Semantic multi-object recognition via attention

1. Object detection

*If you are familiar with object-recognition models feel free to skip to section 2.*

Multi-object detection is a complex, exciting and active research topic within the machine learning community. It can be unfolded in two subproblems: recognition and detection. The first try to answer the question “is the object in the image?” while the second answers “where?”.

Classical computer vision research tried to describe objects with manually defined features. For example a dog is composed of smaller and easier to recognize objects such as eyes, legs, fur, a nose, etc. In reality the features were far more abstract, such as corners, and different combinations of polygonal shapes. Some examples are the Viola-Jones algorithm, bag-of-words, or HOG feature extraction techniques with SVM classifiers. However, in the last few years the field has been dominated by deep learning techniques, in particular, by convolutional neural networks. The major strength of these models are the ability to extract features of the objects automatically. Some of the more popular include R-CNN, fast R-CNN, Deep MultiBox, or YOLO which is capable of detecting more than 9000 objects in real-time.

2. Improving objects classification with semantic information

Most of the object-recognition architectures consist of two stages. Firstly, a subnetwork gives a set of regions of interest (ROI) where objects are likely to be. After, each of these regions bounding boxes are refined and classified with a label. However, the second stage is done independently for each element and information between is vaguely shared through the convolution embeddings. Semantic information is a powerful source of label discrimination. For example a ship it’s very likely to be on water.

The goal of this project is to make this information more explicit to the model by classifying all the ROIs simultaneously. One approach is to leverage the popular transformers architectures which are capable of processing efficiently many inputs at once finding meaningful relations between them. Although initially transformers were applied on language task and sequence modelling, latest research has expanded its application to more generic inputs/output schemes. Specifically for our interest, they have been applied to process bounding boxes as seen in the LayoutLM architecture.
The direction would fuse both architectures to have an end-to-end object recognition pipeline.

**Project outline**

1. Get familiar with Mask-RCNN or any similar OR architecture
2. Find a suitable dataset for multi-object recognition (e.g. COCO)
3. Analyze OR models in terms of accuracy, data efficiency etc..
4. Get familiar with LayoutLM architecture by training a simple masking model
5. Generate a dataset of ROI embeddings
6. Train separately LayoutLM on such embeddings
7. Ensemble and end-to-end pipeline combining OR and LayoutLM
8. Evaluate and analyze results

**Target conférences**

Should this work be successful, it would be submitted to one of the following conferences in priority: ICLR, NIPS, ICML, CVPR, ECCV, ICCV.

**Literature**

[1] Object detection


[2] Attention and Transformers

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin; “Attention Is All You Need”

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou; “LayoutLM: Pre-training of Text and Layout for Document Image Understanding”