Object recognition: Project proposal

Louis Martin and Zaccharie Ramzi

December 4, 2016

Topic E - Joint representations for images and text

Multimodal retrieval: image-to-image search, tag-to-image search, and image-to-tag search.

We chose this project for its interesting combination of visual and textual data.

Technical choices

We will use **git** for code versioning, coworking and to be sure we divide the work equally. Our language of choice is **Python**.

Project description (taken from course project description)

Data: Microsoft COCO dataset.

Automatically producing natural text describing the content of an image is a very hard problem. We will investigate joint representations for images and text suitable for this task. In particular, we will investigate the canonical correlation analysis (CCA) [1], a popular and successful approach for mapping visual and textual features to the same latent space. We will experiment with canonical correlation analysis on several sources of data to find correlations between image features and sentence features.

Work division

We chose to work jointly on step 0 and 1 to be both familiar with the subject, code, and to have a clear overview of the subject in order to define precise common objectives.

Detailed step-by-step instructions (taken from course project description):

0. Design the code architecture.
   We will first create an architecture to divide code into meaningful modules for efficient coworking and code reuse.
   Zaccharie and Louis

1. Implement CCA following [1]. Show it works on toy synthetic data. We will focus only on the standard two-view CCA. The synthetic toy data can be generated as follows:

   - Sample two clouds of points from two different gaussian distributions
   - Given two different mappings ϕ and ψ from $\mathbb{R}^2$ to $\mathbb{R}^2$, we generate two views out of the main view. For example $\phi : (x, y) \mapsto (x + y, x - y)$ and $\psi : (x, y) \mapsto (xy, e^{x+y})$.
   - We will apply CCA on this two view representation and try to retrieve the original view.

   Zaccharie and Louis

2. Extract text word representations from the sentences associated with MS COCO data using [2].
   Louis

3. Extract CNN image representations using [3].
   Zaccharie

4. Apply CCA on the features extracted in 2. and 3.
   Zaccharie and Louis

5. Implement a retrieval pipeline using the computed correlations as in [1].

   Tag-to-image search (T2I) **Zaccharie**

   Image-to-tag search (I2T) **Louis**

6. Show qualitative example results on Microsoft COCO dataset
   Louis

7. Pick 5-10 objects and quantitatively evaluate the results for tag-to-image search using the ground truth object labels provided with MS COCO. Plot a precision recall curve for each object and report average precision (AP).
   Louis

8. Use the object ground truth for MS COCO to quantitatively evaluate the image-to-tag search as described in section 6.7 of [1], i.e. compute the average precision for all tags ranked number 1, number 2, etc. We will do this on a randomly sampled test set.
   Zaccharie

9. Quantitatively compare different CNN features, e.g. [3] with [4] or even [5].
   Zaccharie

References

[1] *Normalized CCA*, http://slazebni.cs.illinois.edu/publications/yunchao_cca13.pdf

[2] *Word2Vec*, http://code.google.com/p/word2vec/

[3] *Overfeat*, http://cilvr.nyu.edu/doku.php?id=software:overfeat:start

[4] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014

[5] M. Cimpoi, S. Maji, A. Vedaldi, Deep Filter Banks for Texture Recognition and Segmentation, CVPR 2015.