Visual Learning and Recognition Course Project

Team:
Jay Maier
Nishanth Mohankumar

Project goal: Investigate adapting video generation models as foundation models for image segmentation tasks.

Motivation: Why is it important and relevant? Why should we care?
Large-scale pretraining on unlabeled data has revolutionized the language modeling world over the last five years. The vision world has leaned in the same direction by training models based on multi-modal data (i.e. CLIP, BLIP, etc), but these approaches still rely on human labeling effort (albeit in the form of organic image captions instead of intentionally labeled data), and as such suffer from issues with labeler intent/bias as well as fundamentally ambiguous text-image pairs. In contrast, since there is no ambiguity in pixel values of the real videos which comprise the training examples, we envision video generation as a fundamentally well-formed task for generative pre-training of a visual modeling system. Furthermore, in order to succeed at the task of video generation, a model needs to implicitly represent information relevant to many downstream vision tasks.

Take for example, generating a convincing video of a tennis match. In order to generate a video of a tennis match, a model needs to somehow implicitly represent which features should stay still and which should move with the tennis players (i.e. instance segmentation) as well as represent which pixels should move like a human and which should move like a tennis ball (i.e. mask classification). Additionally, the model needs to encode relative depth in order to handle occlusions naturally and needs to encode some physical dynamics in order to make the ball appear to bounce convincingly (depth perception). In short, a model which is able to generate coherent videos of dynamic scenes must contain within it representations which encode information relevant to a number of other downstream vision tasks (i.e. segmentation, detection, classification, etc), but it does so without being trained on data which is labeled for any of those tasks.

This is a compelling insight because if the key to true visual understanding turns out to lie in pretraining models at video generation tasks, there is plenty of relevant training data. It is estimated that ~80% of internet traffic is in video form. In fact, the common crawl text dataset (which forms the training set for many generative-pre-training schemes in the language domain) is on the order of a few hundred terabytes, while nearly 200 million terabytes of video data are created each day. Our goal for the project is to investigate the potential of open source video generation models as encoders which can be adapted to downstream tasks such as segmentation, detection, and classification.

Specifically, we intend to focus this effort on the task of instance segmentation on the COCO Dataset. Instance segmentation is a cornerstone of computer vision, underpinning a multitude of applications across diverse domains such as autonomous vehicle systems, medical image segmentation, etc. Its ability to understand the complexities of visual data enhances the capabilities of systems in recognizing objects, understanding scenes and implicit image information.

Prior Work: Briefly mention related works that are relevant to your idea instead of covering all related works.

This work most directly references the MAGVIT (Masked Generative Video Transformer) system[1]. MAGVIT is a single model designed for video generation and manipulation tasks and currently is ranked as SOTA on a number of different video generation and representation challenges. Along with MAGVIT, we also intend to survey other top performing model architectures in video representation tasks, such as MaskGIT[2], TATS[3], and WALT[4]. Beyond the realm of video generation, our approach draws inspiration from work in masked image modeling (such as EVA[5]) and traditional detection models (such as DETR-based systems[6]).

Your idea: What is the idea?
We plan to train the 3D-VQVAE encoder network described in the MAGVIT paper on video sequence data. We will then freeze the weights of the encoder network and create a "segmentation head" which predicts segmentation masks from the latent space of the video encoder network. We will train this segmentation head to predict masks on the COCO dataset from latent representations created by the encoder network. We plan to input still images from the COCO dataset into the encoder network with the same masking strategy which the MAGVIT authors use to seed their model for frame prediction tasks.

Your idea: Why does your idea make sense intuitively?
Video generation models have the undue capability of not only being able to synthesize coherent video sequences but also understand the underlying dynamics of the scene and semantics present in the data, such as segmentation, depth perception etc. Furthermore, one of the remarkable aspects of video generation models is its potential for temporal understanding of moving objects in the image. This implicit understanding can benefit tasks such as image segmentation which we intend to explore in this project.

Your idea: How does it relate to prior work in the area?
The idea of using next token prediction to solve downstream tasks has been greatly popularized in the language domain with the advent of LLM's. This idea has carried on to image/video data with Vision Transformers etc., via tokenization and attention, akin to how LLMs process text. For instance, Vision Transformers have demonstrated success in image classification tasks. Inspired by these developments, we aim to explore the applicability of MAGVIT for creating segmentation masks on images.

Your plan for experiments. What all evaluations are you aiming for?
We plan to break up our training and experiments into two components: image encoding leveraging MAGVIT, and generation of segmentation masks from encoded images. Since the authors of MAGVIT do not release pretrained model weights, we will need to train our own MAGVIT model from scratch. We envision that this section entails substantial technical risk, and thus plan to separate it out from the training of our task-specific head. Specifically, we plan to use publicly available pytorch code along with publicly available video datasets (such as Kinetics-600 and SomethingSomething V2) in order to train the 3D-VQVAE model described in the MAGVIT paper. We will

Baselines. What all baselines are you thinking of? (may not be applicable everywhere)
We plan to use the COCO dataset to assess our model's capability of segmentation. With respect to metrics we shall use Intersection-over-Union and mAP.

---

[1] https://arxiv.org/pdf/2212.05199.pdf
[2] https://arxiv.org/pdf/2202.04200.pdf
[3] https://arxiv.org/pdf/2204.03638.pdf
[4] https://arxiv.org/pdf/2312.06662.pdf
[5] https://arxiv.org/pdf/2211.07636v2.pdf
[6] https://arxiv.org/pdf/2308.03747v1.pdf