SpiderNet: A Modular Approach to Unify Multimodal Modeling

Sanny Kim
Minerva Schools at KGI
sangjin.kim@minerva.kgi.edu

1. Introduction

This paper presents preliminary work on unifying multiple modalities within one model. Based on recent advances in self-attention architectures, we propose a one-encoder-many-decoders Transformer to model multiple modalities such as audio, image, language and music through one unified, but modular architecture.

1.1. Motivation

The advent of Transformers [6] has enabled a multitude of progressions in various modalities, from natural language to image [3] to speech [2]. While these results are promising, they have largely been limited to individual modalities.

1.2. Related Work

Recent works [1, 4] have studied the possibility of unified, multimodal networks. Kaiser *et al*. [1] were the first to use self-attention in multimodal networks through MultiModel. Since then, Pramanik *et al*. [4] have shown promising results with OmniNet, a Transformer capable of solving various tasks within multiple modalities. Lastly, the possibility of using a Transformer for text-to-speech and image captioning tasks have been displayed by Li *et al*. [2] and Yu *et al*. [7] respectively.

2. Method

While the results of MultiModel [1] and OmniNet [4] are encouraging, one potential drawback of these approaches is that the majority of parameters is required to be used across all tasks. This could be a fundamental limitation in circumstances, where only a small fraction of learned parameters are useful for a particular task. Thus, we propose *SpiderNet*, a modular approach to overcome this limitation.

2.1. Initial Experiments

We first tested whether we could train a single encoder sequentially for tasks such as music and language generation. But as common in neural networks, training a model sequentially on considerably different tasks without mechanisms such as episodic memory or replay usually leads to catastrophic forgetting. Thus, we also investigated whether the Transformer could be split into different streams based on modality or task. To test this, we divided the model into individual modality streams such as one music and one language stream. Each stream would then be trained with different sets of attention heads and feed forward networks. While this yielded better results than simply training a model sequentially, this modification can be seen as a fragmented rather than unified method as the model is almost identical to using two separate Transformers, which are used solely for their respective, uni-modal tasks.

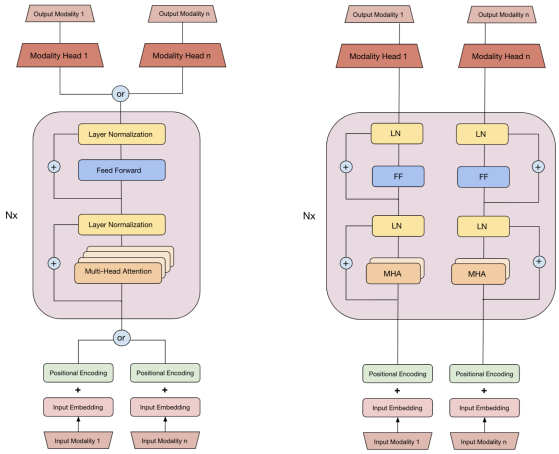

Figure 1. Left shows a Standard Transformer with n input modalities and n output modality heads. Right shows a Transformer split into n modality streams. For the purpose of this figure, $n = 2$.

2.2. Proposed Method

Our approach seeks to take advantage of the Encoder-Decoder structure of the Transformer. While for tasks such as language generation only the encoder is required, for tasks such as image captioning and text-to-speech, in which input and output are not identical, the Transformer has to utilize both the encoder and decoder. Due to this encoder-decoder characteristic and our previous observations about parameter efficiency, we propose a unified encoder with multiple modality streams connected to modular decoders.

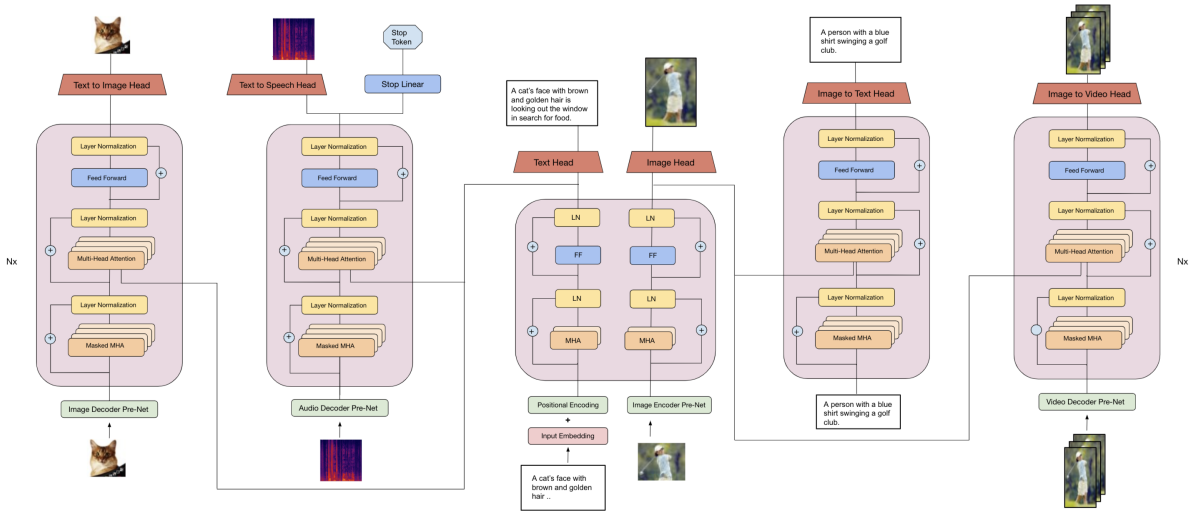

Figure 2. One possible SpiderNet architecture with two modality streams in the encoder and four cross-modal decoders. Images from [5].

This would allow the model to only use the encoding and decoding parameters that are trained specifically for a certain task. Hence, in inference, only a fraction of parameters would be used as only the dedicated encoder stream and decoder module would be used. In training, this means that different modalities as well as tasks can be trained and updated separately. Moreover, this would allow a straightforward addition of modality streams in the encoder. Thus, if we want to add a new modality, we can simply insert a new modality stream within the encoder. And to add a new task, we just append a new decoder and connect it to the corresponding encoder stream. For tasks that share a similar domain, for example, German language generation and German-to-Korean translation, the model would be able to use the encoding stream and share the same parameters. Another possibility for such domains would be to use a methodology similar to curriculum learning, by which the model first learns simpler tasks that require fewer layers. For more difficult tasks, additional layers can then be added and previous layers can be frozen to avoid catastrophic forgetting and enable parameter sharing. During inference, these appended layers would then be also used for their distinct tasks. Finally, as a result of its multi-stream setup, the model is able to receive multimodal input for tasks such as Visual Question Answering.

3. Discussion and Future Work

The described model is one approach to modularize multimodal learning in the decoding process and permits parameter sharing for similar task domains, but lacks cross-domain learning in the encoding part of the model as the streams are disconnected. Furthermore, this work addresses unnecessary weights and continual learning only to a limited extent. Questions such as how to identify a task hierarchy, whether curriculum learning would genuinely benefit training, what constitutes a similar domain and how to avoid catastrophic forgetting when re-training a modality stream on a similar domain still remain to be answered.

In conclusion, this paper proposes one particular Transformer-based approach, but encourages further research in topics such as Lifelong Multimodal Learning, Pruning and Dynamically Adaptable Networks.

References

[1] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv*, 2017.

[2] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. Neural speech synthesis with transformer network. *arXiv*, 2018.

[3] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, ukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv*, 2018.

[4] Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *arXiv*, 2019.

[5] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 2008.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. Attention is all you need. *NeurIPS*, 2017.

[7] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *arXiv*, 2019.