Muster

Multimodal Processing of **S**patial and **T**emporal **E**xp**r**essions

Start date: April 2016

Chist-ERA meeting (Bucharest, Roumania)
3rd of April 2019

Participants

Belgium - KU Leuven (NLP)
Marie-Francine Moens, Guillem Collell Talleda

Switzerland – ETH Zurich (Vision)
Luc Van Gool, Dengxin Dai, Michal Perdoch

France – Sorbonne Université (Machine Learning)
Patrick Gallinari, Benjamin Piwowarski, Laure Soulier, Eloi Zablocki, Patrick Bordes

Spain – University of the Basque Country (NLP)
Aitor Soroa, Eneko Agirre, Oier Lopez de Lacalle

Key challenges

Ground language in perception (visual inputs) and extract representations of meaning tied to the physical world

Multi-modal representation construction

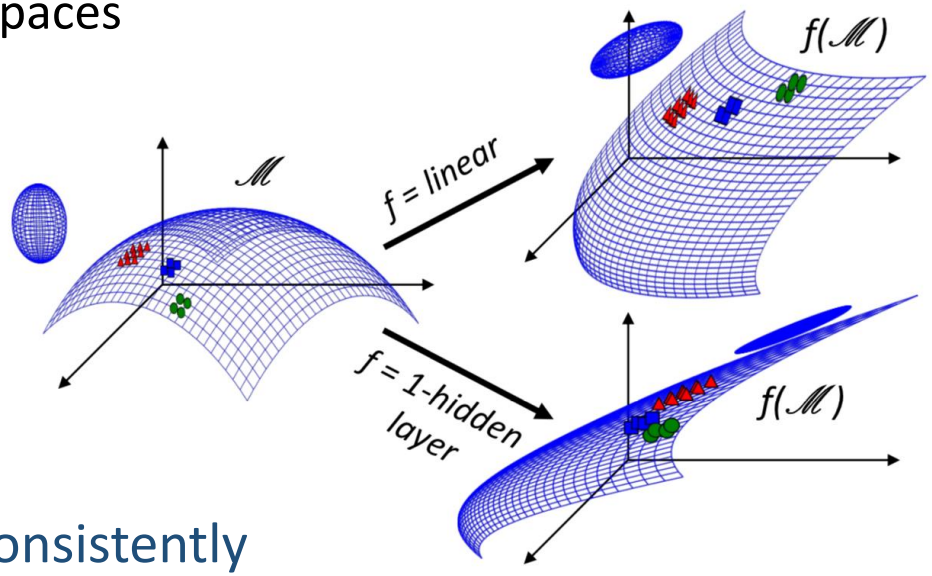
How to automatically create text representations in the form of single-word and multi-word embeddings that ***integrate perceptual knowledge*** *in the representations of objects, actions, their spatial and temporal relations?*

Multi-modal representation integration and usage

How to use the new improved semantic representations (i.e., embeddings) to improve machine understanding of human language?

Achievements – Study of Embeddings

- Does the mapping from text to image spaces work ?
  - Comparison of the neighborhood structure of the two spaces
    - Image → Text and Text → Image
    - Various NN architectures
    - New similarity measure for capturing neighborhood structure

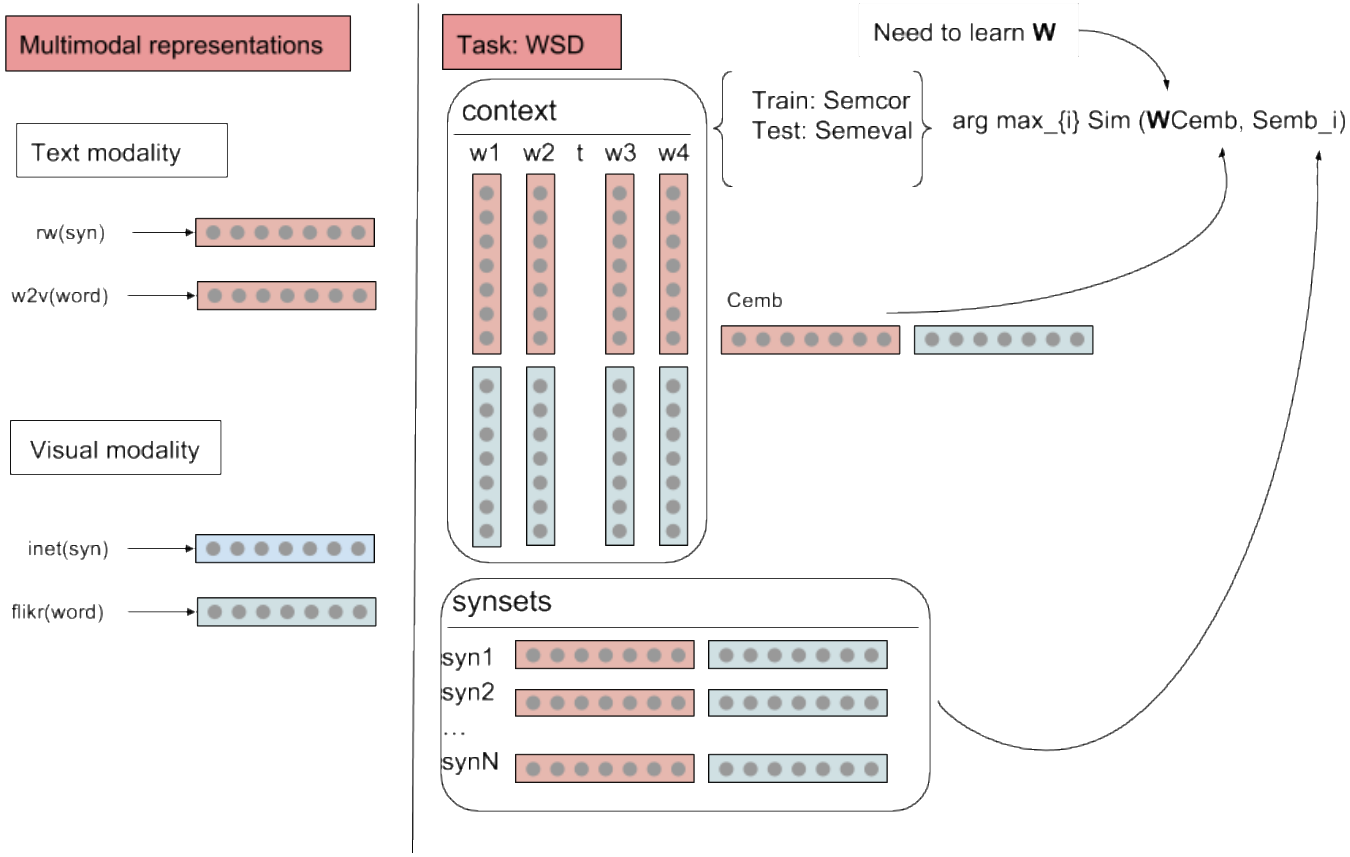


- **Main conclusion**

  The neighborhood structure of the predicted vectors consistently resembles more that of the input vectors than that of the target vectors.

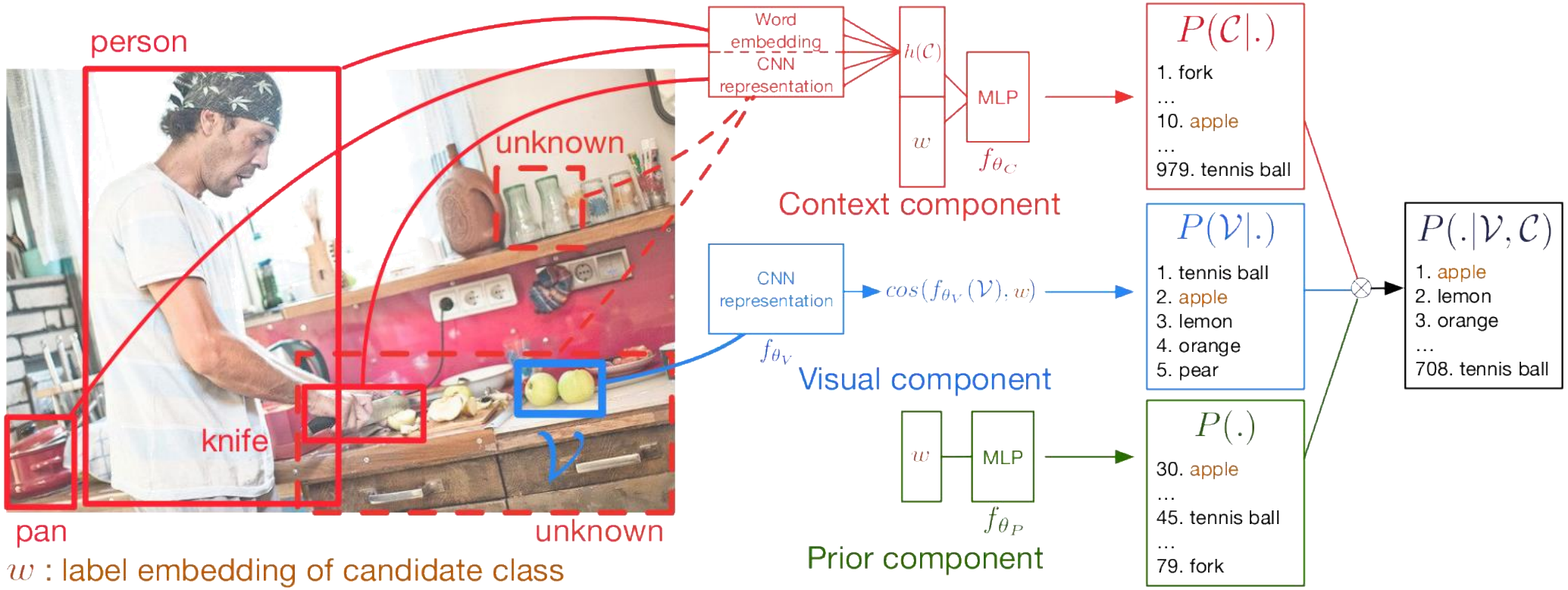
Achievements – Word Sense Disambiguation

- Incorporate visual information into the text representation, and learns correspondences between synset-embedding space and word embeddings
- Visual modality
  - ImageNet for synsets
  - Flickr for words
- The framework is able to leverage multimodal information that is not aligned with the text, and to learn relevant features even for unseen words (transfer learning)



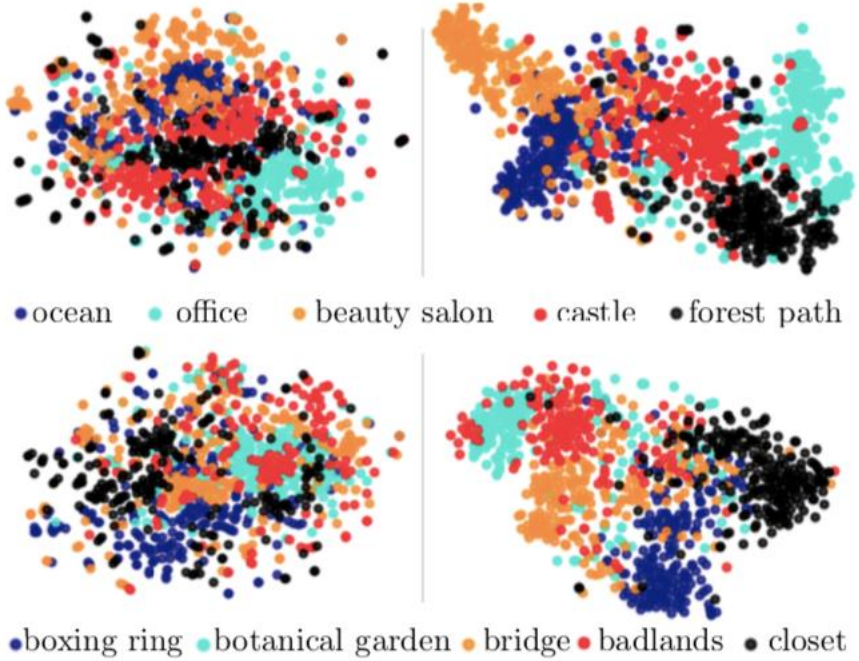
Achievements: Bridging Visual Context with Language

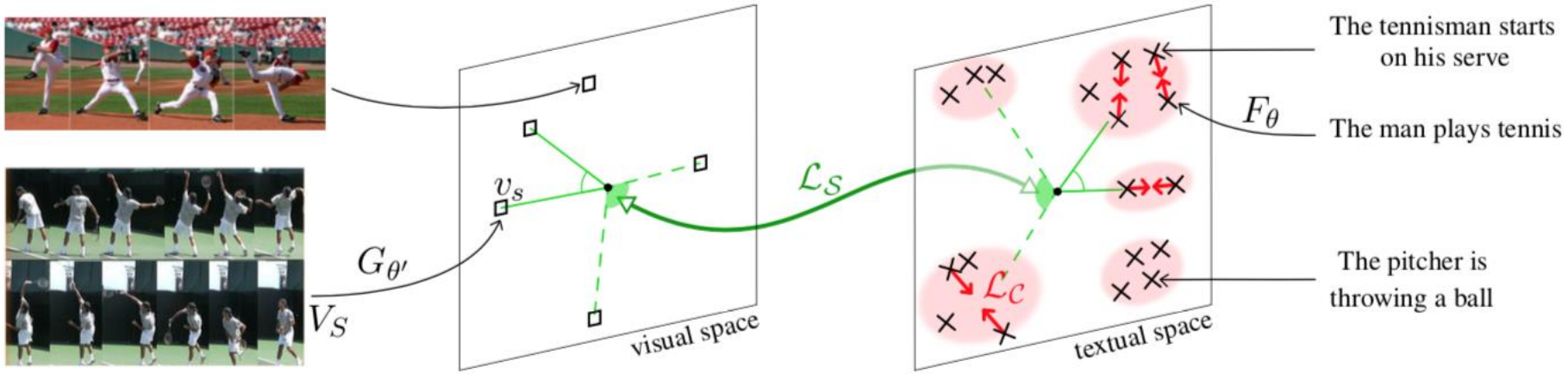
- **Task**: Recognize an **unlabeled** objects given their visual appearance **and** context
- **Idea**: Leverage textual context as a substitute for image context
- Positive results when using visual context (both for seen and unseen objects)



Achievements: Multimodal Sentence Embeddings and Visual Data

- Goal: Learn a grounded **sentence** representation – captures word representations in their context
- Two optimized objectives:
  - Cluster hypothesis: sentences referring to the same image should be close to each other
  - Perceptual similarity: aligns the *similarity* between two images with the *similarity* of two corresponding sentences



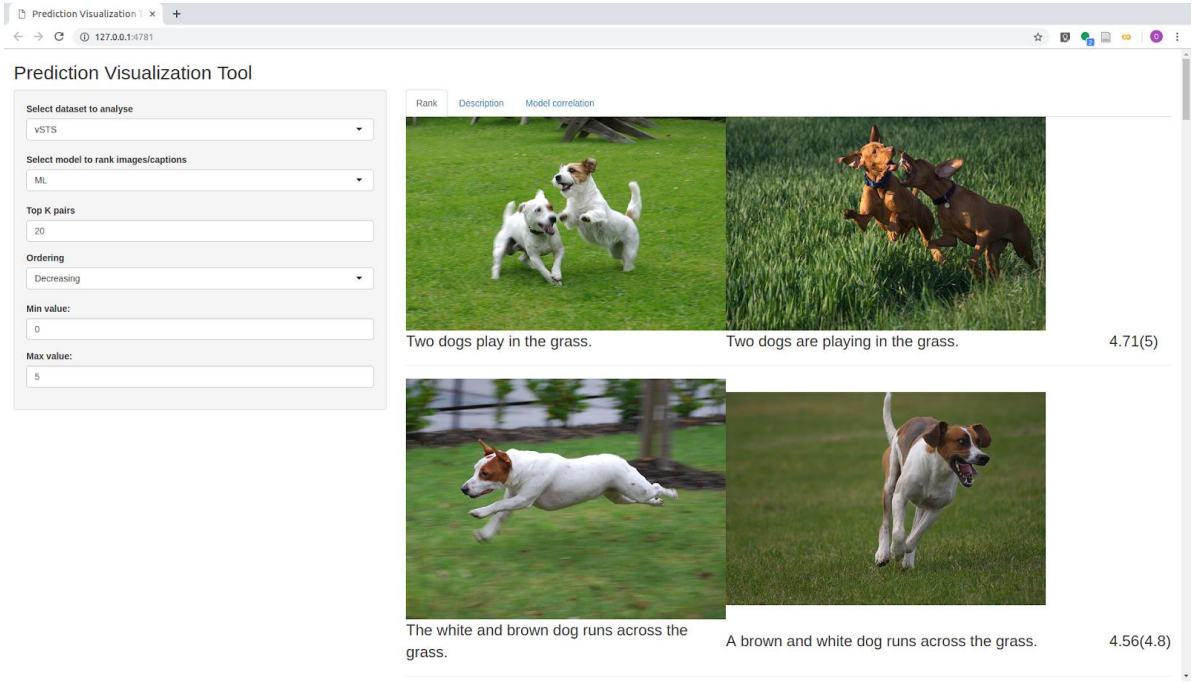


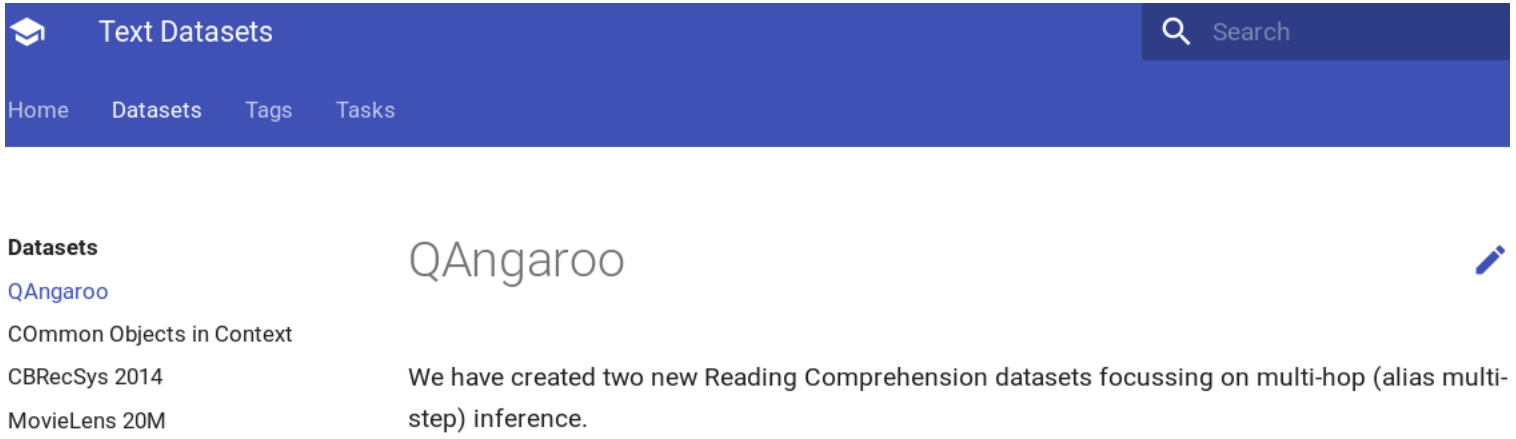
QA with images: How To Do Things

- Question Anwering with Images =
  - Series of steps
  - Each step = a textual answer and a visual answer
- The dataset is constructed by mining data from WikiHow. 200,000 triplets of
  - a question,
  - a sequence of textual answers,
  - and a sequence of corresponding visual answers

Produced Resources

- A dataset for question answering of “how to do things?”
- An interface for querying and browsing the vSTS and related datasets, where images, captions and scores are intuitively shown (https://oierldl.shinyapps.io/vSTS/)





A dataset manager tool

- Collaborative management
- Download and preparation
- Searching (command line / web interface)

Ongoing Work: General Evaluation of Multi-Modal Embeddings

All partners have started working towards a broad study to analyze multimodal representations for words and sentences produced within the project, and evaluating them in semantic tasks such as word and sentence similarity, sentence entailment, etc.

We will address the following research questions:

- Do multimodal word/sentence embeddings use a specific subspace of the embeddings space?
- What category of tasks or sentences are better handled by grounded representations: visual words (words used to describe a picture), concrete words, POS (adjectives, verbs and nouns)?
- How do words/sentence embeddings spaces relate: do they change the topology of the space (i.e. nearest neighbors, etc.)?
- What information is not brought from the visual space with current models?

Overall production

- 26 papers
  - 3 under review + 1 in preparation
  - 15 international conferences + 2 workshops
  - 3 journals
  - 2 national conferences
- Resources
  - Tools:
    - Datasets managers
    - Visualisation for visual sentence similarity
    - Annotation interface for videos (object tracking)
  - An evaluation framework for word embeddings
  - Multimodal and multilingual embeddings (sentence and words)
  - Datasets:
    - Video object tracking
    - Visual How To
    - Visual Semantic Text Similarity (vSTS) and Word Sense Disambiguation (vWSD)