RecVis : Class project proposal

Matthieu Kirchmeyer
matthieu.kirchmeyer@mines-paristech.fr

Sylvain Truong
struong@ens-paris-saclay.fr

December 6, 2016

1 Choice of Project

We will be working on:

Topic E - Joint representations for images and text

The goal of this project is to understand the major concepts surrounding automatic image-text translation and to implement algorithms for automatically producing natural text describing the content of an image.

2 Approach and project steps

Data set: The algorithms will be run on the MS COCO dataset using $5 - 10$ objects for evaluation.
We will focus on joint representations of text and images, which map visual data and tags into a same latent space. This latent space embodies the real nature of the action, or the scene, that is pictured by the text of the image. We will follow the following steps which are of increasing difficulty:

- Form a bibliography
- Implement the canonical correlation analysis (CCA) and try it on personally-generated toy examples using [1]
- Apply CCA on a larger scale using the text and image features extracted from MS COCO in the previous steps
- Implement text-to-image and image-to-text retrieval
- Evaluate the performance of this first attempt on MS COCO dataset (precision recall curves, mAP...)
- Analyse the influence of the CNN features on the retrieval's performance
- Implement full-sentence caption generation and compare it to RNN outputs of [2]

3 Workload sharing

We plan to share the workload the following way:

Matthieu

- Implementation of CNN feature extraction and text-image retrieval full pipeline
- Performance evaluation of the first attempt
- Study of sentence generation from tags representation in the latent space

Sylvain

- Implementation of CCA, and text feature extraction
- CNN feature choice sensitivity analysis
- Implementation of full-sentence caption generation and evaluation

4 Short bibliography

[1] Deep Visual-Semantic Alignments for generating images descriptions - Andrej Karpathy, Li Fei-Fei
[2] A multi-view embedding space for modeling internet images, tags and their semantics - Yunchao Gong, Qifa Ke, Michael Isard, Svetlana Lazebnik