Analysis of Joint Speech-Text Embeddings for Semantic Matching

Muhammad Huzaifah, Ivan Kukanov

Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
{huzaifah_md_shahrin, ivan.kukanov}@i2r.a-star.edu.sg

Abstract

Embeddings play an important role in many recent end-to-end solutions for language processing problems involving more than one data modality. Although there has been some effort to understand the properties of single-modality embedding spaces, particularly that of text, their cross-modal counterparts are less understood. In this work, we study a joint speech-text embedding space trained for semantic matching by minimizing the distance between paired utterance and transcription inputs. This was done through dual encoders in a teacher-student model setup, with a pretrained language model acting as the teacher and a transformer-based speech encoder as the student. We extend our method to incorporate automatic speech recognition through both pretraining and multitask scenarios and found that both approaches improve semantic matching. Multiple techniques were utilized to analyze and evaluate cross-modal semantic alignment of the embeddings: a quantitative retrieval accuracy metric, zero-shot classification to investigate generalizability, and probing of the encoders to observe the extent of knowledge transfer from one modality to another.

Index Terms: cross-modal adaptation, sentence embeddings, semantic matching, probing classifiers

1. Introduction

The popularity of techniques like Word2Vec [1] and Wav2Vec [2] have shown the usefulness of embeddings, a learned low-level representation of high dimensional inputs, in solving human language processing problems. In recent years, with the push to address complex multimodal problems in an end-to-end manner, learning a common embedding space has been seen as a way of reducing the “modality gap” between speech and text data [3], not only with regards to bridging their continuous and discrete natures, but also in terms of widening data availability and complementing their features. This would enable for example, the combination of both acoustic and semantic information in a single model, which would be greatly benefit speech-based tasks such as speech translation (ST), automatic speech recognition (ASR), and spoken language understanding (SLU).

Indeed many state-of-the-art systems in these fields incorporate some form of a joint embedding as part of the model. Shared encoders have been employed together with distance-based losses to aggregate and learn from both text and speech data in several ST systems [4,5]. Dong et al. [7] instead utilized a separate pretrained BERT language model [8] to provide additional supervision to the speech encoder during training as a way of transferring semantic knowledge from a text source. Various models designed for SLU also take this approach, but treat the BERT supervision as a pretraining step [9][11]. Adversarial methods have also been used to bring text and speech representations closer [12][13].

Despite the resultant cross-modal embeddings proving their usefulness in pushing the state-of-the-art, there has been little work directly analysing their properties outside of evaluating them in the context of the downstream task. These same motivations inspired the body of work on probing [14][16], that is checking for the retention of linguistic features in the embeddings via classifiers. However, existing probing literature mostly focus on pure text data. Outside of the speech-text domain, more of such research has been carried out for text and images [17][19]. The CLIP [19] and ALIGN models [20] in particular demonstrated how a simple training task of matching images to captions, if scaled up, could produce embeddings that generalize well to different tasks and datasets.

In this paper, we study an analogous task for speech and text, which in the broader literature is referred to as semantic matching [21]. The goal is to couple semantically-related objects from both domains by learning to map a speech utterance and its corresponding transcription to the same point in embedding space. Since the final objective is the embedding space itself, this allows us to gather some insight without the distraction of a further downstream task. Here, we employ a dual encoder arranged in a teacher-student setup, where the speech embedding space gradually adapts to the text embedding space. We investigate the extent of semantic alignment and knowledge transfer between a teacher text encoder and a student speech encoder directly through a retrieval accuracy metric and more generally via zero-shot classification of several speech datasets. Additionally, to ascertain the occurrence of transfer learning, we extend the probing analysis technique to cover both text and speech encoders. In general, we complement work related to the empirical analysis of embeddings and extend them to the multimodal context.

2. Methods

2.1. Joint embedding model overview

![Figure 1: Architecture for joint speech-text embedding training. A decoder is added when training for the multitask objective.](image-url)

As illustrated in Figure 1, the overall model contains two pipelines that respectively transform speech and text features derived from sentence-level utterance and transcription pairs into a hidden representation. The text pipeline consists of
a pretrained language model acting as the teacher whereas a transformer-based speech encoder \cite{22} acts as the student. A projection head block was utilized after the speech encoder to maintain consistent dimensions between the speech and text embeddings. The final output of each model was mean pooled over the length of the sentence, and the resultant fixed-length vector normalized to unit length. To train for semantic matching, we minimize the L2 distance between the outputs. Teacher-student learning occurs throughout the training by backpropagating the loss only through the speech pipeline while keeping the weights of the text model fixed. This progressively brings the speech embeddings closer to the space defined by the text encoder to eventually construct a joint space.

To further enhance the representational strength of the speech embedding, we explored two techniques, namely pretraining and multitask training. For pretraining, the speech encoder was initialized with the encoder section of an encoder-decoder model trained for ASR. In contrast, the multitask setup trains the speech encoder for ASR from scratch simultaneous with training for the multitask objective. For pretraining, we explored two techniques, namely pretraining the speech embedding for ASR in our experiments, for which the Fairseq checkpoint with the smallest validation loss was chosen for further analysis. In addition to the Librispeech test sets (libri-clean, libri-other), MuST-C (en-zh) \cite{27}, and CoVoST 2 (en) \cite{28} were also used for evaluation. For these, only the English speech segments and transcriptions were used.

3. Experimental Setup

3.1. Text encoder

A pretrained RoBERTa large model \cite{31} from the Fairseq library \cite{24} was adopted as the backbone of the text pipeline. RoBERTa improves on the masked language modeling objective of BERT \cite{18} by tweaking several hyperparameters, removing the next-sentence prediction objective, and training on a much larger dataset, among other innovations. Before input into RoBERTa, byte pair encoding (BPE) was first applied to the text sequence at byte-level. RoBERTa large contains 24 transformer layers and an output embedding size of 1024.

3.2. Speech encoder

We used the s2t_transformer_s architecture from Fairseq, which consists of several convolutional neural network layers to downsample the input followed by 12 transformer layers, with four attention heads per layer and an output embedding size of 256. The post-encoder projection head contains a Gaussian Error Linear Unit (GELU) activation between two linear layers, followed by dropout and layernorm. The speech pipeline for the multitask objective additionally includes a six layer transformer decoder. To ascertain the impact of pretraining, we compared randomly initialized encoders against encoders pretrained for ASR in our experiments.

3.3. Joint speech-text embedding training

The model was trained for the joint embedding objective on the Librispeech dataset using four NVIDIA V100 GPUs for 100 epochs with Adam optimizer, an inverse square root learning rate scheduler with a peak learning rate of 1e-3, and 10000 warmup steps. For the multitask setup, early experiments showed that the contribution of the cross-entropy loss for ASR was about two orders magnitude larger than the L2 loss for embedding distance. The \( \beta \) parameter was tweaked between 1-100 to increase the influence of the L2 loss. For all training variations, the checkpoint with the smallest validation loss was chosen for further analysis. AudioMNIST contains 30000 audio recordings of spoken digits 0-9 with 50 repetitions per digit per speaker. For labels, digits were spelled out instead of using the numerical form (i.e. “one” rather than “1”). A subset of Speech Commands was used comprising 10 auxiliary words: “Bed”, “Bird”, “Cat”, “Dog”, “Happy”, “House”, “Marvin”, “Sheila”, “Tree”, and “Wow”. Each word is spoken once by each speaker, making up 20408 recordings in total. The SX subset of TIMIT was used to explore joint embeddings with multi-word sentences acting as labels. SX sentences were designed to provide a good coverage of pairs of phones. We randomly selected 10 SX sentences among TIMIT’s TEST set for labels, with seven pronunciations each
Table 1: Retrieval accuracy (%) gauging how close speech-transcription embedding pairs are after joint embedding (JE) training. The last two rows refer to speech encoders in the randomly-initializated and pretrained settings prior to the training, and serve as bottomline comparisons. Subsequent discussion will refer to the model variations by their designated letter.

| Model | speech encoder & training conditions | libri-clean | libri-other | MuST-C | CoVoST |
|-------|--------------------------------------|-------------|-------------|--------|--------|
|       |                                      | T-S  S-T    | T-S  S-T    | T-S  S-T | T-S  S-T |
| A     | random                               | 30.23       | 12.02       | 15.89   | 5.78   |
| B     | pretrained                            | 98.82       | 83.42       | 94.25   | 67.88  |
| C     | pretrained, only projection head      | 98.32       | 63.89       | 91.15   | 47.64  |
| D     | random, multitask, $\gamma=1, \beta=1$ | 84.20       | 23.93       | 65.33   | 16.26  |
| E     | random, multitask, $\gamma=1, \beta=10$ | 88.82       | 29.01       | 72.07   | 20.01  |
| F     | pretrained, multitask, $\gamma=1, \beta=100$ | **99.12**   | **91.91**   | **95.75** | **79.58** |
| G     | random, before JE training           | 0.04        | 0.00        | 0.03    | 0.00   |
| H     | pretrained, before JE training       | 0.04        | 0.08        | 0.03    | 0.00   |

4. Results

4.1. Semantic matching using retrieval

To assess the effectiveness of our setup for semantic matching, we measure the retrieval quality of the model given a speech or text input using cosine similarity as a proxy. This analysis gives us a reasonable impression of how close the final embeddings are. For a given speech input, a prediction is counted if the retrieved text embedding with the highest cosine similarity was chosen as the final model. To probe the speech encoder, the test sentences were first transformed into speech using an open-source text-to-speech system (TTS) by Silero [32], with the speaker model 1.16kzh.

for classification.

The text datasets provided by the SentEval toolkit [14] were used for probing. These span 10 semantic and syntactic tasks, from word-level objectives, such as word constituents, surface information like sentence length, to grammatical structures like past and present tense. We refer the reader to the original paper for further details on each task. Each task is split into 100000 training, 10000 validation, and 10000 test examples. The classifier is a neural network comprising of two linear layers with dropout and a tanh activation. It was trained for 10 epochs and the checkpoint with the highest validation accuracy was chosen as the final model. To probe the speech encoder, the test sentences were first transformed into speech using an open-source text-to-speech system (TTS) by Silero [32], with the speaker model 1.16kzh.

4.2. Zero-shot speech classification

The mechanism for retrieving transcriptions given a speech utterance as in S-T above was repurposed for classification by replacing the transcriptions with class labels. Table 2 summarizes the classification results for the top performing models B and F in the previous retrieval task, together with the untrained bottomline model H in Table 1. The accuracy, in particular on AudioMNIST and Speech Commands, were only slightly better than the bottomline, which itself was close to random choice. However, scores were much higher on TIMIT. To provide more insight into the reasons behind this, speech and label embeddings were plotted using t-SNE in Figure 2.

Distinct phenomena were observed for the three datasets. For AudioMNIST, speech embeddings showed clear separation among the different classes (Fig. 2 left). On the other hand, several label embeddings were found to overlap in the t-SNE plot, suggesting that text encoder RoBERTa was projecting them to a similar point in embedding space. This was a reasonable behaviour given that it was trained to produce contextualized word checkpoin used for pretraining.

The multitask setups which train the speech encoder for ASR from scratch (models D, E), did not perform as well as the models initialized with pretrained checkpoints (models B, C, F), especially for S-T retrieval accuracy. The learned speech features conferred by the ASR pretraining were found to be essential in the initial stages of the joint embedding training, leading to smoother optimization and “closer” final embeddings. The best overall model, model F, combines both pretraining and multitask, showing the benefit of retaining ASR ability concurrent to minimizing the embedding distance.

It is noteworthy that the T-S results were consistently higher than S-T. We suspect this asymmetry stems from the teacher-student learning where the text embeddings remain static but the speech embeddings get pulled apart to fit the space. The final position of the speech embedding plane may leave a particular point closer to its text counterpart compared to other points on the speech plane and hence be retrieved correctly given a text input as in T-S. Yet the same speech point may still be at a distance where it is closer to other unrelated points on the text embedding plane compared to its counterpart, which would make the corresponding S-T retrieval wrong. This phenomenon is exacerbated in training setups with relatively poorer convergence, causing S-T accuracy to deteriorate faster than T-S in those cases. We note that such asymmetry was not reported in works without a teacher-student model such as ALIGN [20].
4.3. Probing tasks

To observe the extent of semantic knowledge transfer taking place, we compare the relative accuracy of RoBERTa against Models B, F, and bottomline H on 10 language probing tasks, shown in Table 3. Firstly, RoBERTa itself was poor at several tasks, especially the ones that required word-level knowledge such as word constituents (WC) and semantic odd man out (SOMO). We attribute this to the mean pooling carried out post encoder, which may have resulted in more ambiguous word representations. Coordination inversion (CoordInv), where the order of coordinate clauses were inverted half the time, may also have been impacted by mean pooling the original embeddings.

Overall, we see some evidence of knowledge transfer of some language properties but not all of them. This is seen in tasks for which the speech encoder gets much closer to the RoBERTa score after training, including for sentence length (SentLen), top constituents (TopConstit) which tests for sentence structure, past and present tense (Tense), and plurality of the subject (SubjNum) or object (ObjNum) in a sentence. Some tasks that do not carry over may have been due to the continuous nature of audio in contrast to discrete text. For example, bigram shift (BShift), which inverts two adjacent words at random, may have been much harder to detect directly from speech compared to text. Additionally, we acknowledge that some degree of error stems from the imperfect TTS conversion to create the dataset. For reference, the transformed SentLen test set had a WER of 26.78 with model F.

5. Conclusions

We studied some properties of a joint speech-text embedding space constructed through a semantic matching objective. The model was trained by minimizing the distance between the embedding outputs of a teacher RoBERTa text encoder and a student speech transformer encoder. We found that ASR pretraining of the speech encoder was essential for better semantic matching, measured in terms of retrieval accuracy. Performance could be further improved by combining it with a multitask objective incorporating both the joint embedding training and further ASR fine-tuning. Using the models for zero-shot classification was found to only work well if both speech and text embeddings manage to capture the properties of the classes. Probing showed that some but not all linguistic properties were transferred from the text to the speech model. In future work, we plan to investigate embeddings and training schemes that would enable semantic coupling between less constrained speech-text pairs compared to utterances and transcriptions, which may be applicable to a wider range of downstream tasks.
6. References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, vol. 26, 2013.

[2] 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, 2020.

[3] Y. Liu, J. Zhu, J. Zhang, and C. Zong, “Bridging the modality gap for speech-to-text translation,” arXiv preprint arXiv:2010.14920, 2020.

[4] R. Ye, M. Wang, and L. Li, “End-to-end speech translation via cross-modal progressive training,” in Interspeech, Aug 2021.

[5] C. Han, M. Wang, H. Ji, and L. Li, “Learning shared semantic space for speech-to-text translation,” arXiv preprint arXiv:2105.03095, 2021.

[6] Y. Tang, J. Pino, X. Li, C. Wang, and D. Gentzel, “Improving speech translation by understanding and learning from the auxiliary text translation task,” arXiv preprint arXiv:2107.05782, 2021.

[7] Q. Dong, R. Ye, M. Wang, H. Zhou, S. Xu, B. Xu, and L. Li, “Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 14, 2021, pp. 12749–12759.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[9] P. Denisov and N. T. Vu, “Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning,” in Interspeech, Oct 2020.

[10] Y.-A. Chung, C. Zhu, and M. Zeng, “SPLAT: Speech-language joint pre-training for spoken language understanding,” arXiv preprint arXiv:2010.02295, 2020.

[11] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-S. Lee, “SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering,” arXiv preprint arXiv:1910.11559, 2019.

[12] Y.-A. Chung, W.-H. Weng, S. Tong, and J. Glass, “Unsupervised cross-modal alignment of speech and text embedding spaces,” Advances in Neural Information Processing Systems, vol. 31, 2018.

[13] A. Alinejad and A. Sarkar, “Effectively pretraining a speech translation decoder with machine translation data,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8014–8020.

[14] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, “What you can cram into a single S&!@* vector: Probing sentence embeddings for linguistic properties,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Jul. 2018, pp. 2126–2136.

[15] Y. Adi, E. Kermery, Y. Belinkov, O. Lavi, and Y. Goldberg, “Fine-grained analysis of sentence embeddings using auxiliary prediction tasks,” in International Conference on Learning Representations (ICLR), 2017.

[16] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. V. Durme, S. R. Bowman, D. Das, and E. Pavlick, “What do you learn from context? probing for sentence structure in contextualized word representations,” in International Conference on Learning Representations (ICLR), 2019.

[17] Y. Aytar, C. Vondrick, and A. Torralba, “See, hear, and read. Deep aligned representations,” arXiv preprint arXiv:1706.00912, 2017.

[18] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi, “Learning cross-modal embeddings with adversarial networks for cooking recipes and food images,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11572–11581.

[19] 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 International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[20] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerg, “Scaling up visual and vision-language representation learning with noisy text supervision,” in International Conference on Machine Learning, vol. 139. PMLR, 24–28 Jul 2021, pp. 4904–4916.

[21] F. Grinchuglu, M. Yatskevich, and P. Shvaiko, “Semantic matching: Algorithms and implementation,” in Journal on data semantics IX. Springer, 2007, pp. 1–38.

[22] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, “faireq s2t: Fast speech-to-text modeling with faireq,” arXiv preprint arXiv:2010.05171, 2020.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RobBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[24] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, “faireq: A fast, extensible toolkit for sequence modeling,” in Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.

[26] 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,” in Interspeech, 2019, pp. 2613–2617.

[27] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, “MuST-C: A Multilingual Speech Translation Corpus,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Jun. 2019, pp. 2012–2017.

[28] C. Wang, J. Pino, A. Wu, and J. Gu, “CoVoST: A diverse multilingual speech-to-text translation corpus,” in Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, May 2020, pp. 4197–4203.

[29] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, “Interpreting and explaining deep neural networks for classification of audio signals,” CoRR, vol. abs/1807.03418, 2018.

[30] P. Warden, “Speech commands: A dataset for limited-vocabulary speech recognition,” arXiv preprint arXiv:1804.03209, 2018.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “TIMIT acoustic phonetic continuous speech corpus,” Linguistic Data Consortium, 1993.

[32] S. Team, “Silero models: pre-trained enterprise-grade STT / TTS models and benchmarks,” https://github.com/snakers4/silero-models, 2021.