Multimodality in Machine Translation

Jindřich Libovický

Traditionally, most natural language processing tasks are solved within the language, relying on distributional properties of words. Representation learning abilities of deep learning recently allowed using additional information source by grounding the representations in the visual modality. One of the tasks that attempt to exploit the visual information is multimodal machine translation: translation of image captions when having access to the original image.

The thesis summarizes joint processing of language and real-world images using deep learning. It gives an overview of the state of the art in multimodal machine translation and describes our original contribution to solving this task. We introduce methods of combining multiple inputs of possibly different modalities in recurrent and self-attentive sequence-to-sequence models and show results on multimodal machine translation and other tasks related to machine translation. Finally, we analyze how the multimodality influences the semantic properties of the sentence representation learned by the networks and how that relates to translation quality.