Goals

Multimodal inference across image data and text has the potential to improve the understanding of information in different modalities and the acquisition of new knowledge, which can inform applications such as visual question answering, image captioning, visual reasoning, and the joint classification of multimodal data.

Taking inspiration on textual entailment and natural language inference studies (e.g., see https://nlp.stanford.edu/projects/snli/ and https://cims.nyu.edu/~sbowman/multinli/), previous work has proposed visual-textual entailment tasks, related to determining if a textual hypothesis can be concluded from a premise image, and assigning to each pair of (premise image, textual hypothesis) a label among entailment, neutral, and contradiction (e.g., see https://leonidk.com/pdfs/cs224u.pdf or https://arxiv.org/abs/2004.01894). Neural network methods have also been proposed to address the aforementioned visual-textual entailment task, e.g. combining convolutional networks to encode the visual premise together with recurrent neural networks to encode the textual hypothesis (e.g., see https://arxiv.org/abs/1901.06706 or https://arxiv.org/abd/1806.05645).

Despite the recent progress in the area, there are still many opportunities to be explored. This M.Sc. research project will involve the use of neural models for visual-textual entailment/inferencing, taking advantage of recent developments in terms of Transformer models trained to map natural language in reference to visual concepts (e.g., models such as OpenAI CLIP -- https://arxiv.org/abs/2103.00020). Besides the direct application to visual-textual entailment/inferencing, the candidate will also assess if the models fine-tuned for entailment/inferencing can transfer to other downstream tasks such as multimodal content geo-localization (i.e., assigning images, textual documents, or resources combining both types of content, to the corresponding geospatial locations).

Experiments will mostly leverage the SNLI-VE-2.0 and e-SNLI-VE-2.0 corpora (see https://arxiv.org/abs/2004.03744), although other recently proposed datasets/tasks can also be explored (e.g., https://arxiv.org/abs/1811.00491 , https://www.aclweb.org/anthology/D19-1469.pdf or https://arxiv.org/abs/2005.00908), including datasets derived from Wikipedia and/or from Flickr for evaluating geo-localization performance. The tests will involve assessing model performance under different multimodal scenarios that involve the use of neural networks for (a) determining if a textual hypothesis can be concluded from a premise image, (b) determining if a textual hypothesis can be concluded from a premise composed of a textual description and a context image, and (c) determine if a location hypothesis can be derived from a premise consisting of an image and/or a textual description.

Requirements

Commitment and availability to work on the project (e.g., not recommended for students with other professional activities);

Interest in deep learning, computer vision, natural language processing, and multimodal information retrieval -- previous experience with the specific topics of deep learning for natural language processing and/or computer vision will be valued.

Good commandment of English;

Excellent grades (above 17 values) in courses related to the topics of the project (i.e., courses on machine learning or data science);

Knowledge/experience with tools like Overleaf and GitHub;

Knowledge of Python and machine learning libraries such as Pytorch or Tensorflow.

BERT

Hugging Transformers -> BERT + ViT/ CLIP / VisualBERT