European Master in
Language & Communication Technologies (LCT)

Multimodal Fusion for Combining Textual and Visual Information in a Semantic Model

Nam Khanh Tran

Supervisors: Dr. Marco Baroni
Prof. Dr. Manfred Pinkal

July, 2012

Abstract

In recent times, multimodal fusion that combines information from different data streams such as text and images, has attracted the attention of many researchers due to the benefit it provides for various analytical tasks. In the computational semantics community, recent work has extended distributional semantics to integrate features extracted from images as well as text. In this thesis, we propose a flexible and general framework for the integration of text- and image-based features that includes both a latent mixing phase in which text and image features are mapped together onto a lower-dimensional space, and a multimodal similarity estimation phase, where the two channels are linearly combined either at the feature or at the scoring level. The framework is of sufficient generality that, with specific parameter choices, it encompasses combination techniques that have been previously proposed in the literature. In addition, it can automatically pick the best model for a task given development data.

We evaluate our framework through extensive experiments on simulating similarity judgements, and concept categorization. A further experiment to examine the effect of different image-based features combined with a textual model into a multimodal architecture is also presented. It is shown that our system significantly outperforms a state-of-the-art text-based distributional semantic model on the MEN semantic relatedness evaluation dataset and, using the parameter values tuned on MEN, is also the top performer on the independent WordSim evaluation. In a similar manner, we obtain improvements on AP, a widely used concept classification benchmark, with the use of the parameters tuned on another dataset, BATTIG. The good performance of the multimodal models using a different test set from the one used for parameter tuning demonstrates the robustness of our approach. Moreover, we also show evidence that adding better visual features improves the performance of the system.

Keywords: Multimodal fusion, distributional semantics, latent semantic mixing, multimodal similarity estimation.