Jan Čech
Center for Machine Perception
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague

Doctoral Thesis Review

Prague, April 26, 2019

Title: Multimodality in Machine Translation
Author: Mgr. Jindřich Libovický
Supervisor: doc. RNDr. Pavel Pecina, Ph.D.
Date received: 06/04/2019

The thesis addresses the problem of multimodal machine translation, i.e. a machine translation when an image is available as an auxiliary input to the algorithm besides the textual input. The problem is illustrated on translating image captions from a source language into the target language given both the image and its caption.

Besides that, other multimodal respects of the machine translation are demonstrated and analysed: Experiments when the image is available only in training to capture a better representation, or when the image representation is hallucinated by the system itself in testing, are conducted. Moreover, a related problem of multi-source machine translation (a text in multiple source languages is translated into a target language) is examined.

The manuscript is generally well written and organized. The thesis consists of seven chapters. The first chapter introduces the problem and the thesis structure. The second chapter introduces the methodology of deep learning in computer vision and natural language processing. The third chapter reviews related literature on combining language and vision. The fourth chapter is focused on introducing the multimodal translation problem, reviewing the state-of-the-art works, presenting available datasets, and evaluation statistics. The main contribution of the author is summarized in a technical chapter five. The sixth chapter gives a detailed statistical analysis of the multimodal translation systems. Finally, chapter seven concludes the thesis.

The thesis addresses an interesting an important problem. Understanding a combination of modalities of a very different nature (in dimensionality, discrete symbols/continuous), text and images, is of a crucial importance with an impact beyound the machine translation. The thesis thoroughly analysed options to integrate the visual information, which are nicely categorized on p. 61. Then the proposed solutions in integrating the visual input (or multiple sources) are original and scientifically significant. The novel modifications follow the state-of-the-art machine translation model architectures: recurrent neural networks and self-attentive networks. Moreover, the thesis examines several possible strategies to integrate the auxiliary modality within the model. All the strategies are compared using meticulous quantitative evaluation. The experimental validation is very detailed and thus convincing. I especially appreciate the experiment with adversarial images that was designed to assess the amount of the auxiliary visual information the model takes into account when producing the translation. The statistical correlation analysis of the learned models concludes that despite the translation quality does not improve much having the access to the input image, the visual information provides a stronger training signal for sentence representation than language modelling alone while training on a much smaller dataset. I see this finding as a particularly interesting research result.

To name a few weaknesses of the thesis:

1. The image does not help the machine translation of the captions very much. This is probably caused by the advance of the state-of-the-art machine translation that makes the auxiliary image information rather redundant. However, this seemingly negative result is well balanced by thorough quantitative introspection of the models and deeper insight in the learned representation. Nevertheless, despite the fact that the average BLEU score over the entire test set is not improved significantly by the visual input, there might be particular examples that are disambiguated by the image - possibly out of the test set, and these rare examples should have been shown. See Question 1.

2. Couple of mistakes in equations and around them. For instance, the convolution defined above Eq. (2.3) is said to be a *non-linear* projection, while the convolution itself is a linear operation, that could be written by (Töplitz) matrix multiplication. Probably a coupling with the non-linear activation was meant. The batch normalization defined in Eq. (2.6) omits the trainable parameters β, γ that scale the normalized variables following Ioffe and Szegedy, 2015. The BLEU score definition on p. 53 contains mistakes. Checking with the original paper by Papineni et al., 2012, there should be $\log p_n$ instead of p_n in Eq. (4.2) , and $r < c$ instead of $h > c$ in Eq. (4.1).

3. Sometimes, the text of the thesis is not self-contained in presenting essential concepts and a reader needs to consult with the original papers. For instance, the transformer architecture by Vaswani et al., 2017, described in Chapter 2 and later again in Chapter 5 does not properly explain what certain identifiers (keyes K, values V, queries Q) and symbols ($\oplus$, concatenation) are, e.g. in Eq. (2.27), (2.29), (5.14).

4. I would suggest showing more qualitative examples of the image caption translation. It might be interesting to sort test set samples by their BLEU scores and then show the best 10 results and the worst 10 results. This would provide an intuition what the range of quality is besides the exhaustive statistics computed over the entire test set.

Nevertheless the above weaknesses are minor and does not compromise an overall high quality of the thesis. The author successfully regularly participated in the WMT Multimodal Translation Challenges and published 7 research papers. The author clearly proved his ability to perform research and to achieve scientific results. The papers have in summary 77 citations (without self-citations at the time of thesis submission), which is impressive and confirms the impact of the research to the community.

In conclusion, I do recommend the thesis for presentation with the aim of receiving the Degree of Ph.D.

Ing. Jan Čech, Ph.D.

Questions for the defense

1. Could you find a couple of examples when a visual information helps in resolving ambiguity of the text in the image caption translation? These examples could perhaps be found automatically by comparing BLEU scores between the multimodal and the textual model outputs.

2. Why is there the disbalance among source languages in multi-source translation? All strategies seem to rely prominently (or even almost exclusively) on English, see Tab. 5.4.

3. The imagination model presented in Sec. 5.3.2 is an interesting concept. Would it be possible to visualize the image from the "hallucinated" representation?