Generative Adversarial Network for Videos and Saliency Map

Bat-Orgil Batsaikhan and Catherine Qi Zhao

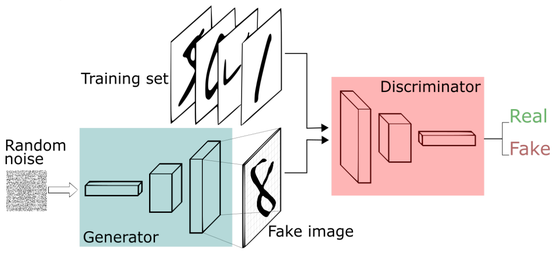
University of Minnesota – Computer Science and Engineering

Introduction

- Generative Adversarial Networks (GANs) are deep neural net architectures introduced by Ian Goodfellow and other researchers at the University of Montreal in 2014 [1].
- GANs achieve incredible results in many important tasks in Computer Vision, such as object detection, 3D modeling, videos, image captioning and so on.
- This project presents implementations of the **VideoGAN** [2], a generative model for videos and **SalGAN** [3], a saliency prediction model. We also provide observations related to the experiments.

Review: Generative Adversarial Network

- Generative models model the data distribution of the individual classes.
- Generator network (G) generates new data instances from a random noise (i.e, 100-dimensional vector from Gaussian distribution).
- Discriminator network (D) outputs the probability of the data being sampled from the distribution.



- We train these networks in a min-max game where G tries to fool D while simultaneously D seeks to detect which examples are real or fake.
- In the perfect equilibrium, the generator would capture the data distribution. As a result, the discriminator would be unsure of whether its inputs are real or not.

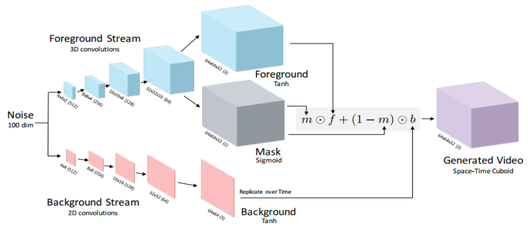
$$\min_{w_G} \max_{w_D} \mathbb{E}_{x \sim p_x(x)} [\log D(x; w_D)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z; w_G); w_D))]$$

- Stochastic Gradient Descent is used to alternate between any differentiable G and differentiable D with respect to parameters ω_G and ω_D to minimize losses.
- The first term of the loss function roughly corresponds to the probability of the data from the sample being real, while the second term is probability of the output of the generator being fake.

VideoGAN

- The implementation is based on the paper *Generating Videos with Scene Dynamics* [2].
- Given a static image, we would like to generate a video of possible consequent frames.
- We used the preprocessed Flickr videos in the Golf category from the original paper [2]. The Golf dataset has 600,000 videos of size 64x64 with 32 frames each (over 1 second)

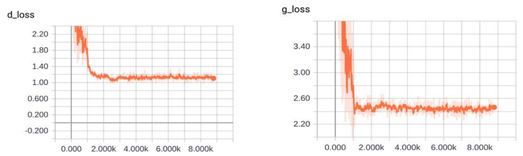
Architecture



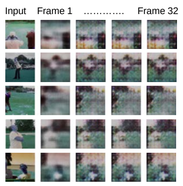
- In the generator network, a five-layer convolutional network is attached to encode the image into the latent space. After each convolution, batch normalization followed by the ReLU is applied.
- The discriminator is a five-layer spatiotemporal convolutional network with 4x4x4 kernels.

Training

- The training typically took several days on a GeForce GTX 1080 Ti GPU. The cross entropy loss, Adam optimizer with a fixed learning rate of 0.0002 and momentum term of 0.5, and a batch size of 64 is used. The following plot for the loss vs iteration indicates the convergence of both the generator and the discriminator.



Results

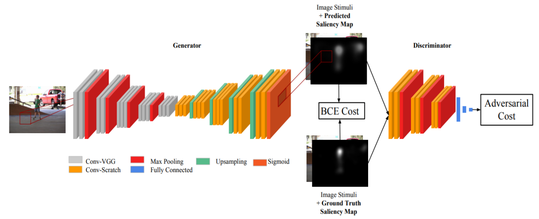


- Generated videos lack resolution, which indicates that the model may be underfitting.
- However, the model seems to capture the main object in the image as blobs.

SalGAN

- The implementation is based on the paper *SalGAN: Visual Saliency Prediction with Generative Adversarial Networks [3].*
- The saliency map refers to a location in an image that attracts the attention of a human.
- The goal is to predict the saliency map given an input image.
- We used SALICON dataset, which offers a large set of saliency annotations [4]. In the preprocessing, images are resized to 256x192.

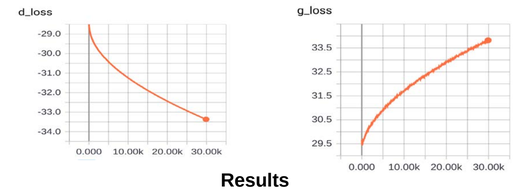
Architecture



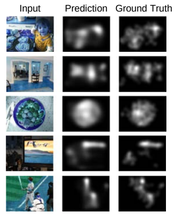
- The generator takes the input image of size 256x192x3 and applies VGG-16 based convolutional network to predict the saliency map.
- The discriminator concatenates a saliency map with its input image and applies a 5-layer convolutional network with 3x3 kernels, ReLU, and pooling interspersed. It ends with three fully connected layers.

Training

- The training took several hours on a GPU. We used AdaGrad optimizer with an initial learning rate of 0.0003, and a batch size of 64. The discriminator seems to dominate the training according to the training loss plot below. This may be due to not training the generator enough and balancing with the discriminator.



Results



- The generated saliency maps are slightly brighter in different areas compared to the expected map.
- Quantitative scores of the model for some of the common criteria:
  - CC score = 0.80
  - AUC_Borji score = 0.73
  - NSS score = 1.98

Observations

- 3D ConvNets used in VideoGAN is more suitable for spatiotemporal feature learning compared to 2D ConvNets in SalGAN. In 2D ConvNet, the filters start to learn-low level features. **In 3D ConvNet, it selectively attends to both motion (spatiotemporal) and 2D features.**
- Normalization is usually used in the preprocessing to deal with the different scales of the input features. **More specifically, batch normalization makes the input to each layer have zero mean and unit variance, which is a regularization method to reduce overfitting.**
- **Visualizing the intermediate outputs, such as masks, foreground, and background in VideoGAN was helpful to understand the problems with the results**. This advice may apply to any other GANs.
- **In SalGAN, even though the total loss for the generator did not converge, BCE part of the loss will helps to make accurate enough predictions**.
- **There are creative ways to enrich the loss functions**. VideoGAN model adds L1 loss between the first frame of the generated video and the input image. SalGAN model adds a content loss, which is another measure of how good is the output of the generator given the input image saliency map.

Conclusion

This project has the following contributions:

- It presents first PyTorch implementations of VideoGAN and SalGAN models. The research community can make use of the open source implementations available on Github in the following repositories:
  - VideoGAN: *https://github.com/batsa003/videogan/*
  - SalGAN: *https://github.com/batsa003/salgan/*
- It provides experimental results and observations for training GANs.

References

[1]. Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[2]. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." Advances In Neural Information Processing Systems. 2016.

[3]. Pan, Junting, et al. "Salgan: Visual saliency prediction with generative adversarial networks." arXiv preprint arXiv:1701.01081 (2017).

[4]. Jiang, Ming, et al. "Salicon: Saliency in context." Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.

Acknowledgements

I would like to thank professor Catherine Zhao for her guidance and Ming Jiang for his technical advice on the project. This work was founded by the University of Minnesota Undergraduate Research Opportunity Program (UROP).