Self-supervised learning of visual features through embedding images into text topic spaces

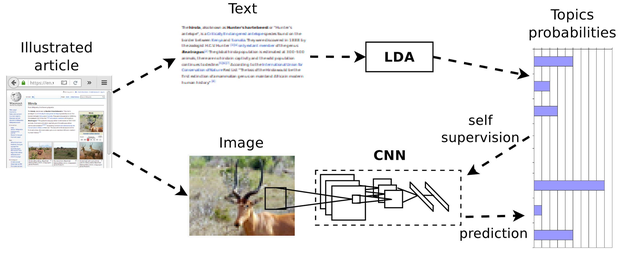
Lluis Gomez*[1], Yash Patel*[2], Marçal Rusiñol[1], Dimosthenis Karatzas[1], & C.V. Jawahar[2]
[1]Computer Vision Center, UAB, Spain [2]CVIT, KCIS, IIIT Hyderabad, India

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

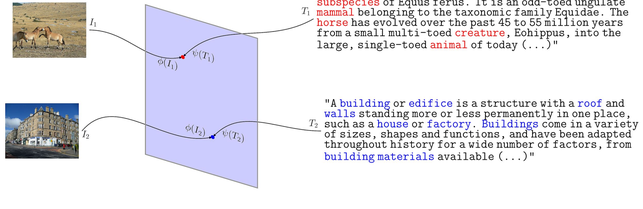
Motivation

- 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 strong 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.
- Then we use this semantic level representation as the supervisory signal for CNN training.
- This way the CNN learns to predict the semantic context in which images appear as illustration.



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.

Top-5 most relevant words and top-5 most relevant images for three of the discovered topics.

Top-5 activations for three units in fc7 layer.

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].

Method	max5	pool5	fc6	fc7
TextTopicNet	-	47.4	48.1	48.5
Sound [3]	39.4	46.7	47.1	47.4
Texton-CNN	28.9	37.5	35.3	32.5
K-means [6]	27.5	34.8	33.9	32.1
Tracking [4]	33.5	42.2	42.4	40.2
Patch pos. [7]	26.8	46.1	-	-
Egomotion [1]	22.7	31.1	-	-
AlexNet [8]	63.6	65.6	69.6	73.6

PASCAL VOC2007 %mAP image classification.

Method	Img2Txt	Txt2Img	Avg.
TextTopicNet	39.6	38.2	38.9
CCA [2]	19.7	17.8	18.8
PLS [11]	30.6	28.0	29.3
SCM [2]	37.1	28.2	32.7
GMMFA [5]	38.7	31.1	34.9
CCA-3V [10]	40.5	36.5	38.5
GMLDA [5]	40.8	36.9	38.9
JFSSL [9]	42.8	39.6	41.2

MAP comparison on Wikipedia dataset with unsupervised (middle) and supervised (bottom) methods.

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

[1] Agrawal et al. "Learning to see by moving." In ICCV, 2015.
[2] Rasiwasia et al. "A new approach to cross-modal multimedia retrieval." ACM-MM, 2010.
[3] Owens et al. "Ambient sound provides supervision for visual learning." In ECCV, 2016.
[4] Wang et al. "Unsupervised learning of visual representations using videos." In CVPR, 2015.
[5] Sharma et al. "Generalized multiview analysis: A discriminative latent space." In CVPR, 2012.
[6] Krahenbühl et al. "Data-dependent initializations of convolutional neural networks." In ICLR, 2015.
[7] Doersch et al. "Unsupervised visual representation learning by context prediction." In ICCV, 2015.
[8] Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." In NIPS, 2012.
[9] Wang et al. "Joint feature selection and subspace learning for cross-modal retrieval." TPAMI, 2016.
[10] Gong et al. "A multi-view embedding space for modeling internet images, tags, and their semantics." IJCV, 2014.
[11] Rosipal et al. "Overview and recent advances in partial least squares." In Subspace, latent structure and feature selection. 2006.

Code & Models

https://git.io/vSotz