Review 3 - 1

Learning Aligned Cross-Modal Representations from Weakly Aligned Data

Summary

People can recognize scenes from different kind of modalities, but it is not a trivial task for computer. Today state of the art computer vision techniques can map the features of images into a high-dimensions space, but the problem is the intermediate representations are not aligned across modalities. To tackle this problem, this paper proposes two complementary models and combines them. One is applying fine-tuning on modality, and the other is to encourage intermediate layers to have similar statistics across modalities.

In Section 1, this paper presents basic concepts like why computer vision models don't have the cross-modal capability and what the difference between strong and weak alignments. Then it states the goal of this paper is to learn cross-modal representations for scenes that have strong alignment using only data with weak alignment. In Section 2, the paper presents several related research area. Section 3 describes the new data set the authors built for this task, which covers five different modalities. Section 4 describes the details of the model. Firstly they extend single-modality classification networks in order to handle multiple modalities. They build one network for each modality and let the high-level layers to be shared across all modalities. This structure can be used to learn cross-modalities representations from different image sources. Then to address the problem the represents in later layer are mostly for one modality, the paper presents two methods. For the modality tuning, they replace the earlier layers of the network instead of replacing the last layers of the network which are done by regular fine-tuning. It's good because the higher level features are trained for only one modality. The other method is statistical regularization. The basic idea is to encourage activations in the intermediate hidden layers to have similar statistics across modalities. The model learns the distribution over the hidden activations by using activations in the hidden layers of CNN. Finally, the model combines the two methods by control the activation of shared layers and weights in different stages. Section 5 is about experiments and results. They compute the distances between extracted features and the representations of sampled images to rank the sources from different modalities.

Main Contribution

1. Propose a model that can learn aligned representations from cross-modal
2. Can learn a representation for scenes that has strong alignment using only data with weak alignment.
3. Show they can reconstruct natural images from other modalities using the features in the aligned representation as a qualitative measure of which semantics are preserved in our cross-modal representation.
4. Can generate natural images from different modal information source.

Pros and Cons

Pros: 1. Provide a structured neural network to learn aligned cross-modal representations. 2. 1. Can learn a representation for scenes that has strong alignment using only data with weak alignment. 3. Prove the interpretability of the model.

Cons: This paper doesn't get an obvious performance improvement on the task of cross-modal retrieval of semantically-related content.

Review 3 - 2

Summary

This paper provides model learning not only from the plain text, but also from the hypernymy and visual-semantic hierarchy structure. The basic idea is to exploit the partial order structure of the visual-semantic hierarchy by learning a mapping which is not distance-preserving but order-preserving between the visual-semantic hierarchy and a partial order over the embedding space. To achieve, they first abstract all hypernym prediction, caption-image retrieval and textual entailment problem into the problem of partial order completion. Then to model the semantic hierarchy, they come up with a method of choosing y the reversed product order on R^N_+ to make sure it is rich enough to embed all pair relations. After learning order-embeddings from hypernymy, they combine it with an encoder-decoder structure to conduct image caption. Finally they compare the performance of other state of the art models and their model. It's turned out there is outperform than others.

Main Contribution

1. Abstract different tasks into a same problem.
2. Provide a general method for learning ordered representations in different tasks.

Pros and Cons

Pros:

1. Provide new state of the art in image caption retrieval task.
2. Can caption images better even there are many details in different levels.

Cons: