Motivation

- The goal of this paper is to propose an alternative solution to fully supervised training of CNNs by leveraging the correlation between images and text found in illustrated articles.

- Our main motivation is to explore how strongly are language semantics as a supervisory signal to learn visual features.

Proposed Approach

- Given an illustrated article we project its textual information into the topic-probability space provided by a topic modeling framework.

- This way the CNN learns to predict the semantic context in which images appear as illustration.

- Then we use this semantic level representation as the supervisory signal for CNN training.

- We train our models on a subset of Wikipedia articles.

- 35,582 unique articles and 100,785 images.

Rich visual features from freely available data

- We train our models on a subset of Wikipedia articles.

- 35,582 unique articles and 100,785 images.

Image Classification and Multimodal Retrieval

- Image Classification is done by using one-vs-all linear SVMs trained on max5, pool5, fc6 and fc7 feature maps.

- Multi-modal retrieval : (1) Image query vs. Text database, (2) Text query vs. Image database on Wikipedia Dataset [2].

Qualitative Results on Multimodal Retrieval

- Top 4 nearest neighbors for a given query image (left-most). Each row makes use of features from different layers: prob, fc7, fc6, pool5 (from top to bottom).

- Top 10 nearest neighbors for a given text query (from left to right: “airplane”, “bird”, and “horse”).

Conclusions

- We can use freely available multi-modal content to train a CNN without human supervision.

- CNNs can learn rich visual features from noisy and unstructured textual annotations.

- Our results are comparable with state of the art self-supervised algorithms for visual feature learning.

References


Code & Models

https://git.io/vSotz