NYU

Future Frame Prediction & Segmentation

Final Project for NYU Deep Learning - Spring 2024

Prithviraj Murthy
pkm5789

Satya Chillale
sc9960

Karan Kumar G
kk5409

Problem Statement

Given the first 11 frames of a video, predict the semantic segmentation of the 22nd frame. The dataset consists 20000 videos split into train, validation, unlabeled and hidden dataset.

Approach

- Diffusion based future frame prediction model to generate future 11 frames for each video using the sliding window of past 11 frames.
- UNET model to perform segmentation and generate mask for the 22nd frame.

NYU

MCVD

- Score-based diffusion model, which work as a stochastic process adding noise to image and then reversing that process to go from noise to image generation.
- The network conditionally masks block of past frames with a probability of 0.5 using binary mask, which enable the network to learn implicit models of spatio-temporal dynamics to generate future block of frame.
- The model is a UNet architecture with 2D convolutions and multihead self-attention with positional encoding.

MCVD

- The model implements a sliding window block-wise autoregressive conditioning diffusion procedure to allow coherent long-term generation. In our case, the sliding window was of length 12 frames generating one frame at a time. We used DDPM method in our sampling.
- The model was trained on 13000 unlabelled videos i.e the given 11 frames for all videos, the model was trained in a manner where it randomly and independently mask all the past frames.

UNET

- Model Structure:
  - Sequentially reduces image dimensions while increasing feature depth through layers of convolutions and max pooling. We used down-conv upto 1024 for the given 160x240 images.
  - Sequentially increases image dimensions and refines features through up-convolutions and concatenations with corresponding down-convolution output.
  - Output layer generates segmentation maps from refined features using a 1x1 convolution to 49 classes.
- The segmentation model was trained on training videos and validated on validation videos for which masks were provided.

Training & Inference

- UNet model Training was performed with following configuration:
  - Model size: 31M parameters
  - Learning rate: 0.00005
  - Batch size: 8
  - Optimizer: RMSProp
  - Epochs: 10
- Diffusion model training was performed with following configuration:
  - Model size: 110M parameters
  - Image size: 128*128
  - Steps: 105000
  - Batch size: 24
  - Generator and attention head channels: 128
  - Spade dim: 256
  - sampling steps: 1000
  - Optimizer:Adam
- Model was trained for 18 hours on a A100 machine.
- Total of 5 A100s for 20 hours each used for video prediction inference on 5000 hidden videos.

Results

- Achieved >96% accuracy on the UNet Segmentation Model on Validation Set
- Future frame prediction with 100 sampling we achieved a Jaccard score of ~0.12 and with 500 sampling score of ~0.2
- Future frame prediction using ⅔ of model size ~0.06
- Future frame prediction and applying the segmentation mask on it, we achieved a ~0.36 Jaccard score on the subset of validation dataset.

Alternative Approaches

- SIMVP, MagVIT and MaskVIT future frame generation.
- Using the diffusion model on already generated segmentation masks instead of actual images to remove the complexities involved in diffusion process

References

- Mcvd-pytorch
- Pytorch-Unet
- MagVit
- MaskVit