Semantic Pyramid for Image Generation

Brief Summary

Inspired by the Self-Attention Generative Adversarial Networks and the utilisation of the semantic information obtain from classification models, this paper proposes a GAN based model which utilizes the semantic information from the deep features of an image learned by a pretrained classification model, to generate diverse image samples with features that match at each semantic level of the classification model. To work with the feature space of an image, the features responses from different layers of the classification model are extracted and manipulated. These modified features are then inverted to a real image; However, there isn't a one-to-one mapping between features and an image and therefore inverting the modified features to an image is a difficult task. Previous methods use a regularization prior on the reconstructed image but lead to blurry and unrealistic images. However, GAN based models are capable of generating realistic distributions of natural images and therefore perform better than the previous methods.

The proposed model uses the VGG-16 classifier, pretrained on the Places365 dataset to obtained the features at different semantic levels. The GAN model is modified to have a mirror structure with the classification model and takes a random noise vector *z*, a set of masks *M* as well as the deep features *F* as inputs wherein the masks control the extent to which each of the deep feature maps are used. The generator is trained against a class-conditional discriminator and includes an adversarial loss term to account for it. The model utilizes a *semantic reconstruction loss* term to encourage the output to preserve the features from which the output was generated. It also includes a diversity loss term to ensure that the generated results vary significantly from one another.

The authors use the *FID score* and *Real-Fake* user study metrics to evaluate the results while also testing the model for varied tasks such as *Repainting, Semantic Image Composition* and *Relabelling.*

Main Contribution of the paper

One of the most important differentiators of the proposed model to previous work is the utilization of a novel GAN based architecture to leverage features from different semantic levels to generate realistic distributions of images.

Eariler work either utilizes a regularization prior on the reconstructed image (*giving blurry and unrealistic results)* or for the work that uses GAN based models, the image generation is only conditioned on the object's class information. These methods lead to a single image unlike the proposed model which generates a distribution of images and therefore the new features match at each semantic level of the classification model.

List of strengths and weaknesses

Strengths	Weaknesses
The same model can be used without significant manipulation to complete a variety of tasks such as *semantic image composition* and *image repainting*	The proposed model is specific to utilizing images related to nature and cannot be adopted for images that include people in them
The model generates a set of possible images that match the features instead of a single image	When testing the model on applications such as *repainting* or *semantic image composition*, only the qualitative results have been provided
The model is robust due to the inclusion of the adversarial loss, the semantic reconstruction loss and the diversity loss	

Explanation of strengths and weaknesses

A major advantage of using *masks* in the model is that they allow user controllability. To adopt to different tasks such as *generation from unnatural reference images*, the model must allow features from different semantic levels or from different images to be combined and setting the values of the masks accordingly helps the user to do so.

Generating a distribution of possible images that match the features of the image from the classifier certainly increase the probability of finding images that semantically match the deep features generated from the classifier as opposed to generating a single image.

As a property of GANs, humans are not synthesized well by GANs and therefore this model is limited in its application on a certain category of images (*nature-related*) images. Also, when discussing the application of the model on varied tasks in *Section 4.2*, the authors only provide the qualitative results of the model's performance on each of the tasks and therefore make it difficult to compare the results with previous state-of-the-art models in that task space.

Comment on the experiments

To perform quantitative analysis, the authors made use of the *Frechet Inception Distance* (FID) metric along with a *Real-Fake* user study (Amazon Mechanical Turk). The *FID* metric compares the distribution of the generated images with the real images that were used to test the network and are typically used to test the images generated by GANs. The authors used 6000 random samples from the *Places365* dataset and extracted their features from the VGG-16 pretrained classifier. They then generated random image samples from each semantic level to evaluate the *FID* scores. For deeper features (example: from CONV 4), they found that the features of the generated images deviated more from those of the original image whereas for the features from CONV 1, the features of the generated images aligned almost perfectly with those of the original images, leading to a low *FID* score.

For the *Amazon Mechanical Turk* user study, the authors conducted a

- Paired Test: the generated image is presented against the reference image
- Unpaired Test: the generated image is presented against a real-unrelated image

These tests were conducted with 100 raters and 75 images, randomly sampled from the *Places365* dataset were shown to the raters for 1 sec. A confusion rate (% of turkers fooled) was used as a metric and the tests were conducted for images generated from semantic features at each level of the model. It was found that for images generated from the earlier layers fooled more turkers than the one generated from features of the deeper layers.

How could the work be extended?

As discussed in the *Weaknesses* of the paper, the authors must conduct quantitative experiments to test their results on varied applications against the state-of-the-art models for those applications.

The GAN model could also be modified and further developed for it to generate images in which humans are included.