Multimodal fusion for datasets with only one modality

Josiah Bjorgaard[1]

[1]*Syntensor, Inc.*

December 6, 2023

1 Abstract

2 Introduction

The main contribution of this article are

1. x
2. y

3 Related Work

In this work, we develop a model which can create a multimodal embedding space from data with missing modalities through pretraining and inference with missing data. By training on some datasets which have samples spanning two or more modalities, and additional datasets with only one modality, multimodal fusion inference can be made using only a single modality. Therefor the model is called Multimodal Fusion for Datasets with Only One Modality (MFDOOM).

Everything At Once by Svetsova et al. uses an early fusion transformer with combinatorial loss. The fusion tranformer is applied $N(N-1))$ times, sequentially, to formulate embeddings for eatch modality and each pair of modalities. These embeddings are applied in a combinatorial contrastive loss function. This method is unable to scale due to the exponential increase in applications of the model as more modalities are added. Nonetheless, they acheive good results on multimodal retrieval tasks. They do not create unique unimodal representations in their model.

FLAVA uses multiple loss models for data which is missing a modality, for example, for language they use MLM, while for image-text pairs, they use contrastive loss. This approach allows for unimodal and multimodal inference, but it doesn't prepare an embedding space which can function for many different

modalities.v [Why?] They also use initial unimodal encoders which are combined in a multimodal encoder, such that it is a mid-fusion model.

Zhang et al. developed a model which can be trained with missing modality combinations. The model projects unimodal encodings into a modality-aligned feature sapce, and then performs weight-shared dual attention-based prediction of two sets of outputs. The first output is a supervised prediction to class labels, while the second prediction is a supervision of unimodal predictions accross epochs, which they claim improves predictions with unseen modalities.

VATT -

Zorro is a masked multimodal transformer which produces unimodal and multimodal outputs.

4 Methods

4.1 Transformer Architecture

The network architecture is derived from Zorro, a multimodal transformer with masked attention. Zorro allows both unimodal and multimodal inputs without representation collapse by using masked attention. Each modality undergoes self-attention, while cross attention occurs between modalities and a fusion representation. Outputs of the model are include unimodal, fusion, and global representation heads.

Zorro is trained with multimodal benchmarking datasets comprised of video, audio, and text modalities. In this model, we instead embed continuous RNA count measures for spliced and unspliced RNA as two separate modalities. Using the Zorro network architecture, we can include additional data modalities in the future.

4.1.1 Embeddings

4.1.2 Masked Attention

4.2 Training

4.2.1 Contrastive Pre-Training

This network architecture is based on the Zorro framework.

4.2.2 Expression Decoder

4.2.3 Velocity Decoder

4.2.4 Generative Decoder

4.3 Data

4.3.1 Data Sources

4.3.2 Data Preprocessing

5 Acknolwedgments

Thanks to several people for useful discussions and potentially to AWS for support and Trainium usage.

6 References

References

[1] Kasia Zofia Kedzierska, Lorin Crawford, Ava Pardis Amini, and Alex X Lu. Assessing the limits of zero-shot foundation models in single-cell biology. *bioRxiv*, pages 2023–10, 2023.