Challenges of Self-Supervised Learning for Unified, Multi-Modal, Multi-Task Transformer Models

Graham Annett
Computer Science
Boise State University
Boise, USA
grahamannett@boisestate.edu

Tim Andersen
Computer Science
Boise State University
Boise, USA
tandersen@boisestate.edu

Robert Annett
Health Sciences Center
University of New Mexico
Albuquerque, USA
rdannett@unm.edu

***Abstract*—The recent success of multi-modal multi-task transformer models combined with their ability to learn in a scalable self-supervised fashion has presented evidence that omnipotent models trained with heterogeneous data and tasks are within the realms of possibility. This paper presents several research questions and impediments related towards the training of generalized transformer architectures.**

***Index Terms*—multi-modal, multi-task, self-supervised learning**

Short Paper — Research Track: CSCI-RTAI

I. Background

As transformer [24] based language models in conjunction with self-supervised learning has come to the forefront of deep learning research, the application of these models and training methods for multi-modal and multi-task data has become an emergent topic. There are many reasons why an omnipotent model can be intriguing to researchers, the most lofty of these being general human level intelligence, but many researchers see the ability to integrate multiple modalities and multiple tasks as a way to create unified models leverage computation and data to scale to new challenges.

Much like the language based transformers these models have evolved from, they are data-hungry and have found success in part because of their ability to learn in a self-supervised manner. Individually this learning paradigm has been shown to be effective for a variety of domains and tasks [16, 14, 25]. With recent success, it seems likely that these modalities can be combined to train models capable of data and task agnostic goals and capable of unseen downstream tasks [17].

Unified multi-modal, multi-task transformer models are applied to datasets which include heterogeneous inputs and outputs, both of which are dictated by how the self-supervised training is implemented. Typically these models learn from data that is sequentially ordered, but have been applied to non-sequential based tasks as well [5]. Previously models that were capable of these tasks were trained individually (and often different model architectures) for the task at hand and required time consuming curation and annotation of data. These new models differ due to the abundance of unlabeled data they can be trained with and capable of generalizing without the associated data labeling bottleneck. For these previous task specific models which would often use fine-tuning, models pretrained on one particular dataset would suffer from issues related to distributional shift and data-alignment for new downstream tasks [19, 7]. In contrast, the ability for multi-modal multi-task models to generalize well to new downstream tasks and data is possible due to the variety of modalities and amount of scalable data with intrinsic structure they can simultaneously learn from [7].

The ideas that underpin multi-modal multi-task models have been applied in various manners individually for unimodal data and shown to work quite well (e.g. large language models (LLMs) [3, 5]), but researchers have now applied these methods to models that concurrently learn from a variety of datasets and tasks [26, 20]. These models perform well on multiple tasks with minimal extension to the model (e.g. a specialized output decoder layer) or performing well on differing tasks but utilize one unified model architecture [20, 30, 13]. Although the multi-modal distinction may refer to models that use a dataset with multiple modalities on a single task, the ability for a model to use heterogeneous samples which come from discrete datasets is the most intriguing aspect of these new omnipotent models. With this integration of multiple data sources, experimental issues about how best to train them along with technical issues that impact training speed and model complexity have arisen that are more significant than previous transformer models.

II. Problems and Challenges

We discuss several of the open problems and challenges related to training multi-modal, multi-task transformer models in terms of experimental open problems as well as a technical open problem. Experimental open problems refer to topics and questions that may help guide researchers with empirical results for working on multi-modal, multi-task models. As some form of self-supervised training is common for these models, the questions relate to aspects of the modalities which guide decisions about the model architecture and in what manner cross modality interactions will happen.

The technical open problem section mentions one challenge pertaining to multi-modal data dimensionality which potentially hinders multi-modal training and research. This technical issue has workarounds but they are not ideal due to the necessary scale for which self-supervised training of transformer models require.

A. Experimental Open Problems

Experimental open problems refer to topics that have not been fully explored or while explored have yet to be fully understood in the context of multi-task multi-modal transformer based models. These problems can be thought of either in terms of best practices, how the underlying mechanisms relate to the largely unlabeled datasets they are trained with, or in relation to the self-supervised methods of how these models are trained.

1) Cross Modality Interactions: Although integrating multiple modalities is an aspect of multi-modal models in general, for self-supervised learning where the target may be generated in various ways (compared to supervised learning), the decision of where these cross modal interactions take place for the target output is not fully understood. These decisions related to model architecture and model output will be guided by where "fusing" of these modalities happens. Whether there is a best place for this interaction in transformer based models to happen or how these modalities interact are still open problems and probing these different cross-modal interactions would be helpful for guiding architecture developments. Below we give three common ways in which these cross-modality interactions may happen and although for clarity they are discussed as distinct phases, these interactions can be combined in various ways. For these differing types of interactions, there has been no deep comparison about how these interactions may impact the model's abilities and could be explored by comparing model performance depending on where fusing happens.

Model Input (Early Fusion)

Fusing the modalities at model input or prior to input generally relies on projecting each modality to a latent representation within a sample such that modalities have similar inner dimensions. A common way models may do this is by generating an embedding for each modality and then combining the modalities into single sequence. Early fusion can allow for the model architecture to be a single-stream whereby the learned parameters for the model are trained and shared amongst all data modalities. [17] [20] [26] [15]

Intermediate Representations (Mid-Fusion)

Cross-modality interaction during the intermediate layers can be done by simple operations (adding, concatenating) but recent work has used cross-attention between modalities [9].

Late Fusion

With late fusion, the model is effectively processing each modality in a modality specific stream until the models output or final layers where the streams are combined. Where there is no interaction between modalities until the end, this is similar in many ways to learning a discrete model for each modality. Late fusion may imply that modalities will have differing levels of expressiveness due to computation and dedicated parameters for modality specific streams.

2) Self-Supervised Data Alignment: While the majority of multi-modal transformer based pre-training works in a self-supervised fashion, most of the research for previous multi-modal models has relied on text-image pairs of well-aligned data [23, 27, 21]. For new unified, task agnostic models, the nature of this data is much more likely to be weakly aligned[29] and may either contain samples within a batch that are distributionally very different (for instance one sample being a video-audio pair and another an observation from an offline reinforcement learning dataset), or due to the flexibility of how self-supervised targets can be generated [7], different targets applied for similar data (e.g. one sample of image-text may be captioning, another may be an image with a related observation-action pair).

One aspect that has yet to be rigorously studied for multi-modal self-supervised trained models is the relative importance of each modality during training. Some preliminary research [4] gives credence to the notion that text may be more attended to for transformer based multi-modal models but this has yet to be explored fully when including more modalities and differing dataset sample weights. Another way this could be examined for single-stream models is by utilizing pre-trained embeddings for some modalities while other modality embeddings are trained from scratch and comparing relative trade-offs on training time and computational costs.

3) Transfer-ability: While LLMs have been shown to be successful when training on downstream tasks (along with zero-shot and few-shot learning), the problem of how well these multi-modal and task-agnostic models may transfer to new problems is largely under-explored [18]. Researchers have begun to explore how well a model may generalize to new unseen tasks [17, 26], but the modality of these downstream tasks has been mostly constrained to text based tasks that are known to be possible under previous LLMs [3]. Presumably this could be explored with new test-set only benchmarks [8] but the breadth of these benchmarks should be expanded to give a better indicator of the true generality of unified models [8].

Recent research has given insight into trade-offs relative to scaling of compute versus data for LLMs [11]. In the case of applying self-supervised training to multiple datasets in unison, scaling a particular dataset or modality of interest may not be optimal or feasible due to limited availability of data. Many convolutional based networks are trained in a supervised fashion on large datasets and fine-tuned for downstream tasks where less data is available, but outside of language based transformers this same process has not been deeply studied for multi-modal multi-task transformer models. The ability to scale distributionally similar data with self-supervised learning and fine-tune a model on other downstream tasks seems plausible given recent success [17, 13, 20] and could allow researchers to utilize a universal pre-training schema [6] to reach local minima in relation to the model and data.

B. Technical Open Problems

While there are various technical problems related to transformer models (i.e. altering some part of the attention mechanism [28] or how a larger transformer can be distilled to

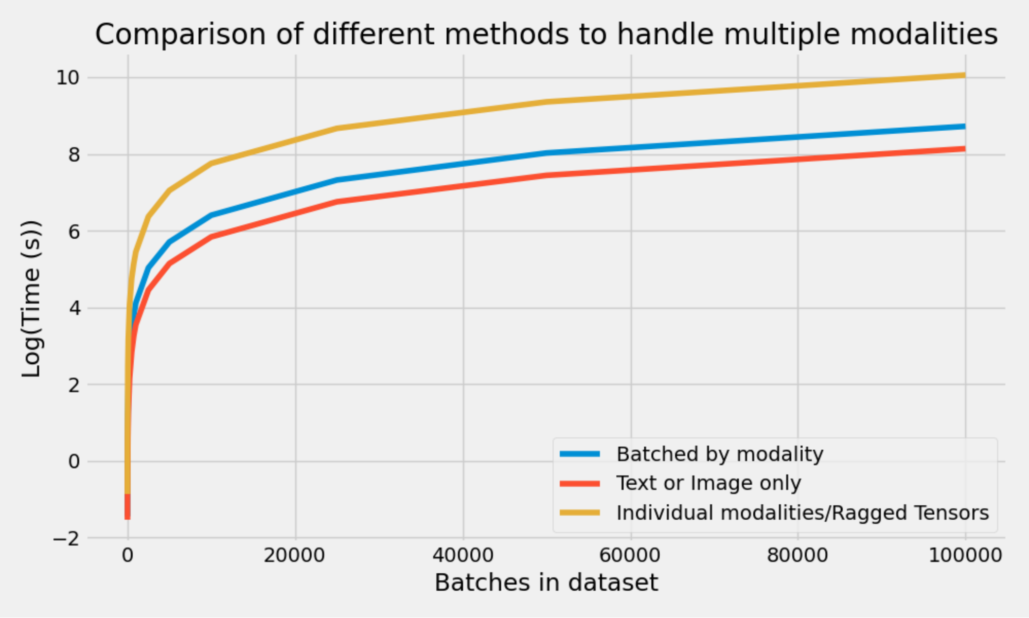

Fig. 1. Comparison of speeds when processing multi-modal data with differing methods.

a smaller model [22]), the issue of dealing with heterogeneous modalities and samples is the most relevant as this problem is greatly exacerbated for research and development of multi-modal multi-task self-supervised learning.

1) Handling/Processing Multi-Modal Data: When using multiple datasets where samples have differing modalities, various decisions about how to handle the data must be made and requires data loading and collation to be carefully considered when compared to using unimodal data. For language only self-supervised training, using the data generally entails tokenizing then masking in batches, which due to the the uniformity of the data, allows for it to be padded and contained within multi-dimensional tensors of a maximum sequence length. This processes allows for the parallelization of data and reduces slowdowns from issues related to data bandwidth (i.e. to and from GPU) and computational overhead [10]. Training with multiple modalities may still follow a similar process, but the inability to combine disparate modalities easily in parallelizable tensors results in slower training and processing when compared to unimodal data. Differing research approaches to this problem have included the following:

Handle modalities independently By handling modalities independently, this is effectively equivalent to a forward and backward pass with a batch size of one until the modality fusing takes place.

Batch by samples with similar modalities This entails that per batch modalities are the same. Because the model can process each modality individually until the cross-modality interaction takes place, this is similar to how multi-modal training (when all contains the same number of modalities) is currently done [1]. It is unclear how the backward pass of a training loop may be impacted when per batch modality contains only similar types of samples.

Use precomputed embeddings By using embeddings that are not learned directly (e.g. pretrained image feature extractors or 3rd party APIs), we save on computational costs related to training embedding sections of the model [12, 15]. These embeddings may still have dimensionality mismatch and still need to be transformed (either via preprocessing or through the model).

In figure 1 we show how naive implementations of handling multi-modal data will have different speeds solely related to transferring data to and from the GPU. Without advanced knowledge of how best to handle heterogeneous data in a training loop, these differences will become amplified and slow model training further. While the methods mentioned in II-B1 have sufficed so far, new work related to handling variable sized tensors could improve upon this handling while also easing the development of models handling multi-modal data and be more on-par with training times associated with unimodal training. These developments would utilize functionality built into deep learning frameworks but need further development to be fully usable in this manner. For TensorFlow and PyTorch the name of this functionality is RaggedTensors and NestedTensor respectively. While parts of these are both actively being developed, this along with ideas from other frameworks [2] which put priority on high performance optimized code will greatly speedup development and training for multi-modal self-supervised learning.

III. Conclusion

By highlighting some of these challenges, our aim is to give an approachable manner in which research can be furthered and future results will elucidate further issues and paths forwards for omnipotent models that are capable of generalizing to any task.

REFERENCES

[1] Jean-Baptiste Alayrac et al. *Flamingo: A Visual Language Model for Few-Shot Learning*. Apr. 29, 2022. arXiv: 2204.14198 [cs]. URL: http://arxiv.org/abs/2204.14198 (visited on 06/26/2022).

[2] James Bradbury et al. *JAX: Composable Transformations of Python+NumPy Programs*. Version 0.3.13. 2018. URL: http://github.com/google/jax.

[3] Tom B. Brown et al. *Language Models Are Few-Shot Learners*. July 22, 2020. arXiv: 2005.14165 [cs]. URL: http://arxiv.org/abs/2005.14165 (visited on 09/19/2022).

[4] Jize Cao et al. *Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models*. July 18, 2020. arXiv: 2005.07310 [cs]. URL: http://arxiv.org/abs/2005.07310 (visited on 09/06/2022).

[5] Alexey Dosovitskiy et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021. arXiv: 2010.11929 [cs]. URL: http://arxiv.org/abs/2010.11929 (visited on 09/22/2022).

[6] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. *How Well Do Self-Supervised Models Transfer?* Mar. 29, 2021. arXiv: 2011.13377 [cs]. URL: http://arxiv.org/abs/2011.13377 (visited on 09/29/2022).

[7] Linus Ericsson et al. "Self-Supervised Representation Learning: Introduction, Advances and Challenges". In: *IEEE Signal Processing Magazine* 39.3 (May 2022), pp. 42–62. ISSN: 1053-5888, 1558-0792. DOI: 10.1109/MSP.2021.3134634. arXiv: 2110.09327 [cs, stat]. URL: http://arxiv.org/abs/2110.09327 (visited on 09/06/2022).

[8] Tanmay Gupta et al. *GRIT: General Robust Image Task Benchmark*. May 2, 2022. arXiv: 2204.13653 [cs]. URL: http://arxiv.org/abs/2204.13653 (visited on 09/19/2022).

[9] Tanmay Gupta et al. *Towards General Purpose Vision Systems*. Apr. 19, 2022. arXiv: 2104.00743 [cs]. URL: http://arxiv.org/abs/2104.00743 (visited on 09/19/2022).

[10] Horace He. "Making Deep Learning Go Brrrr from First Principles". In: (2022). URL: https://horace.io/brrr_intro.html.

[11] Jordan Hoffmann et al. *Training Compute-Optimal Large Language Models*. Mar. 29, 2022. arXiv: 2203.15556 [cs]. URL: http://arxiv.org/abs/2203.15556 (visited on 09/06/2022).

[12] Zhicheng Huang et al. *Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers*. June 22, 2020. arXiv: 2004.00849 [cs]. URL: http://arxiv.org/abs/2004.00849 (visited on 09/22/2022).

[13] Andrew Jaegle et al. *Perceiver IO: A General Architecture for Structured Inputs & Outputs*. Mar. 15, 2022. arXiv: 2107.14795 [cs, eess]. URL: http://arxiv.org/abs/2107.14795 (visited on 07/01/2022).

[14] Salman Khan et al. "Transformers in Vision: A Survey". In: *ACM Computing Surveys* 54 (10s Jan. 31, 2022), pp. 1–41. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3505244. arXiv: 2101.01169 [cs]. URL: http://arxiv.org/abs/2101.01169 (visited on 09/29/2022).

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision*. June 10, 2021. arXiv: 2102.03334 [cs, stat]. URL: http://arxiv.org/abs/2102.03334 (visited on 09/22/2022).

[16] Tianyang Lin et al. *A Survey of Transformers*. June 15, 2021. arXiv: 2106.04554 [cs]. URL: http://arxiv.org/abs/2106.04554 (visited on 09/29/2022).

[17] Jiasen Lu et al. *Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks*. June 17, 2022. arXiv: 2206.08916 [cs]. URL: http://arxiv.org/abs/2206.08916 (visited on 09/12/2022).

[18] Sharan Narang et al. *Do Transformer Modifications Transfer Across Implementations and Applications?* Sept. 10, 2021. arXiv: 2102.11972 [cs]. URL: http://arxiv.org/abs/2102.11972 (visited on 08/17/2022).

[19] Oxford VGG. *Self-Supervision as a Path to a Post-Dataset Era - Alexei Alyosha Efros*. Sept. 1, 2020. URL: https://www.youtube.com/watch?v=iTbfEXFwDJc (visited on 09/06/2022).

[20] Scott Reed et al. *A Generalist Agent*. May 19, 2022. arXiv: 2205.06175 [cs]. URL: http://arxiv.org/abs/2205.06175 (visited on 06/01/2022).

[21] Jabeen Summaira et al. *Recent Advances and Trends in Multimodal Deep Learning: A Review*. May 24, 2021. arXiv: 2105.11087 [cs]. URL: http://arxiv.org/abs/2105.11087 (visited on 09/06/2022).

[22] Hugo Touvron et al. *Training Data-Efficient Image Transformers & Distillation through Attention*. Jan. 15, 2021. arXiv: 2012.12877 [cs]. URL: http://arxiv.org/abs/2012.12877 (visited on 09/30/2022).

[23] Yao-Hung Hubert Tsai et al. "Multimodal Transformer for Unaligned Multimodal Language Sequences". Version 1. In: (2019). DOI: 10.48550/ARXIV.1906.00295. URL: https://arxiv.org/abs/1906.00295 (visited on 09/28/2022).

[24] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 09/22/2022).

[25] Prateek Verma and Jonathan Berger. *Audio Transformers:Transformer Architectures For Large Scale Audio Understanding. Adieu Convolutions*. May 1, 2021. arXiv: 2105.00335 [cs, eess]. URL: http://arxiv.org/abs/2105.00335 (visited on 09/29/2022).

[26] Peng Wang et al. *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework*. June 1, 2022. arXiv: 2202.03052 [cs]. URL: http://arxiv.org/abs/2202.03052 (visited on 09/16/2022).

[27] Peng Xu, Xiatian Zhu, and David A. Clifton. *Multimodal Learning with Transformers: A Survey*. June 13, 2022. arXiv: 2206.06488 [cs]. URL: http://arxiv.org/abs/2206.06488 (visited on 08/09/2022).

[28] Shuangfei Zhai et al. *An Attention Free Transformer*. Sept. 21, 2021. arXiv: 2105.14103 [cs]. URL: http://arxiv.org/abs/2105.14103 (visited on 09/11/2022).

[29] Xunlin Zhan et al. *Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-modal Pretraining*. Aug. 9, 2021. arXiv: 2107.14572 [cs]. URL: http://arxiv.org/abs/2107.14572 (visited on 08/18/2022).

[30] Xizhou Zhu et al. *Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks*. Dec. 2, 2021. arXiv: 2112.01522 [cs]. URL: http://arxiv.org/abs/2112.01522 (visited on 09/09/2022).