Deep learning of textual and visual inference - IXA group, CHISTERA

Proposers: Oier Lopez de Lacalle, Aitor Soroa, Eneko Agirre (IXA http://ixa.eus)

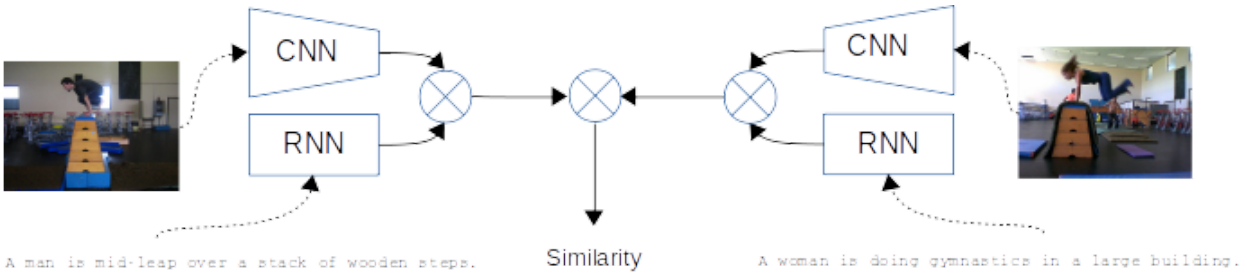
Contact: e.agirre@ehu.eus http://ixa2.si.ehu.eus/eneko

Description

The success of word representations (embeddings) learned from text has motivated analogous methods to learn representations of longer sequences of text such as sentences, a fundamental step on any task requiring some level of text understanding (Pagliardini et al., 2017). In order to evaluate sentence representations, intermediate tasks such as Semantic Textual Similarity (STS) (Cer et al., 2017) or Natural Language Inference (NLI) (Bowman et al., 2015) have been proposed. STS assesses the degree to which two sentences are semantically equivalent to each other. The STS task is motivated by the observation that accurately modeling the meaning similarity of sentences is a foundational language understanding problem relevant to numerous applications including: machine translation (MT), summarization, generation, question answering (QA), short answer grading, semantic search, dialog and conversational systems.

In another strand of related work, tasks that combine representations of multiple modalities have gained increasing attention, including image-caption retrieval, video and text alignment, caption generation, and visual question answering. Approaches to theses tasks are commonly based implemented using the latest Deep Learning techniques such as Convolutional Networks, recurrent neural networks and sophisticated attention mechanisms.

In this project will apply multimodal deep learning methods to build models that able to approach to Visual Semantic Textual Similarity (vSTS) task(Lopez de Lacalle et al., 2017). In this dataset the model have access to the corresponding images and the caption that describes the image, in contrast with having access to text alone. Figure below shows one possible architecture that combine CNN and RNN modules with attention mechanism (Peng et al., 2017).



The student will acquire:

- Deep learning applied to text and visual modalities
- Knowledge about interdisciplinary field of computer vision and natural language processing.

This project is in the context of the CHISTERA project http://www.chistera.eu/projects/muster involving four universities.

Goals

The student will apply deep learning in order to build a Multimodal (image+text) Deep Learning model able to assess the similarity degree of a given pair of text and image. The key objectives are the following:

1. Analysis of the state of the art techniques for multimodal deep learning systems.
2. Design of an multimodal deep learning architecture that provides similarity scores
3. Implementation and evaluation of the model on VSTS

Requirements

English. Machine learning. Good programming skills, basic math skills.

Although it is not a requirement, taking the course "**Seminar on language technologies. Deep Learning**" (see below) will allow the student to accomplish more ambitious goals. Contact us for further details.

The dissertation can be written in Basque, English or Spanish.

Framework

Python, Tensorflow (but this is not a hard constraint, the student may use other framework like Pytorch)

Tasks and plan

Dec-Jan: Study literature

Feb: Attend course "Seminar on language technologies. Deep Learning" (see below)

Mar-May: Development and experiments

June: Write down and presentation

References

M. Pagliardini, P. Gupta, and M. Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. CoRR, abs/1703.02507, 2017

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity multi-lingual and crosslingual focused evaluation. In Proceedingsof the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

Oier Lopez de Lacalle, Eneko Agirre, Aitor Soroa. Evaluating Multimodal Representations on Sentence Similarity:vSTS, Visual Semantic Textual Similarity Dataset. ICCV17: second workshop on Closing the Loop Between Vision and Language. Venice, Italy, 2017

Y. Peng, J. Qi, and Y. Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. CoRR, abs/1708.04776, 2017.

Seminar on Language Technologies. Deep learning. (LAP18)

Deep Learning neural network models have been successfully applied to natural language processing. These models are able to infer a continuous representation for words and sentences, instead of using hand-engineered features as in other machine learning approaches. The seminar will introduce the main deep learning models used in natural language processing, allowing the students to gain hands-on understanding and implementation of them in Tensorflow .

Topics

- Introduction to machine learning and NLP with Tensorflow
- Deep learning
- Word embeddings
- Language modeling and recurrent neural networks
- Convolutional neural networks
- Attention mechanisms

Prerequisite. Basic programming experience, a university-level course in computer science and experience in Python. Basic math skills (algebra or pre-calculus) are also needed.