Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Hoa Trong Vu
Název práce	Grounding Natural Language Inference on Images
Rok odevzdání	2018
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Jindřich Libovický **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

The thesis introduces a new task of visually informed natural language inference (NLI). It uses the fact that one of the commonly used NLI dataset was created using simple sentences which come from an image captioning dataset. The original images are used as an additional source of information in experiments trying to improve the NLI system performance using the visual information.

The experimental methodology is appropriate and well-managed. Except for standard quantitative analysis of the model results, the thesis also contains a very laborious analysis of the results.

The results and the subsequent analysis are negatively affected by the choice of the dataset. The original NLI dataset that is extended with the images was designed in such a way that it could be solved without visual information. I suppose (although I was not able to verify that) that the annotators creating the dataset did not see the original images. Under such circumstances, we cannot expect the large gain in the model performance, even if we were able to develop a model that is capable of modeling the multimodal aspects of the task. Creating a dataset where both the textual and visual information are required for the inference might lead to more interesting experimental results.

The biggest weakness of the thesis is its poor structure. The related work should not be part of the introduction, but rather a standalone chapter. The thesis attempts to solve a novel task, description of the task including the evaluation should have a more prominent place in the thesis, rather than a few paragraphs within the Introduction chapter.

At many places, the author mixes the task description and approaches taken to solve the tasks. For instance, the second sentence of paragraph introducing language-and-vision problems mentions multiplication of vector representation. The NLI task description is interleaved with a description of outdated systems explicitly utilizing lexical overlap features. The Introduction also mentions details neural architectures used to solve NLI, although the building blocks and basic concepts of the architectures are explained later in the thesis in Chapters 2.

Chapter 2 introducing basics of neural networks and does not explain the notation that is used through the chapter, readers need to guess that x is input, y output, t target etc. The chapter is also not well structured. For instance, the back-propagation algorithm that is used for training all types of neural networks is presented as a subsection of section introducing feed-forward networks. In general, the chapter provides many details, however important pieces of information are missing.

Chapters describing neural architectures used in the experiments and the experiments themselves are better structured than the previous chapters. The choice of the architectures is well justified. The qualitative analysis of the results is outstanding. However, due to the choice of the task, the results do not lead to particularly interesting conclusions.

Questions for the author:

1. Did you do any prior analysis that would suggest that the dataset consist examples where the visual information might helpful for infrerence?

2. Experiments with multimodal machine translation and visual question answering show that using image representation from the ResNet-based architectures (first version were available since 2015) lead to much better results than using the VGG network (from 2014). Do you have any specific reason for using VGG network?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 4. 9. 2018

Podpis: