stolcke. Recurrent neural network and LSTM models for lexical utterance classification. In *Interspeech*, 2015.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages. In *ICASSP*, 2020.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, 2019.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. Towards end-to-end spoken language understanding. In *ICASSP*, 2018.

Xingchen Song, Guangsen Wang, Zhiyong Wu, Yiheng Huang, Dan Su, Dong Yu, and Helen Meng. Speech-XLNet: Unsupervised acoustic model pretraining for self-attention networks. In *Interspeech*, 2020.

Gokhan Tur and Renato De Mori. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Weiran Wang, Qingming Tang, and Karen Livescu. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *ICASSP*, 2020.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, 2018.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020a.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. In *AAAI*, 2020b.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. SDNet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*, 2018.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. Make lead bias in your favor: A simple and effective method for news summarization. *arXiv preprint arXiv:1912.11602*, 2019.