**Paper: Attention on Attention (AoA) for Image Captioning**

Name: Milon Hossain

ID: 1610497042

**Abstract:** Attention mechanism are widely used in current image captioning encoder/decoder frameworks, where a weighted average is generated to guide the caption decoding process on encoded vectors at each time step. The decoder, however, has no knowledge of whether or how well the vector attended and the attention question given are related, which may cause the decoder to give false results. AoA first generate an “information vector” and an “attention gate” using the attention result and the current context, then adds another attention by applying element-wise multiplication to them and finally obtains the “attended information”, the expected useful knowledge. In this project AoA apply to both the encoder and the decoder for the image captioning model which named as AoA

**Introduction:** Image captioning is one of the primary goals of computer vision which aims to automatically generate natural description for images. It requires not only to recognize salient objects in an image, understand their interactions, but also to verbalize them using natural language, which makes itself very challenging. The attention mechanism plays a crucial role in such a system that must capture global dependencies. A model for the sequence to sequence learning task like image/video captioning, since the output is directly conditioned on the attention result.

**Dataset:** A popular MS COCO dataset have been used in this project. MS COCO dataset contains 123,287 images labeled with 5 captions for each, including 82,783 training images and 40,504 validation images. MS COCO also provides 40,775 images as test set for online evaluation. This dataset is used to train the model.

**Attention on Attention:** An attention module fatt (**Q, K, V**) operates on some quires, keys and values and generates some weighted average vectors. It first measures the similarities between **Q** and **K** and then uses the similarity score to compute weighted average vector over **V.** The attention module outputs a weighted average for each query, no matter whether or how **Q** and **K**/**V** are related. Even when there are no relevant vectors, the attention module still generates a weighted average vector, which can be irrelevant or even misleading information.

**Method:** first implement the Attention on Attention (AoA) module and then show how they derive AoANet for image captioning by applying AoA to the image encoder and the caption decoder. Later Built a model, AoANet, for image captioning based on the encoder/decoder framework, where both the encoder and decoder are incorporated with an AoA module. For the accurate captioning evaluation metrics different approaches “BLEU”, “METEOR”, “ROGUE-L”, “CIDER-D” and “SPICE” have been used.

**Analysis:** Here we discuss the evaluation from different point of view.

1. **Offline Evaluation:** The model includes LSTM which encodes the image using CNN and decodes it using LSTM. SCST is used to optimize the evaluation metrics. Up-Down employs a two LSTM layer model with Bottom-Up features extracted from faster RCNN. For fair comparison, all the models are first trained under XE (Cross Entropy) loss and then optimized for CIDEr-D score.
2. **Online Evaluation:** It also evaluate their model on the online COCO test server. The results of AoANet are evaluated by an ensemble of four models trained on the “Karpathy” training split. AoANet achieves the highest scores for most metrics.
3. **Qualitative Evaluation:** From various examples with images and captions generated by the AoANet model and a strong baseline as well as the human-annotated ground truths. The baseline model generates captions which are in the line with the logic of language but inaccurate for the image content, While AoANet generates accurate caption with high quality. AoANet count the objects of the same kind more accurately and AoANet figure out the interactions of the objects in an image.
4. **Ablative Evaluation:** To quantify the impact of the proposed AoA module, they compare AoANet against a set of other ablated models with various settings. First design the “base” model which doesn’t have a refining module in its encoder and adopts a “base” decoder using linear transformation to generate the context vector.

**Conclusion:** In conclusion, An Attention on Attention (AoA) module, an extension to conventional attention mechanism, to address the irrelevant issues. By using this model, we can reduce the length of a caption which had been generated when we give an image as an input. It removes repetitive word which are unnecessary for our caption and generate the appropriate meaningful caption.

**Reference:**

1. <https://openaccess.thecvf.com/content_ICCV_2019/html/Huang_Attention_on_Attention_for_Image_Captioning_ICCV_2019_paper.html>