Review 4 - 1

Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions

Summary

This paper introduces a Zero-Shot Learning model that learns to predict unseen images classes from texts (encyclopedia articles). The overall goal of this paper is to learn an images classifier from natural language descriptions. To achieve this, this paper need to map text description features to image features. However, dimensions of images and text feature space are usually extremely highly, which could make it difficult to estimate the large number of parameters in map function. Thus this paper introduces a second mapping parameterized by a multi-layer neural network that transforms the visual features to a lower dimensional space, which can reduce the amount of parameters dramatically and speed up the training a lot. In addition, considering the structure of image features, both the low-level convolutional weights and the fully connected weights are predicted from the text feature using a single multi-task neural network with shared layers. Finally the paper shows the model outperforms previous zero-shot methods on the ROC-AUC metric.

Pros:

1. Provide a flexible Zero-Shot model can lean to predict unseen images classes from text descriptions.
2. Improve the state-of-the-art on two main-stream datasets using only raw images and text articles.
3. Map images features to a lower-dimension space.
4. Use both low-level and high-level features of CNN.

Cons:

1. The multi-class recognition performance on the zero-shot class is still lower than some of the attribute-based methods.
2. Doesn't use an LSTM RNN to extract text features

Review 4 - 2

Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Summary

This paper proposes a method of visual question answering combining an internal representation of the content of an image with information extracted from a general knowledge base to answer a broad range of image-based questions. The core idea of this paper is to combine images information with the information representations retrieved from a KB in a RNN. Given an images, a CNN is first applied to produce the attribute-based representation. The internal textual representation is made up of images captions generated based on the images-attributes. The last vector representations in image caption RNN are aggregated and be used as caption representation. The external knowledge is mined from the KB and encoded by Doc2Vec as knowledge representation. Finally the 3 vectors are combined into a single representation and put into VQA LSTM model to generate a possible answer with highest probability. At the time of writing this paper, our system performs the best on two large-scale VQA datasets and produces promising results on the VQA evaluation server.

Pros:

1. The model can answer the question about the images, even when the answers are not covered in the contents of the image.
2. Outperform previous state-of-the-art models on visual question answering.

Cons:

1. Performance highly depends on the knowledge base.
2. The model generates key words and retrieves information from knowledge base before it sees questions.
3. Ignore the structured representation.