Visual Learning and Recognition Course Project

Team:
Jay Maier
Nishanth Mohankumar

Motivation: Why is it important and relevant? Why should we care?

Over the past five years, large-scale pretraining on unlabeled data has significantly transformed the landscape of language modeling. While strides have been made in the realm of visual understanding through models like CLIP and BLIP, which leverage multi-modal data such as organic image captions, these approaches still contend with issues related to labeler intent and bias, as well as inherent ambiguity in text-image pairs. However, the realm of video generation presents a distinct advantage due to the absence of ambiguity in pixel values within real video data. This suggests that video generation could serve as a well-structured task for generative pretraining of visual modeling systems, implicitly encapsulating information relevant to various downstream vision tasks such as instance segmentation, mask classification, and depth perception.
Moreover, the abundance of video data available online, estimated to constitute approximately 80% of internet traffic, offers a vast resource for training such models. This stands in stark contrast to text data, where even the largest datasets are dwarfed by the sheer volume of video content created daily. Given this landscape, our project aims to explore the potential of open-source video generation models as adaptable encoders for downstream tasks like instance segmentation, particularly focusing on the COCO Dataset. Instance segmentation, pivotal in computer vision applications ranging from autonomous vehicles to medical image analysis, enhances systems' abilities to comprehend visual complexities and glean implicit information from images.

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], Video GPT[13], and WALT[4]. Beyond the realm of video generation, our approach draws inspiration from work that mentions object segmentation to be akin to viewing how objects move temporally, i.e.parts of the same object have negligible relative velocity with each other.

Your idea: What is the idea?

[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

We plan to train the 3D-VQVAE encoder network described in the MAGVIT paper on video sequence data, 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 which greatly assists with the task of object segmentation i.e. the same object moves at a relatively same velocity temporarily through frames. 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?

We plan to leverage prior work in the video generation space as generalized encoders which we can use to infer segmentation masks on still images. In this we focus on the MAGVIT (Masked Generative Video Transformer) system. MAGVIT stands out as the current best performing model (SOTA) for various video generation and analysis tasks. In addition to MAGVIT, we'll also explore other leading video representation models like MaskGIT, TATS, Video GPT13, and WALT. We plan to design a decoder for whichever video encoding model we select that is reminiscent of the U-Net architecture or something similar to segformer in order to leverage recent advances in transformer architectures.

Your results: If it worked, how much did you improve?If it did not work, why did you expect it to work? what are the next steps to try?

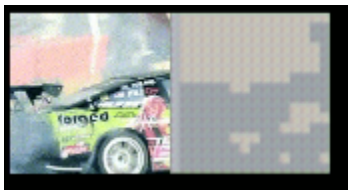
Our project has been streamlined into distinct phases aimed at the development and training of a 3D-VQVAE model and subsequent implementation of a Segmentation head. The outlined steps are as follows:

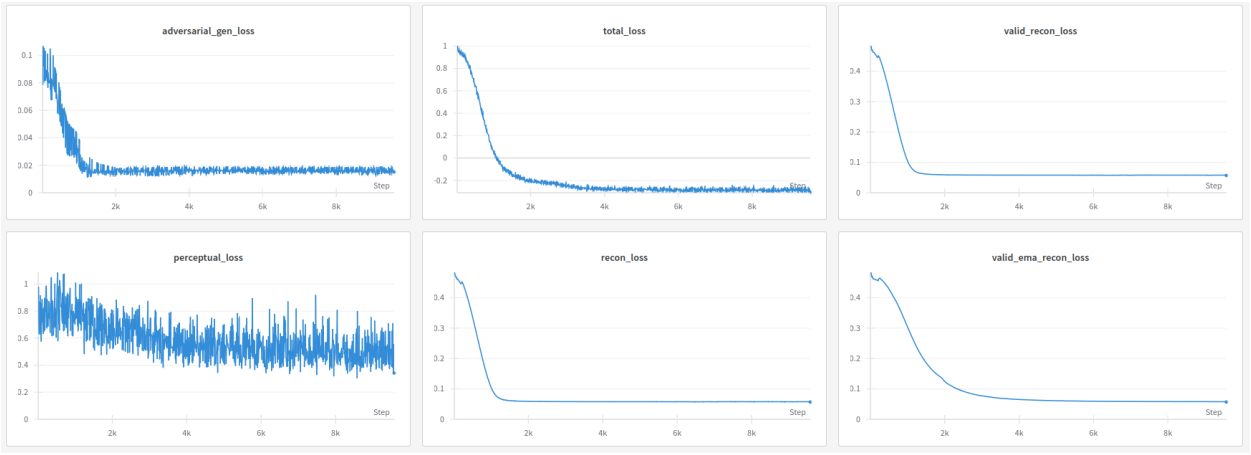
1. Development and Training of the 3D-VQVAE:
    a. Initial training and overfitting on a small dataset to validate model architecture and functionality.
    b. Subsequent training on an expanded dataset to enhance model generalization and performance.
2. Development and Training of the Segmentation Head:
    a. Creation and optimization of a Segmentation head architecture suitable for integration with the 3D-VQVAE.
    b. Integration of the encoder component of the 3D-VQVAE with the developed Segmentation head.
    c. Training of the integrated model on still images sourced from the COCO dataset to achieve segmentation objectives.

Presently, our progress has advanced to a midway point in step 1, specifically in part b, where the 3D-VQVAE is undergoing training on a larger and more diverse dataset of videos(GIF's and MP4). This stage marks a critical juncture in the project, poised to lay the foundation for subsequent phases, including the integration and training of the Segmentation head.

Your results: Any negative results? Maybe you figured after trying that it did not make sense.

We had originally intended to focus our effort for the project on the problem of transferring latent representations from a video generation model to a still-image segmentation model, but unfortunately the video generation model which we had originally chosen as our encoder did not have publicly available model weights (MAGVITv2). Because of this, we have mostly focused on training a video generation model from scratch, and this has proven to be difficult. We are trying to replicate the training process described in the official MAGVIT paper, but are having difficulty recreating their experiments due to computational resources. Because our goal for the project is to investigate the transfer potential of video models to segmentation tasks, we may look into different video generation models to see if we can find something with open source code as well as weights. See below a screenshot of a generated gif using our 3d-VQVAE trained for 10,000 steps. This image, along with our loss curves (also below) suggest that we have not been successful in training a network that is able to accomplish our goal of video generation.





Summary

Our original goal was to use a video generation model that was pre-trained on a large scale video dataset as a general image encoder, and then use the encodings from that video generation model as a latent representation to infer class conditional segments. We have struggled because the video generation model that we originally chose did not have publically available weights, and so we have been attempting to train our own model ourselves (which has proven difficult). Our original goal for the project was to focus on designing a custom segmentation head to attach tonto an off the shelf video generation model, so we may look into switching our encoder architecture to take advantage of pre-trained video generation model weights.

Future directions

In the next stages of our project, we plan to refine the MAGVIT video tokenizer's performance through fine-tuning and enhancing training efficiency. Exploring techniques like mixed-precision training, gradient accumulation, and various learning rate schedulers that will further enhance training efficiency and generalize model performance.

Subsequently, we will proceed with the development and training of a segmentation head, employing Transformer or Convolutional network (CNN) architectures that are compatible with the latent space of the MAGVIT video tokenizer. To validate the functionality of the segmentation head, we plan to utilize a subset of still images from the Common Objects in Context (CoCo) dataset to train and test the model. Finally, we will integrate the encoder component of the 3D-VQVAE (Vector Quantized Variational Autoencoder) with the segmentation head, leveraging its capabilities to extract high-level features from video data and enhance segmentation performance. Through these strategic steps, we aim to advance our project's capabilities and achieve superior results in video understanding and segmentation tasks.