Introduction

Transformers have recently created a lot of excitement in visual representation learning. Several new transformer-based architectures have been very quickly iterated upon, in the short time span of the last 10 months since the public distribution of Vision Transformers. Aimed to minimize strong inductive biases such as convolution, the transformer architecture is a refreshingly new framework to learn from visual data such as videos and images. In the last year, we've made instrumental contributions in progressing the state of the art in Transformer models with our Multiscale Vision Transformer (MViT) models. MViT amalgamates the seminal idea of hierarchical visual representation in the transformer architecture, allowing much more accurate, flop and data efficient learning from both videos and images.

In the process of exploring the transformer architectural landscape during MViT explorations, we also found several other promising threads of both algorithmic and model improvements that can further significantly improve the capabilities of current transformer models in computer vision. These directions include joint co-training with both images and videos in the same training loop, top-down modulation of structured information in the earlier layers modulated by the deeper layers and more. In this following year, we will further explore these directions and continue improving video representation learning with the rich and successful collaboration that has produced the Multiscale Vision Transformer in the past.

Related Work

We start with a brief review of the recently proposed video transformer architectures and discuss how they relate to solving larger problems in video recognition such as robust co-learning of nouns and verbs and long-term video understanding.
Much of the recent excitement of Transformers in vision can be credited to the work of Dosovitskiy et al. on Vision Transformers [1]. Dosovitskiy et al. demonstrate that the powerful Transformer architecture [2] that has revolutionized the field of NLP is also extremely effective for visual recognition tasks and can perform competitively to well established and mature convolutional neural networks. Subsequently, a flurry of papers have iterated over several design choices such as training recipes [3], patch embedding stem [4, 5], attention patterns [6], convolutional biases [7, 8], feature transpositions [9] and more. These advances enable training Transformers to state of the art accuracies across all major visual recognition benchmarks with minimal or no pre-training required [6,12].
Transformer based architectures allow modelling long term dependencies in data directly owing to the global attention mechanisms in transformer blocks [10,11]. Further owing to the flexibility with respect to the sequence length, Transformers can process different length inputs such as videos and images with minimal model surgery. Also, with the proposed hierarchical representation structure in MViT, transformers can develop a multi-scale feature stack that can be used to provide structured low level but high resolution signal directly into the deeper high level but low resolution features. All these observations enable different exploration directions which are discussed next.

Proposed Method

We aim to exploit the above mentioned synergies between long standing video recognition problems and architectural design of Multiscale Vision Transformers to significantly advance state-of-the-art on video recognition and detection benchmarks. Concretely, we want to explore the following directions:

(a) **Image and Video co-training:** Currently, there exists a dichotomy in video model for recognition. State-of-the-art methods such as ViViT [13] and TimeSformer [14] pretrain on large datasets like IN21k

and finetune on Kinetics 400. Because of short finetuning, these models are cheaper to train. However, they lack a robust notion of temporal cues and hence are *bag of frame classifiers in disguise*. On the other hand, MViT [12] trains from scratch on K400 and achieves competitive performance as well. Further, the trained models exhibit robust dependence on time. However, due to lack of direct supervision on objects, the models lack a good sense of space/objects. We aim to merge this dischomoy by training models with a robust sense of both space/objects and time/actions.

In general, co-training a single model on both images and videos runs into either issues of heavy model surgery such as separate spatial and temporal convolutional kernels specific to each modality or falls into trivial solutions such as treating either a video as a bag of frames or an image as a static video (with frame replication). MViT however is invariant to the length of input sequence and hence since through patchificiation, a video only results in a longer sequence than an image, a single transformer model can train jointly on image and videos within the same loop. This can lead to remarkably more effective video models that effectively fight scene bias by focusing simultaneously on objects and actions and improve performance significantly.

(b) **Top-down supervision**: MViT develops a hierarchical feature stack through gradual channel upsampling and resolution downsampling. While this brings in the seminal ideas of hierarchical features to the Transformer backbone, losing high resolution in important areas such as hands and face early on in computation can be detrimental to action understanding. Consider for example, the complementary actions of pickup and put down. Since most of the visual signal is very similar between these actions, intricacies of the hand pose plays a crucial role in distinguishing them. While high resolution information is present in lower layers, it gets pooled in the residual connections as well as in the main backbone before additional computation. In contrast, having high resolution on the entire image present for more layers is computationally expensive and leads to a bad compute accuracy tradeoff.
Hence, we propose to modulate information processing in later layers with the use of unpooled high resolution information from lower layers directly. Such a mechanism can unblock the model from treating every image pixel equally and allow spending additional compute on image areas such as hand and faces where high resolution details matter more. This can be instantiated as an unpooled attention mechanism directed from higher layers onto the lower layer tokens that are selected according to prior structured information such as hand/object/face boxes or keypoints. More complex visual info such as 3D pose can also supervised this mechanism either implicitly through loss on learnt attention masks or explicitly through extracted feature maps. Such selective information processing, can improve performance on the long tail of action recognition classes where the overall RGB scene is not as effect an indicator as subtle high resolution details such as pointing, throwing, pickup, putdown, opening/closing an object etc.

Milestones

(a) **Image and Video co-training:** 1. Setup separate baselines for both datasets ImageNet (IN) and Kinetics (K400) and both training settings -- when trained from scratch and fine tuned. [**DONE**]
2. Setup a joint model for co-training jointly and reproduce accuracies of each of the separate models on similar setups (ie. SOTA performance on IN/K400 on the same model) [**DONE**]
3. Prepare training recipes for scaling up co-training to larger datasets such as IN21k and K700
4. Exploit the distribution overlap between IN and Kinetics for improved transfer performance than pre-training on ImageNet and fine tuning on Kinetics.
(b) **Top-down supervision**: 1. Extract face and body keypoints/boxes for Kinetics & AVA [**DONE**]
2. Implement structured information extraction from low level high resolution features
3. Implement modulation mechanism between extracted low level features & low rez deeper features.
4. Train on K400 and finetune on AVA aiming to significantly improve transfer on long tail classes.

References:

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[3] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." arXiv preprint arXiv:2012.12877 (2020).

[4] Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." arXiv preprint arXiv:2101.11986 (2021).

[5] Xiao et al. “Early Convolutions Help Transformers See Better” rXiv preprint arXiv:2106.14881 (2021).

[6] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).

[7] Graham, Ben, et al. "LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference." arXiv preprint arXiv:2104.01136 (2021).

[8] d'Ascoli, Stéphane, et al. "Convit: Improving vision transformers with soft convolutional inductive biases." arXiv preprint arXiv:2103.10697 (2021).

[9] El-Nouby, Alaaeldin, et al. "XCiT: Cross-Covariance Image Transformers." arXiv preprint arXiv:2106.09681 (2021).

[10] Dai, Zihang, et al. "Transformer-xl: Attentive language models beyond a fixed-length context." arXiv preprint arXiv:1901.02860 (2019).

[11] Rae, Jack W., et al. "Compressive transformers for long-range sequence modelling." arXiv preprint arXiv:1911.05507 (2019).

[12] Fan, Haoqi, et al. "Multiscale vision transformers." arXiv preprint arXiv:2104.11227 (2021).

[13] Arnab, Anurag, et al. "Vivit: A video vision transformer." arXiv preprint arXiv:2103.15691 (2021).

[14] Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?." arXiv preprint arXiv:2102.05095 (2021).