Inspecting and Debugging a VQA System through Explanations

Background

Visual Question Answering (VQA) is a research area concerned with the construction of computer systems that can answer questions in natural language from the contents of an image. VQA carries potential applications in multimodal information retrieval.

Current VQA solutions rely on deep learning techniques. Being a problem on multimodal data, this implies to merge both images and questions into a common representation space. This is challenging because images and texts are very different data types, treated by means of different neural network architectures: CNNs (Convolutional Neural Networks) are the state of the art for image classification and representation, whereas text processing often resorts to RNNs (Recurrent Neural Networks). VQA solutions must orchestrate both technologies leading to systems that are extremely complex.

The complexity of existing VQA solutions makes the task of inspection and debugging very hard. In particular, neural networks are black-box models: one requires a significant amount of work and solid expertise to understand the inner-workings of the network. It becomes therefore very difficult to understand why a VQA system makes a mistake. Such a task, however, is vital for the progress of research in VQA.

Main activities

The post-doctoral researcher in charge of this project will work on methods to explain the output of a VQA system in order to understand why it erred (or not).

In this regard, we aim at deploying post-hoc interpretability modules that can provide hints on the logic behind a VQA module for a given case. Techniques such as LIME [Ribeiro et. al., 2016], SHAP [Lundberg and Lee, 2017], or Anchors [Ribeiro et. al., 2018] provide the foundations to generate such explanations for any type of black-box model. That said, those techniques deliver explanations in terms of an interpretable space that depends on the data type (e.g., tabular data, texts, images) and is usually very different from the representation space of the black-box. For instance, pixel channels and feature maps in images are replaced by super-pixels (image segments) in explanations, whereas word embeddings are substituted by word occurrences. No research work until now has tried to reconcile those representations in a multimodal setting, thus the main challenge is to find a common interpretable representation for explanations that encompasses both images and texts in a VQA setting. We are particularly interested in representations that resort to semantically

meaningful units such as known objects or patterns (e.g., pointy borders, a nose, limbs, vehicules) as studied in network dissection. This would allow us to yield explanations in terms of the presence or absence of those objects. Such explanations could be enhanced with semantic knowledge, such as a taxonomy, in order to signal interesting associations automatically, e.g., pointy borders in a bird may refer to its beak. Other recent approaches [Shi et al., 2019], [Yi et al., 2018] have focused on the extraction of knowledge from the input before defining a reasoning program to execute. This knowledge may be a start in the definition of an interpretable space for explanations of VQA systems.

Traditional post-hoc interpretability modules do not make any assumptions about the architecture of the model they try to explain, i.e., they are model-agnostic. In the context of VQA systems, a possible solution is to drop this assumption and mine neuron activation patterns that identify the instances for which the system fails. We could use contrast pattern mining techniques for this purpose.

Our research will make use of publicly available VQA datasets such as GQA, VQA-E, C-VQA, and CLEVR.

Skills

We are searching for motivated candidates with a PhD degree in Computer Science and with competences in machine learning (preferably with focus on deep learning). Knowledge in data mining, e.g., sequence and itemset mining, will be also appreciated.

The candidate should be proficient in written and spoken English (at least B2 level according to the CEFR system).

Assignment

To apply for the position, the candidate must send an email to the list of contacts below. The email should include:

- A CV
- A statement letter explaining the candidate's motivations to apply for the position
- At least two recommendation letters

Context

This position is open at Inria, in France and is part of the HyAIAI Inria challenge. More specifically, this position is part of a collaboration between the Lacodam and SequeL Inria teams. The post-doc will share her time between the two groups (Lacodam in Rennes, SequeL in Lille). She will actively participate in the activities of HyAIAI, in particular report on her work in the HYAIAI meetings.

Contact people

Philippe Preux (philippe.preux@inria.fr)
Luis Galárraga (luis.galarraga@inria.fr)