Review 6 - 1

Generating Visual Explanations

Summary

Current two existing main deep visual models are image classification models and image caption methods. Deep classification methods have had tremendous success in visual recognition, but they don't provide enough evidence to show why a specific images are supposed to be classified into specific class. On the other side, a standard captioning system might provide a description, but it doesn't mention which features in this sentence are discriminative. To give a better explanation, this paper presents a new method that not only provides class labels, but also explains why the predicted label is discriminating properties. For this goal, this paper proposes a novel loss function. This loss function consists of two parts. The relevance loss is learnt to generate a good description based on the image and discriminative loss is to focus sentence generation on discriminative visual properties of an image. The learning process of the relevance loss is similar with traditional image caption models. One modification made in this paper it they include the object category as an additional given input. More impressively, this paper trains the discriminative loss by a reinforcement method. First, they sample a sentence and then input the sample sentence into the discriminative loos function. By gradient descent, the model can maximize a reward function to determine if the sampled sentences are discriminative.

Main Contributions:

1. Includes the object category as an additional input to get a more discriminative description.
2. Incorporate a reinforcement learning based discriminative loss to make the final explanation include only discriminative features.
3. Propose a method by employing reinforcement learning to optimize the discriminative loss over sampled sentences.

Pros:

1. Provide an explainable description for images.
2. Use pre-trained models.

Cons:

1. The sentence-based explanations might not be structured enough and hard to use for other model.
2. Can use a smarter sampling method.

Review 6 - 2

Learning What and Where to Draw

Summary

Generative Adversarial Networks (GANs) have recently achieved a great success in synthesizing real-world images. An obvious limitation of existing models is they are only based on global features such as class label or caption and they don't provide control over pose or object location. To address this issue, this paper proposes a model learning to perform location and content-controllable image synthesis. This paper also presents two ways to encode spatial constraints. One way is to use a bounding box and a mask to detect and keep the features, and the other is synthesizing images based on key points. For the bounding box method, the model first generate a caption from a pre-trained model and convert the feature vector of the sentence into a smaller size concatenated with a noise vector. Then the model uses a global generator to generate general image. At the same time, it also employs a local deconvolutional pathway and a mask to specifically generate target objects. As for the key-point-conditional model, a bunch of pre-tagged keypoints are used to focus on generation of key point of the objects. Besides, by using generative methods, this model can generate missing keypoints given observed keypoints.

Main Contributions:

1. Propose a novel architecture for text- and location-controllable image synthesis, yielding more realistic and higher-resolution image samples
2. Present a text-conditional object part completion model enabling a streamlined user interface for specifying part locations.
3. Try this model for pose-conditional text to human image synthesis.

Pros:

1. During training, the 1024-dimensional text embedding for a given image was taken to be the average of four randomly-sampled caption encodings corresponding to that image.
2. Can generate higher-resolution images.
3. Suggest that providing additional conditioning variables in the form of location constraints is helpful for learning to generate high-resolution images.

Cons:

1. In some cases lack clearly defined parts such as a beak.