Video Signal Processing Assignment

CHEN SHANG

October 2022

1 Data

There is a dataset in kaggle cosisting of 3,020 photos of Donald Trump. And we can find raw data from search enginees. There are 480,669 images from gettyimages.com. We assume that we can get them easily with web crawler.

And data augmentation should not be negelected. We can use Mixup, Cutmix, label smoothing and many other ways to have more data.

The most important thing is that we not only need data of Trump, but also need many categories of text and image pairs. We decide to use diffusion models with contrastive learning, so if we want to learn from scratch, we will need other types of data to support our self-supervise learning.

2 Network architecture

We need to carefully control the length of the assignment, so there are no pictures here.

Our whole network is also a encoder-decoder architecture. Encoder produces image embedding from given text. Decoder uses image embedding to generate images. In detail, we use diffusion models as our decoder [1].

In the first part of our encoder, we use CLIP to extract image features which proves to have great performance in many aspects. And then we train a priori encoder to transfer embedding from text to another type of embedding which is more like embedding from image and more acceptable for image decoder.

In decoder, we use diffusion models which learn from noise and add noise into images as its forward process and generate images from noise as its reverse process. We train the diffusion model to use its reverse process to generate images. In detail, we use UNet with Transformer to implement the network, which adds Transformer block with attention operation in the bottleneck part between encoder and decoder compared with traditional UNet.

3 Any other resources we need

If we need to train a general text to image model from scratch, we need huge amount of GPUs. It is said that Stable Diffusion v1 needs 150000 hours of training with A100.

After all, the more GPUs we have, the better performance we get.

4 Our design and thoughts

For generative task, specially generation of images, the most frequently used models are different types of GANs in the last decade. However, since 2021, diffusion models have got more attention [2] and show many advantages compared with GAN.

Diffusion models always need more data and huge computing resources. Luckily we do not need to have a detailed result with real image output, so we choose to use diffusion models, otherwise we may use GAN. In our task, another advantage of diffusion models is that they can generate images of diversity. Unlike GAN, e.g. StyleGAN, diffusion model can generate images of Donald Trump with many things which do not exist in real world.

Another reason is that it is not easy to generate images of Trump with StyleGAN. Though GAN can directly utilize elements in latent space to change image style while diffusion model cannot, it is hard to say Donald Trump belongs to what type of style. StyleGAN can transfer images of people from a race to the other beacuse they have different styles, but it is hard for network to distinguish facial features of Donald Trump from other features as a style.

Like other generation models, diffusion models also make use of latent variables, but latent variables here have less restriction compared with GAN, VAE and flow-based models. Recent popular work of OpenAI including diffusion model, e.g. CLIP, all show this trend that models do not need much priori knowledge to get great performance.

We would like to fine-tune on a large pre-trained model which can use natural language labels [3] to generate image required. This may get better performance compared with training from scratch. Though we may have millions of images as our dataset, it is still much smaller than the pre-trained dataset which those open-source huge models use.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4