Efficient Training & Inference with Joint Multimodal Models for Vision and Language

Motivation: Distributional semantics have been critical to the development of state-of-the-art neural language models. However, by relying solely on lexical context, current models neglect the fundamental relationship between language and the physical world [4]. For example, based on word co-occurrences and the statistical distribution of language, there is little to distinguish directions, such as left and right, or colors, such as red and orange. However, the differences between these concepts are obvious once allowed to account for their perceptual properties. Joint training of multimodal models over both images and language carries the potential to ground language representations with richer world semantics.

Intellectual Merit: In order to develop systems able to interact with their environments, models must be able to reflect and reason about the real world relationships between objects in language. Furthermore, for these models to be applicable to practical settings, such as on edge devices like mobile phones, they must not impose excessive computational costs in either training or inference that would be prohibitive to development or deployment. **The goal of this proposal is to increase the computational efficiency of multimodal models for vision and language and to improve the quality of their cross-modal representations.** In particular, this proposal will present: (1) efficient learning algorithms to reduce the computational overhead of training multimodal models and (2) adaptive techniques for accelerating inference predictions.

Research Plan: Efficient Multimodal Training Algorithms & Architectures Many current approaches to multimodal modeling of language-and-vision leverage transformer-based neural architectures and generative pretraining objectives to learn cross-modal representations from input image and language embeddings. Generative training objectives, such as masked language modeling and masked object prediction, learn by predicting a hidden word or object in the context of a larger image or text sequence. Despite widespread success, there are two fundamental limitations to current approaches: (1) Inefficient usage of data by the generative pretraining objectives and (2) inefficient parameter usage by the transformer architecture's selfattention mechanism, which is quadratic-complexity in terms of sequence length. My focus is to improve both the data and parameter efficiency of these models. To improve the data efficiency, I will build on recent innovations in language-only training which have shown that discriminative training objectives achieve comparable performance while using substantially less computing resources [1]. Models trained with the masked language modeling objective are data inefficient as they only learn from the subset of tokens that are masked. I propose a new discriminative training objective composed of replaced word detection and replaced object detection which will allow models to learn from the entire sequence. In my proposed objective, the model will use its cross-modal representations to predict which input words and image features were corrupted with synthetic replacements. In this setting, the model is learning across all features from both modalities which will enable it to fully leverage each training iteration. To improve the parameter efficiency of these models, I will employ techniques to make use of the self-attention matrix's sparsity and low-rank structure to remove the dependency on sequence lengths, e.g. [5]. By computing self-attention in lower dimensions, these techniques reduce the number of parameters needed.Multimodal models offer new opportunities for dimensionality reduction by taking into consideration interactions between modalities, e.g. computing sparse attention over one modality by conditioning on representations from the other. These improvements will significantly reduce the cost of training multimodal models via reductions in the number of training iterations and in the cost of each pass through the network.

Research Plan: Adaptive Computation for Multimodal Inference: While model training demands a substantial upfront cost, the large size of models and long wall-clock times needed for inference can impose prohibitive barriers to using these models on hardware-constrained edge devices. Existing methods for performing inference fail to account for the difficulty and structure of queries which leads models to perform the same amount of computation regardless which example is being processed. To accelerate multimodal inference computations, I will explore techniques that dynamically adjust both the number of layers and number of features used in a prediction based on the difficulty and structure of a

query. Previous adaptive computation techniques for language-only models have shown that the number of network layers can be adjusted based on the difficulty of samples to reduce the required inference-time computation [2]. I will extend these techniques to the multimodal setting with early exiting strategies to perform predictions without needing a full forward pass through the network. To inform these strategies, I will use model probing techniques to analyze the importance of the individual modalities at each layer by examining the magnitude of network activations. Based on both the downstream task and the intermediate representations, I will add a classifier to each layer in the architecture that both outputs an intermediate prediction alongside a confidence estimate for the prediction quality. To take advantage of the structure of multimodal queries, I will design prefiltering strategies to remove extraneous visual features from inference predictions. Instead of computing predictions using all visual features, I will introduce a lightweight classifier which predicts the utility of each visual feature to the query text and filters for the most relevant features. By reducing the number of input features, this technique will also reduce the ensuing models size and total computing requirements.

Techniques & Evaluation: Models will be evaluated on image captioning (e.g. Conceptual Captions, COCO) and visual question-answering (e.g. VQA, GQA). Additionally, models will be evaluated on visual reasoning tasks (e.g. VCR, NLVR ಠ) which tests its ability to align images to text and its capacity for judging relationships between objects in an image. Computational efficiency will be evaluated by examining performance under fixed computational budgets (wall-clock time and FLOPS). To evaluate efficiency improvements, I will compare the wall-clock time and FLOPS needed to achieve a fixed validation accuracy on both the pretraining objective and the downstream task for a baseline model and one with my proposed modifications. Broader Impacts My goal is to enable the democratization and ubiquity of machine learning technologies without incurring the substantial real-world costs which come along with naive scaling. Improvements in computational efficiency will make onboard machine learning feasible for hardware-constrained devices, such as robots and mobile phones. Adaptive inference will enable model size to be dynamically adjusted to meet the limitations of a specific application. The lowered costs of inference will allow multimodal models to be embedded across a variety of edge devices and provide these devices with the capacity for grounded language understanding. By understanding the relationships [between] language and objects in their environment, these systems will be able to reason and communicate about their environments with human users. Such capabilities will facilitate new interfaces between intelligence systems and humans, empowering applications such as: assistive technologies for the visually impaired and language-based navigation.

Crucially, improvements in computational efficiency will combat the increasingly prohibitive financial and environmental costs of neural network development through reductions in the required training iterations and corresponding power consumption. The ballooning size of neural language models has culminated in enormous and expensive state-of-the-art models that consist of hundreds of billions of parameters and require millions of GPU-hours to train [3]. For example, training of OpenAI's GPT-3 model is estimated to have cost tens of million of dollars in computing resources and nearly 200,000 kWh in power. These prohibitive training requirements have real world consequences, imposing heavy financial burdens on research institutions and entailing substantial greenhouse gas emissions. Increased training efficiency will remove the financial barrier to training and allow more institutions and practitioners to engage in model research and development along with reducing the negative environmental impact of this research.

[1] Kevin Clark et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *ICLR. 2019.*
[2] Maha Elbayad et al. "Depth-adaptive transformer". In: *arXiv preprint arXiv:1910.10073 (2019).*
[3] Emma Strubell et al. "Energy and Policy Considerations for Deep Learning in NLP". In: *ACL. 2019.*
[4] Yonatan Bisk et al. "Experience grounds language". In: *EMNLP. 2020.*
[5] Nikita Kitaev et al. "Reformer: The efficient transformer". In: *ICLR. 2020.*
[6] Gabriel Ilharco et al. "Probing Text Models for Common Ground with Visual Representations".