Title

Learning Deep Representations of Fine-Grained Visual Descriptions

1st Author

Scott Reed

Submission Date

17 May 2016

Publisher

-

You can download this paper from **here**

Abstract:

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manually encoded vectors describing shared characteristics among categories. Despite good performance, attributes have limitations: (1) finer-grained recognition requires commensurately more attributes, and (2) attributes do not provide a natural language interface. We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images. Natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. Our model achieves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification on the Caltech-UCSD Birds 200-2011 dataset.

My View:

The paper introduces a new method for zero-shot visual recognition. Instead of using attributes, the authors train neural language models from scratch using only text. This approach aligns with fine-grained image content and uses natural language to describe image details, leading to better performance in zero-shot classification and image retrieval compared to attribute-based methods. The main idea is to leverage neural language models to improve fine-grained visual recognition using natural language descriptions.

Previous Idea:

- Deep networks improved visual recognition, with transferable features.
- Multi-modal learning combines different data modalities effectively.
- Models generate text tags from images using LSTMs or character networks.
- CNN-RNN architecture extended for character-level visual embeddings.
- Prior work improved label embeddings for image classification.
- Assumes substantial text data for training high-capacity models.
- Expands on character-level language models and fine-grained zero-shot learning.
- Demonstrates text-based embeddings outperform attributes for zero-shot recognition.

Previous Idea weakness:

Weaknesses of previous ideas include: limited labeled data for fine-grained classification, overfitting risk in captioning models, potential lack of context in image descriptions, text data dependency, complexity of character-level embeddings, reliance on attributes for zero-shot learning, lack of a natural language interface, potential generalization limitations of image-text features, unexplored zero-shot text-based retrieval, dependence on specific modalities in multi-modal models, and complexity of building joint embeddings from scratch.

Paper solution:

The paper proposes training neural language models from scratch for better zero-shot fine-grained visual recognition. By aligning with image content, these models generate accurate natural language descriptions, outperforming attribute-based methods. Using a high-quality dataset, the authors show that their approach achieves state-of-the-art recognition accuracy on the Caltech-UCSD Birds 200-2011 dataset. This strategy also improves performance across different training data sizes, enabling effective retrieval systems based on language.