ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747759

TIE YOUR EMBEDDINGS DOWN: CROSS-MODAL LATENT SPACES FOR END-TO-END SPOKEN LANGUAGE UNDERSTANDING

Bhuvan Agrawal, Markus Müller, Samridhi Choudhary, Martin Radfar
Athanasios Mouchtaris, Ross McGowan, Nathan Susanj, Siegfried Kunzmann

Alexa AI, Amazon

ABSTRACT

End-to-end (E2E) spoken language understanding (SLU) systems can infer the semantics of a spoken utterance directly from an audio signal. However, training an E2E system remains a challenge, largely due to the scarcity of paired audio-semantics data. In this paper, we consider an E2E system as a multi-modal model, with audio and text functioning as its two modalities, and use a cross-modal latent space (CMLS) architecture, where a shared latent space is learned between the 'acoustic' and 'text' embeddings. We propose using different multi-modal losses to explicitly align the acoustic embedding to the text embeddings (obtained via a semantically powerful pre-trained BERT model) in the latent space. We train the CMLS model on two publicly available E2E datasets and one internal dataset, across different cross-modal losses. Our proposed *triplet loss* function achieves the best performance. It achieves a relative improvement of 22.1% over an E2E model without a cross-modal space and a relative improvement of 2.8% over a previously published CMLS model using L_2 loss on our internal dataset.

Index Terms— Spoken Language Understanding, Signal to Interpretation, End-to-end Neural Model, Cross-modal Learning

1. INTRODUCTION

Spoken language understanding (SLU) is the task of inferring the semantics of user-spoken utterances and is the core technology behind voice assistant systems. The traditional approach to SLU uses two distinct components to sequentially process a spoken utterance: an automatic speech recognition (ASR) model that transcribes the speech, followed by a natural language understanding (NLU) model that predicts the intent given the transcript [1].

An increasingly popular approach is to employ models that predict SLU output directly from a speech signal input [2, 3, 4, 5, 6, 7, 8]. This class of models, also referred to as signal-to-interpretation (S2I) models, are trained in an end-to-end (E2E) fashion to maximize the SLU prediction accuracy. Such models typically have smaller footprints than their cascaded counterparts, making them attractive candidates for performing SLU in resource constrained environments. However, availability of sufficient good quality speech data with associated semantic labels is key to achieving comparable performance to the traditional, cascaded counterparts. A paucity of such datasets becomes a major bottleneck for these SLU systems.

Curriculum and transfer learning strategies are used in [9, 10] to gradually fine-tune the SLU model on increasingly relevant datasets, followed by finally training it on low-resource in-domain data. Authors in [5, 11] leverage the large amount of transcribed speech data to pre-train an ASR model on phoneme and word-level targets. However, a principled study performed by the authors in [12], revealed that most of these methods report high performance on SLU datasets of relatively low semantic complexity, often representing targeted SLU use-cases. As the complexity of the dataset increases and the SLU task becomes more sophisticated, the performance of these models starts degrading.

Authors in [13] attempt to address this problem by combining component pre-training, knowledge transfer, and data augmentation approaches to create a robust SLU model. BERT-based text embeddings in [14] are used to 'guide' acoustic embeddings. An L_2 loss between the text and acoustic embeddings is used to explicitly tie the cross-modal latent representation space, leading to better intent classification accuracy.

In this work, we further explore the multi-modal view of SLU models by experimenting with different approaches to learn a robust cross-modal latent space (CMLS). Specifically, we experiment with different loss functions to tie the acoustic and text embeddings together. Our goal is to have the encoders generate embeddings in the same latent space, so that the origin of the embeddings becomes indistinguishable.

We train the so-called CMLS E2E SLU model architecture defined in [13] on an internal dataset and two publicly available SLU datasets [5, 15], and we study the effect of these cross-modal embedding losses. Our results indicate that the triplet loss has the best performance across both these datasets.

To the best of our knowledge, this is the first attempt to

systematically apply losses that have shown success in learning cross-modal representations in relatively mature textual-visual or speech-visual multi-modal domains to SLU models to improve the SLU tasks. We hypothesize that using an appropriate loss function and cross-modal training methodology is key to achieving tighter coupling and hence better performance, as the SLU tasks increase in complexity.

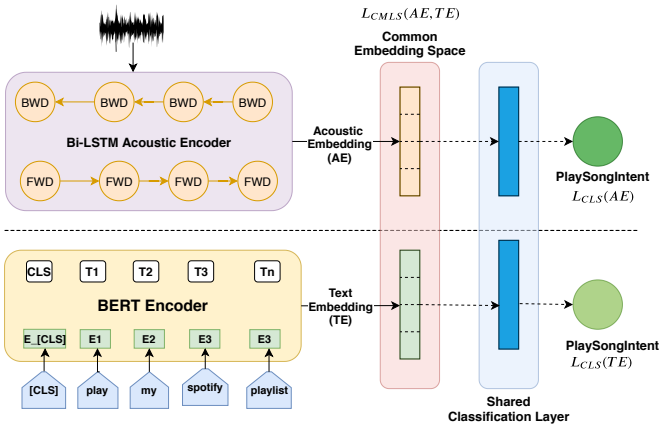

Fig. 1. The CMLS SLU model architecture. $L_{CE}(AE/TE)$ is the cross-entropy classification loss for the acoustic/text embedding, whereas, $L_{CMLS}(AE, TE)$ is one of our cross-modal embedding loss to bring the acoustic and text embeddings together. The text-encoder component below the dotted line is removed for inference

2. CROSS-MODAL LATENT SPACES

A major challenge in building an efficient E2E SLU system is the availability of appropriate datasets. While there is a plethora of ASR and NLU specific datasets available, there is a scarcity of good quality E2E SLU datasets that have paired utterance audio and semantic labels, to allow for an E2E training of the model. An approach to overcome this is building a system that has a tied space for multiple modalities. This presents the advantage of leveraging modality specific datasets that can help achieve better generalization for the final SLU task. The tied cross-modal space can be achieved by unifying representations across modalities. An example would be projecting the acoustic and text-only data into the same embedding space and learning the parameters of this joint space to optimize the SLU task-loss. This enables us to train an SLU model in an E2E fashion on the smaller dataset, but also leverage the large amount of ASR- or text-only datasets to learn a robust latent space for the final task.

One primary first solution to learning such a CMLS between the acoustic and text modalities in an E2E SLU model was proposed in [13]. The text modality latent space is represented by using the encoded representation of an utterance text from a pre-trained BERT model, whereas the acoustic modality is represented by using a multi-layer Bi-LSTM model to create an acoustic embedding of the utterance audio. Along with the task-specific classification loss[1], they use an L_2 loss between the text and acoustic embeddings to explicitly tie the cross-modal space. A shared classification layer is jointly trained on both the acoustic embedding (AE) and text embedding (TE).

While this method of an explicit embedding loss added to the task-specific loss led to an improved performance over the baseline model, there are more effective ways of mapping representations from different modalities into a common space. We employ the same E2E SLU model setup as in [13], but apply three different losses to learn a robust CMLS model. In the next subsection 2.1, we outline our model architecture and describe each loss in detail in subsection 2.2.

2.1. Model Architecture

Our proposed CMLS model consists of three sub-modules as shown in Figure 1: an acoustic encoder, a BERT encoder, and a shared classification layer. The acoustic encoder is a multi-layer Bi-LSTM network that computes the acoustic embedding from the acoustic features of the utterance audio. This is done by max-pooling the last layer Bi-LSTM states across the time dimension to obtain a fixed-dimensional vector that summarizes the input audio, independent of the length of the input signal. In order to obtain the text embedding of the input utterance, we use a pre-trained BERT model that takes the utterance text as an input. As is common in BERT-based encoders, the last layer transformer-encoder representation of the [CLS] token is used as the text-embedding of the utterance. The shared classification layer is a fully-connected network, followed by a softmax to predict the semantic label (intent in our case). It produces an intent prediction using both the AE and TE separately, resulting in computing an embedding specific classification loss $L_{CLS}(AE/TE)$ as shown in Figure 1.

The BERT-based text pipeline is only used during training to guide the AEs and is discarded for inference. This architecture has the advantage of being easily extensible to support more modalities during training while at the same time keeping the resource footprint constrained during inference. The CMLS model is not limited to using a BERT model; the method is agnostic to the type of embeddings used.

2.2. Embedding losses

We evaluate three different loss functions (L_2 loss, pairwise ranking loss, and triplet loss) to tie the embeddings originating from two modalities together into a common latent space.

[1] A cross-entropy loss on intent classification is used.

2.2.1. L_2 loss

Introduced in [13] to tie the latent space, the L_2 loss is computed as follows:

$$\mathcal{L}_E(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||_2^2 \quad (1)$$

where $\mathbf{x}_1$ denotes the acoustic embedding and $\mathbf{x}_2$ denotes the text embedding of the same utterance. $d(\mathbf{x}_1, \mathbf{x}_2)$ expresses the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$.

2.2.2. *Pairwise Ranking Loss*

Moving beyond the L_2 loss, the pairwise ranking loss [16] allows us to train the network using the following loss formulation:

$$\mathcal{L}_E(\mathbf{x}_1, \mathbf{x}_2, t) = td(\mathbf{x}_1, \mathbf{x}_2) + (1-t)\max\{0, m - d(\mathbf{x}_1, \mathbf{x}_2)\} \quad (2)$$

where $\mathbf{x}_1$ denotes the acoustic embedding and $\mathbf{x}_2$ denotes the text embedding. $\mathbf{x}_1$ and $\mathbf{x}_2$ need not necessarily represent the same utterance. t is a binary variable indicating if $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same intent ($t = 1$) or a different one ($t = 0$), and m is the margin which controls the minimum distance between the negative pairs. This is a tunable hyperparameter.

If both samples belong to the same intent, the loss forces them to be closer together (similar to the L_2 loss), but if they originate from different intents, then the loss is $\max\{0, m - d(\mathbf{x}_1, \mathbf{x}_2)\}$, and this pushes the embeddings further apart.

2.2.3. *Triplet Loss*

In contrast to other losses, the triplet loss [17, 18] uses three examples: the current training example, called anchor $\mathbf{x}_a$ and $\mathbf{x}_+$, $\mathbf{x}_-$ are positive and negative examples, respectively. Formula 3 shows how the loss is computed.

$$\mathcal{L}_E(\mathbf{x}_a, \mathbf{x}_+, \mathbf{x}_-) = \max\{0, m + d(\mathbf{x}_a, \mathbf{x}_+) - d(\mathbf{x}_a, \mathbf{x}_-)\} \quad (3)$$

In our case, the anchor is the acoustic embedding and the positive example is the text embedding of an utterance with the same intent. Similarly, the negative example is the text embedding of an utterance with a different intent. For example, the anchor could be the acoustic embedding of 'could you please increase the brightness' and the positive example could be the text embedding of 'it's too dark in here' which both have the `IncreaseBrightness` intent. The negative example, on the other hand, could be the text embedding of 'change the lights to green' which has the `ChangeColor` intent. We also experiment with the bidirectional triplet loss with structure preserving constraints which was introduced in [19].

2.3. Combination of the Losses

The entire model is trained jointly and the total loss can be written as shown in Formula 4, where $\mathcal{L}_{cls}$ denotes the classification loss and $\mathcal{L}_E$ denotes the embedding loss. λ_1 and λ_2 are hyperparameters that control the weights of the different losses in the sum.

$$\mathcal{L} = \mathcal{L}_{cls}^{acoustic} + \lambda_1 \mathcal{L}_{cls}^{text} + \lambda_2 \mathcal{L}_E \quad (4)$$

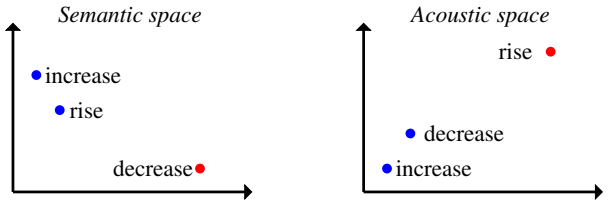

Fig. 2. Leveraging the triplet loss in a CMLS task: In contrast to previously used losses (L_2 loss), the triplet loss not only minimizes the distance between semantically related words but also maximizes the distance between those target words that are acoustically close but semantically opposite.

3. EXPERIMENTAL SETUP

3.1. Datasets

We use an internal SLU dataset along with two publicly available datasets to train and evaluate our model – Fluent Speech Commands (FSC) [5] and Snips SLU *SmartLights* [15]. All three datasets contain utterance text, corresponding audio, and a semantic class label.

The internal dataset is curated by taking a random slice of proprietary real-world, de-identified utterances from interactions with a voice assistant. It contains over 200 hr of audio. We filter the dataset to consist of 15 high frequency intents in order to have a reasonable test bed to the public dataset. We further create a 'hard' subset of the test set that exclusively consists of utterances containing bigrams that do not appear in the training set. This is done to measure the performance of the model on 'never seen' phrases.

FSC is one of the largest public SLU datasets, containing $\approx 30,000$ utterances. Each utterance text is associated with a triplet - action, object, and location. This triplet functions as the intent label of the utterance and becomes our target semantic class to be predicted. There are a total of 31 unique intent classes in the dataset. Snips is a smaller SLU dataset, making the prediction task challenging. It contains $\approx 3,000$ utterances and 6 intent classes. Since this dataset does not have its own train/test/validation splits, we created an 80-10-10 split for train, validation, and test, respectively.

3.2. Model Training and Hyperparameters

For FSC and Snips SLU, we use a 4-layer and 3-layer Bi-LSTM acoustic encoder respectively with 512 hidden units per layer. The text-encoder comprises of a pre-trained BERT model[2] from [20], where the encoded representation of the

[2]We use the BERT-base-cased model.

[CLS] token from the last encoder layer is used as the text-embedding of the utterance. The shared classification layer has an input size of 768 (the size of the embeddings) and the number of outputs depends on the dataset (15 for the internal dataset, 31 for FSC, and 6 for Snips SLU).

In order to show the efficacy of tying the text-acoustic cross-modal embedding spaces, we also train a *baseline model* that just contains the multi-layer Bi-LSTM acoustic encoder with a fully connected layer to perform intent classification without any text encoder present.

4. RESULTS

4.1. Fluent Speech Commands

The results on the FSC dataset are shown in Table 1. We train three different CMLS SLU models on FSC, one per loss type. All the loss types are able to perform better (or equal to, in the case of ranking loss) than the baseline model. Triplet loss achieves the best performance with a relative improvement of 1.4% over the baseline model and 0.4% relative over the L_2 loss model from [13]. When compared to the original accuracy reported on the FSC dataset by the authors in [5] for their non-pre-trained model (96.6%), we see that the triplet loss achieves the best performance here, too, with a relative improvement of 1%. The L_2 loss is also able to beat the reported accuracy from [5] by 0.7% relative. Both these results indicate the importance of having a cross-modal latent space, using a pre-trained text encoder in an E2E SLU model.

4.2. Snips SLU

We repeat the experiments on the Snips SLU dataset, the results of which are shown in Table 1. Similar to our previous observation, the CMLS SLU model, is able to beat the performance of the baseline model, for all the embedding loss types, achieving a relative improvement of 3% on average. The triplet loss has the best performance on this dataset, too, with a relative improvement of 4% over the baseline and 1% over the L_2 loss model from [13]. Since this dataset is a more challenging dataset owing to its small size, these performance improvements are notable.

Table 1. Intent accuracy on the public datasets. The baseline model is just the acoustic encoder with no CMLS while the CMLS model with L_2 loss is similar to the model proposed in [13].

Model	**Dataset**	
	FSC	Snips
Baseline (Only acoustic, no CMLS)	96.31	70.84
CMLS - L_2 loss (model in [13])	97.31	72.89
CMLS - Pairwise ranking loss	96.31	72.29
CMLS - Triplet Loss	**97.65**	**74.10**

4.3. Internal Dataset

The results on the internal datasets are shown in Table 2. We used the triplet loss variations only on the internal dataset since it is much larger than the public datasets available, giving us much more flexibility to experiment with the losses. Vanilla triplet loss achieves the best performance with a relative improvement of 22.1% over the baseline model and 2.8% over the L_2 loss model. The versions of triplet loss which updated the BERT parameters also performed subpar, possibly due to the noise introduced to the pre-trained BERT weights as the model now has to forget and re-learn on a much smaller dataset.

Table 2. Relative intent accuracy on the internal dataset.

Model	**Accuracy (Rel)**	**Hard Test Set Accuracy (Rel)**
Baseline (Only acoustic, no CMLS)	—	—
CMLS - L_2 loss	18.26%	17.02%
CMLS - Triplet loss	$\mathbf{22.1}\%$	$\mathbf{19.71}\%$
CMLS - Triplet loss with BERT gradients	-1.2%	-3.39%
CMLS - Bidirectional Triplet loss with structure preserving constraints	-0.54%	-2.51%

5. CONCLUSION

In this work, we take a multi-modal view of E2E SLU model and propose a CMLS setup wherein the model learns a shared latent space between the two modalities of the SLU model – speech and text. The CMLS setup enables us to achieve a higher SLU performance even on a smaller E2E dataset because the text embeddings are extracted from a BERT-based text encoder that has been trained on massive amount of textual data and has shown to capture the semantics very well across a variety of tasks. We show that triplet loss has the best performance based on our experiments across all the datasets.

6. REFERENCES

[1] Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee, "Recent approaches to dialog management for spoken dialog systems," *Journal of Computing Science and Engineering*, vol. 4, no. 1, pp. 1–22, 2010.

[2] Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann, "End-to-end neural transformer based spo-

ken language understanding," in *Interspeech 2020,[th] Annual Conference of the International Speech Communication Association*, 2020, pp. 500–505.

[3] 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 *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 720–726.

[4] Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua 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.

[5] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[6] Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore, "Spoken language understanding without speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6189–6193.

[7] 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 *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 569–576.

[8] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin, "End-to-end named entity and semantic concept extraction from speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 692–699.

[9] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," *arXiv preprint arXiv:1906.07601*, 2019.

[10] Natalia Tomashenko, Antoine Caubrière, and Yannick Estève, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," 2019.

[11] Swapnil Bhosale, Imran Sheikh, Sri Harsha Dumpala, and Sunil Kumar Kopparapu, "End-to-end spoken language understanding: Bootstrapping in low resource scenarios.," in *Interspeech*, 2019, pp. 1188–1192.

[12] Joseph P. McKenna, Samridhi Choudhary, Michael Saxon, Grant P. Strimel, and Athanasios Mouchtaris, "Semantic complexity in end-to-end spoken language understanding," *arXiv preprint arXiv:2008.02858*, 2020.

[13] Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons, Kartik Audhkhasi, Brian Kingsbury, Ron Hoory, and Michael 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.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alaa Saade, Alice Coucke, Alexandre Caulier, Joseph Dureau, Adrien Ball, Théodore Bluche, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, and Maël Primet, "Spoken language understanding on the edge," 2018.

[16] Yuncheng Li, Yale Song, and Jiebo Luo, "Improving pairwise ranking for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3617–3625.

[17] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, "Large scale online learning of image similarity through ranking," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2009, pp. 11–14.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[19] Liwei Wang, Yin Li, and Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5005–5013.

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, pp. arXiv–1910, 2019.