Conditional Generative Adversarial Models for Food Images

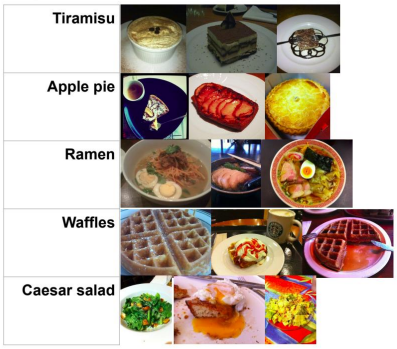
Eric Gene Wu, wueric@stanford.edu; Sudip Guha, sudipg@stanford.edu

Introduction

Conditional generative adversarial networks (GANs) for generating images are an active area of research. In particular, conditional generative models of food are an interesting problem because they generally struggle to reproduce the structure of food and instead tend to produce images containing swirls and blobs. Here we use several different model architectures to generate conditional images of food using the Food-101 dataset [1].

Food-101 Dataset

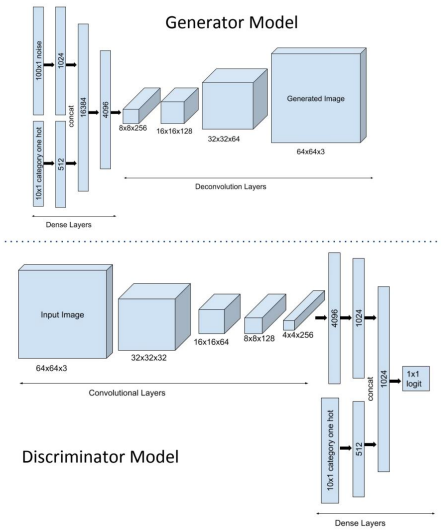
We used the Food-101 dataset provided by ETH Zurich [1]. The dataset contains 101000 images of food split evenly into 1000 classes. The dataset is not cleaned, and so contains noise and mislabeled images. There is substantial variability in the content, lighting, and background of images even within the same class.



References

1. https://www.vision.ee.ethz.ch/datasets_extra/food-101/
2. Arjovsky, M. *et al.* Wasserstein GAN. https://arxiv.org/abs/1701.07875
3. Gulrajani, I. *et al.* Improved Training of Wasserstein GANs. https://arxiv.org/abs/1704.00028
4. Odena, A. *et al.* Conditional Image Synthesis with Auxiliary Classifier GANs. https://arxiv.org/abs/1610.09585
5. Salimans, T. *et al.* Improved Techniques for Training GANs. https://arxiv.org/abs/1606.03498

GAN Network Architecture

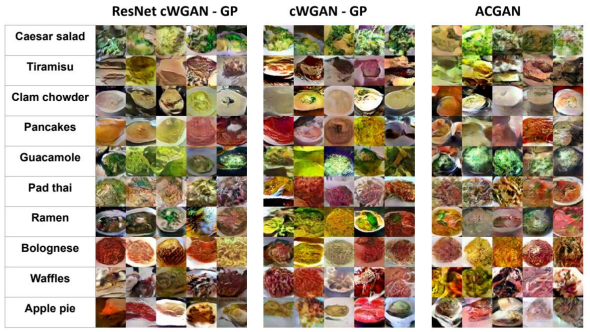


- Residual blocks instead of regular convolutional blocks help produce sharper images.
- Conditioning using labels are concatenated to the learned encodings in the discriminator and to the noise in the generator.
- We found that the Wasserstein loss [2] with gradient penalty [3] to be the best option as it avoids significant mode collapse and allows for a higher learning rate.

$$L = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$

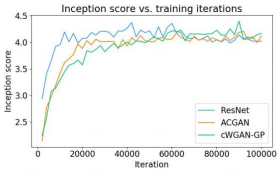
- We also experimented with ACGAN model architectures [4] to use the conditioning, however, we found WGAN-GP to generally work better.

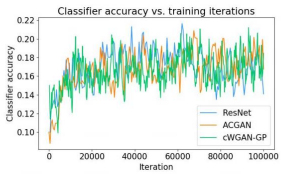
Results and Discussion



Quality of generated conditional images is evaluated using the following methods:

- Inception score [5], to assess quality and diversity of the generated images
- Accuracy of an independent classifier (we trained a separate 10-class classifier using the same dataset), to assess quality of conditioning.





- Inception scores for the ResNet (cWGAN-GP) model were generally better than for the ACGAN and the non residual version of the cWGAN-GP model, corresponding to sharper, more well defined images.
- Classifier accuracy hovered around 20%, which is better than random chance. Classifier accuracy was generally low due in part to the poor quality of the generated images, and due to overfitting in the classifier.