A Tale of Two Models: Comparative Analysis of Image Captioning Models with Varying Attention Techniques

Charlie Sun **Adam Musa** **Raymond Li**
{charlie.sun, adamm.musa, raymondlf.li}@mail.utoronto.ca

Abstract

Image Captioning involves analyzing features and objects within an image to generate a textual description. The automation of this process through neural networks has led to remarkable innovations and advancements, resulting in numerous deep-learning algorithms that achieve human-like performance. In this study, we aim to evaluate various image captioning algorithms and methods to understand the strengths and weaknesses of different models. Specifically, we focus on the earlier model presented in the paper Show, Attend and Tell, and the more recent model described in Attention on Attention for Image Captioning. Furthermore, we investigate the robustness of these models through a series of analyses.

1 Introduction

The fields of computer vision (CV) and natural language processing (NLP) have experienced rapid development over recent years, resulting in advanced object recognition systems and highly interactive natural language chat models. Combining CV and NLP has given rise to research areas such as image captioning. Over the past years, image captioning models have exhibited consistent improvements in both syntactic correctness and caption richness. Nevertheless, image captioning remains a challenging task, despite the individual advances in CV and NLP.

In this paper, our primary goal is to reproduce and analyze past models in the field of image captioning. We focus on one earlier model and one newer model, training both on the same dataset to compare their performances. Furthermore, we conduct additional experiments and analyses to evaluate the robustness of these models under different conditions.

2 Related Works

In the realm of image captioning, various methodologies have been proposed and explored. The basic encoder-decoder framework, for instance, employs a convolutional neural network (CNN) as the encoder to extract image features and a recurrent neural network (RNN) as the decoder to generate captions. This framework was later enhanced by the attention mechanism, forming the basis of a series of attention-enhanced methods, which are able to concentrate on relevant image regions during the caption generation process. Such methods encompass a range of attention types, including hard and soft attention, area-based attention, adaptive attention, and bottom-up and top-down attention. The Bottom-Up and Top-Down paper [2] is a notable example of this category, in which the model selectively focuses on salient image regions by integrating low-level visual features with high-level semantic information, enhancing the overall description quality.

More recent advancements have seen the incorporation of transformer-based methods, where these approaches replace either the encoder or decoder with self-attention. For example, the Global Representation paper [8] introduces the Global Enhanced Transformer, which extracts a comprehensive global representation from the image, it then acts as a guide to the decoder. Another set of methods,

Preprint. Under review.

post-editing based approaches, utilize a two-pass decoding framework for caption generation. The first-pass model generates an initial caption, which is subsequently corrected or refined using diverse strategies. This assortment of techniques demonstrates the diverse landscape of image captioning methodologies within the machine learning community [3].

3 Methodology

In this section, we present and summarize the two models that will be trained and assessed.

3.1 Show, Attend and Tell

The Show, Attend and Tell model was introduced in 2016 as one of the attention-enhanced methods for image captioning [15]. At a high level, it employs a Convolutional Neural Network (CNN) for encoding visual information and a Recurrent Neural Network (RNN) to decode the information into a sentence.

3.1.1 Encoder

First, a CNN is used to extract a set of feature vectors, referred to as *annotation vectors*:

$$a = \{a_1, \dots a_L\}, a_i \in \mathbb{R}^D$$

The feature extractor does not use any fully connected layers, ensuring that the feature vectors correspond to positions in the original image. This allows the decoder to focus on different parts of the image.

3.1.2 Decoder

A Long Short-Term Memory (LSTM) network is used as the decoder, with attended features acting as the context vector.

3.1.3 Attention mechanism

The authors presented two attention mechanisms. Both compute the attention score by using the LSTM state as the query and annotations as the key:

$$e_{ti} = f_{\text{att}}(a_i, h_{t-1})$$

where e is the set of unnormalized scores, a is the set of annotation vectors, and h_t is the decoder hidden state at time step t:

- Soft (Bahdanau) attention: This mechanism computes a weighted sum of the annotation vectors, with the weights determined by a learned compatibility function.
- Hard (Luong) attention: This mechanism selects a single annotation vector as the context vector at each time step, rather than assigning a probability to each annotation.

3.2 Attention on Attention (AoA)

The Attention on Attention model is an RNN-based image captioning model released in 2019 [7]. The model pre-processes the images using a ResNet [5] or BottomUp [2] attention module to extract the feature vectors and attention from the images. The key feature of this model is the AoA module, from which the model derives its name, used in both the Encoder and Decoder.

3.2.1 AoA Module

AoA first acquires the Attended result by using self-attention with Query, Key and Value, and then generates an information vector and an attention gate using the attention result and the attention query. Then, more attention is introduced by applying the gate to the information and obtains the attended information (Figure 3a):

$$AoA(A, Q) = \sigma(W^g concat(A, Q)) \odot (W^i concat(A, Q))$$

3.2.2 Encoder

The AoANet encoder is constructed by stacking multiple encoder layers and applying a layer normalization at the end. Let $\boldsymbol{A} = \boldsymbol{a_1}, ..., \boldsymbol{a_k} \in \mathbb{R}^{k \times d}$ be the attention vectors computed for all k input images. Then our encoder layer can be described as, (Figure 3b)

$$EL(\boldsymbol{A}) = LN(AoA(f_{mh-att}(\boldsymbol{W}^Q\boldsymbol{A}, \boldsymbol{W}^K\boldsymbol{A}, \boldsymbol{W}^K\boldsymbol{A}), \boldsymbol{W}^Q\boldsymbol{A}) + \boldsymbol{A})$$

where LN represents a layer normalization operation. The encoder can be simply described as

$$E(\boldsymbol{A}) = LN(EL(EL(...(EL(\boldsymbol{A})))))$$

3.2.3 Decoder

The Decoder uses an LSTM to construct an RNN to decode the information. It uses the sum of the c_i and $\bar{a} = MeanPool(E(\boldsymbol{A}))$ to denote $c_{i+1} = c_i + \bar{a}$ (c_i is context vector of ith layer). The output of the LTSM is fed into an Attention module then an AoA module in order to produce the output. This can be seen in Figure 3c.

3.3 Evaluation metrics and benchmarks

We will be using various evaluation metrics to evaluate the models' performance, including BLEU [11], METEOR [4], ROUGE-L [9], CIDEr-D [14], and SPICE [1]. These metrics measure different aspects of the generated captions, such as n-gram precision, recall, and semantic similarity, providing a comprehensive evaluation of the model's performance.

Popular datasets for image captioning tasks include Flickr8k [6], Flickr30k [16], and Microsoft COCO [10]. These datasets contain thousands of images along with multiple human-annotated captions for each image. However, due to resource constraints, only a subset of the Flickr8k dataset is used for training in our experiments.

4 Experiments

In this section, we describe our implementation and adaptation of the AoA and Show, Attend and Tell models for our experiments. We re-implemented the AoA model from scratch and adapted the existing implementation of the Show, Attend and Tell model.

Table 1: Evaluation of models using the MS COCO dataset. B#, M, R, and C refer to BLEU-#, METEOR, ROUGE-L, and CIDEr-D respectively. (—) indicates an unknown metric. s indicates the Show, Attend and Tell model.

Model	B1	B2	B3	B4	M	R	C
Author-Reported							
Soft Attention^s	70.7	49.2	34.4	24.3	23.90	—	—
Hard Attention^s	71.8	**50.4**	**35.7**	25.0	23.04	—	—
AoANet	**80.2**	—	—	**38.9**	**29.2**	**58.8**	**129.8**
Replicated							
Soft Attention^s	**47.1**	**28.6**	**15.8**	**8.3**	15.9	36.1	**22.5**
Scaled Dot Attention^s	44.6	28.3	15.6	8.0	15.4	35.9	21.4
AoANet	37.3	22.9	10.4	5.2	**19.3**	**37.3**	6.3

4.1 Training and evaluation

Both models were trained on Google Colab, and training parameters are specified in the notebooks attached in the References section. We report results based on the Microsoft COCO training dataset comprised of 82,783 images. All captions were tokenized by converting to lowercase and splitting on whitespace. All experiments used a vocabulary consisting of words that appeared more than 5 times in the Flickr8k dataset, consistent with our preprocessing during model training.

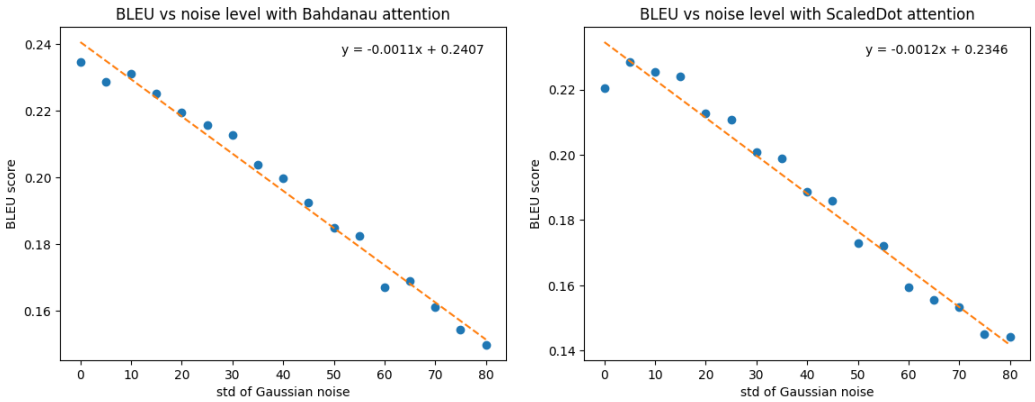
Experiment results are presented in Table 1. SPICE is omitted due to computational constraints. We find that AoANet performed better according to METEOR and ROUGE-L, but worse according to CIDEr-D and BLEU, which is likely due resource constraints.

4.2 Different attention mechanisms

We explored the use of the scaled-dot attention mechanism, an alternative to the original SA&T model's hard or soft attention mechanisms, as proposed in [13]. The primary difference between Bahdanau (soft) attention and scaled-dot attention lies in the computation of attention weights: a separate neural network learns the compatibility function in Bahdanau attention, while scaled-dot attention calculates the compatibility function as the scaled dot product of the query and key vectors. As seen in Table 1, scaled-dot attention yields **slightly lower scores** compared to soft attention, which is expected due to its simplicity. However, under our training conditions, scaled-dot attention **reduced training convergence time** by approximately 20%.

4.3 Performance under noisy conditions

Next, we evaluate the sensitivity of the Show, Attend and Tell model to noise. We add random noise ϵ to each RGB value in each image, where $\epsilon \sim \text{Gaussian}(0, \sigma)$ and σ is of our choosing. We then evaluate the model on the test set of Flickr8k using BLEU, assigning a weighting of 0.25 for each n-gram up to $n = 4$.



(a) Bahdanau attention

(b) Scaled dot-product attention

Figure 1: Performance of SA&T model with different attention mechanisms under noise.

As can be seen from Figure 1, both attention mechanisms **perform similarly** under noisy conditions, with **performance decreasing linearly** for both. An example caption that demonstrates the effect of noise can be found in Figure 6.

5 Conclusion

Our experiments successfully recreated the models presented in the original papers, reaching a reasonable degree of efficacy on test images. However, due to hardware limitations, we couldn't fully reproduce the reported results, as the GPU on Google Colab was insufficient for training on larger datasets like COCO. We also demonstrated that earlier models, such as SA&T, are sensitive to image noise, affecting performance. Moreover, on a smaller dataset like Flickr8k and with less computational resources, a simpler model like SA&T performs similarly to the more complex AoA model, which is undertrained in our implementation.

In conclusion, this study underscores the importance of selecting appropriate models and training strategies for image captioning tasks, considering factors such as available resources. Additionally, it highlights the need for robust models capable of effectively handling noise and diverse input data to ensure consistent performance across various image captioning scenarios.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822, 2016. URL http://arxiv.org/abs/1607.08822.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018.

[3] Feng Chen, Xinyi Li, Jintao Tang, Shasha Li, and Ting Wang. A survey on recent advances in image captioning. *Journal of Physics: Conference Series*, 1914(1):012053, 2021.

[4] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL https://aclanthology.org/W14-3348.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[6] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013. doi: 10.1613/jair.3994.

[7] Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. Attention on attention for image captioning. *CoRR*, abs/1908.06954, 2019. URL http://arxiv.org/abs/1908.06954.

[8] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network, 2020.

[9] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[12] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015. URL http://arxiv.org/abs/1505.04870.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[14] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL http://arxiv.org/abs/1502.03044.

[16] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

Code and Data

SA&T:

- Training: https://colab.research.google.com/drive/1xOrEoq8nSQTN0_n5Hx314joN-hTN4ICm?usp=sharing
- Evaluation: https://colab.research.google.com/drive/1mkNQgrBwvrEFnfB8uAWI7EdMQyM69u2g?usp=sharing
- Experiment: noise: https://colab.research.google.com/drive/1bNQPnLBgLE3W5wGHrdFOqut26L6h_bfj?usp=sharing

AoA:

- Training: https://colab.research.google.com/drive/1cFChbkwyAMlJmdsirH_cpAFs8Pe5eDF6?usp=sharing
- Evaluation: https://colab.research.google.com/drive/1W8Gupqw6u9QgJ4sgqNKQ21sogyrPJKjP?usp=sharing

Microsoft COCO Evaluation:

- https://colab.research.google.com/drive/1_s8pJcVs59ZY4IdBaw3rbK6tsNYplzsk?usp=sharing

Data:

- Flickr8k: [6]
- Flickr30k: [12, 16]
- COCO: [10]

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See section 5
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See section 5, the hyperparameters are listed in notebook
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See section 4.1

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] See references

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See section 5

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

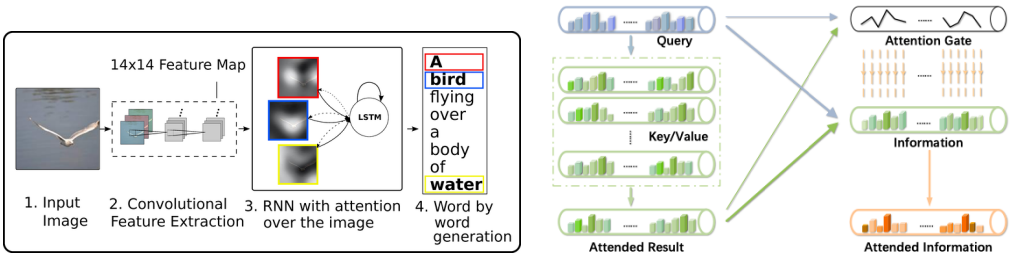
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Contributions of each team member

- **Raymond Li**: Data loading; Adapted Microsoft COCO evaluation; Write report
- **Adam Musa**: Implementing AoA model; Train and Adapt AoA model; Write report
- **Charlie Sun**: Adapt and train SA&T model; Implement and train scaled dot attention for SA&T; Perform noise analysis; Write report

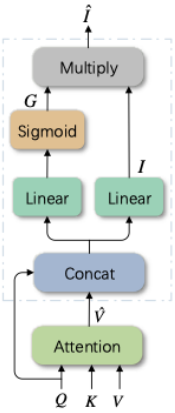
A.2 Architecture diagrams



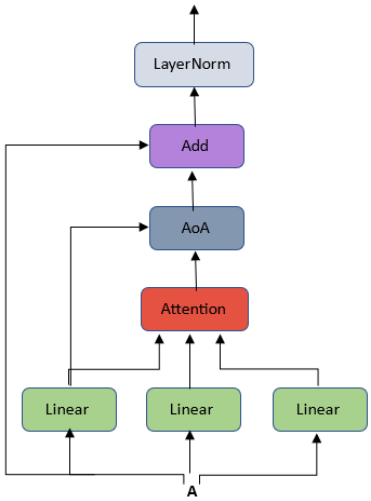
(a) Show, Attend and Tell model architecture. Adapted from `https://kelvinxu.github.io/projects/capgen.html`.

(b) Attention on Attention module, from the original paper. [7]

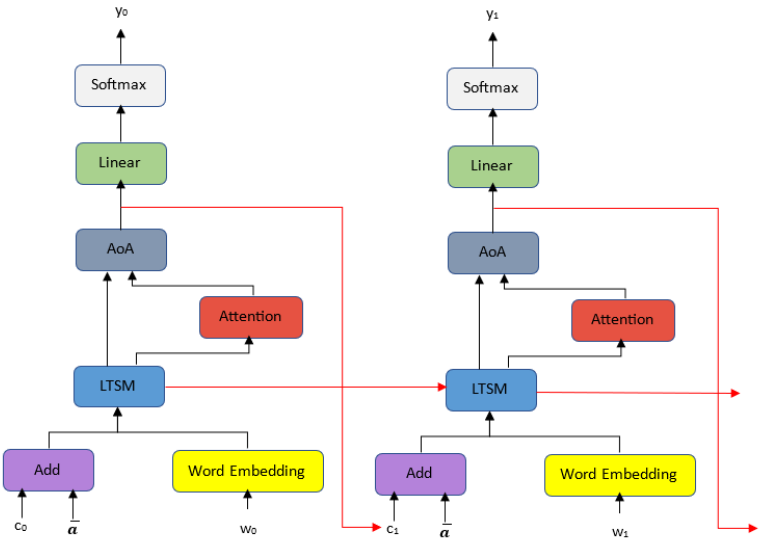
Figure 2: Model architecture diagrams.



(a) The AoA module.



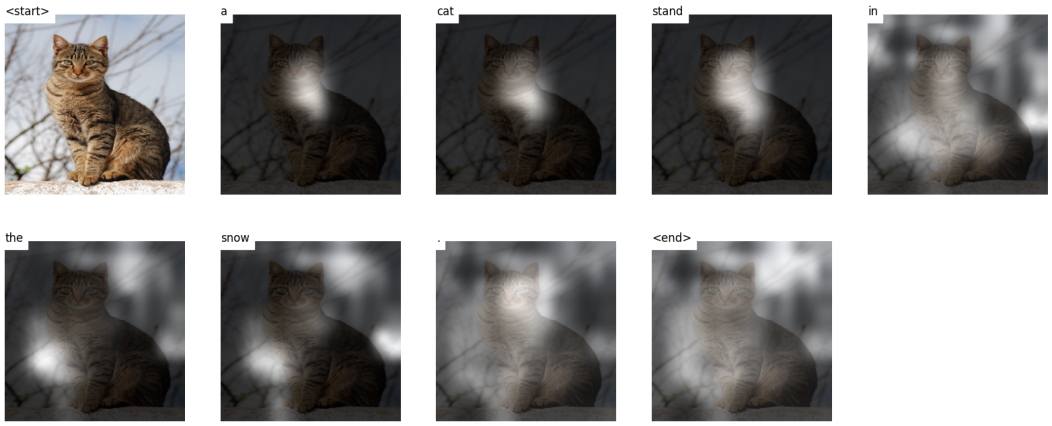
(b) A single Encoder Layer for AoANet.



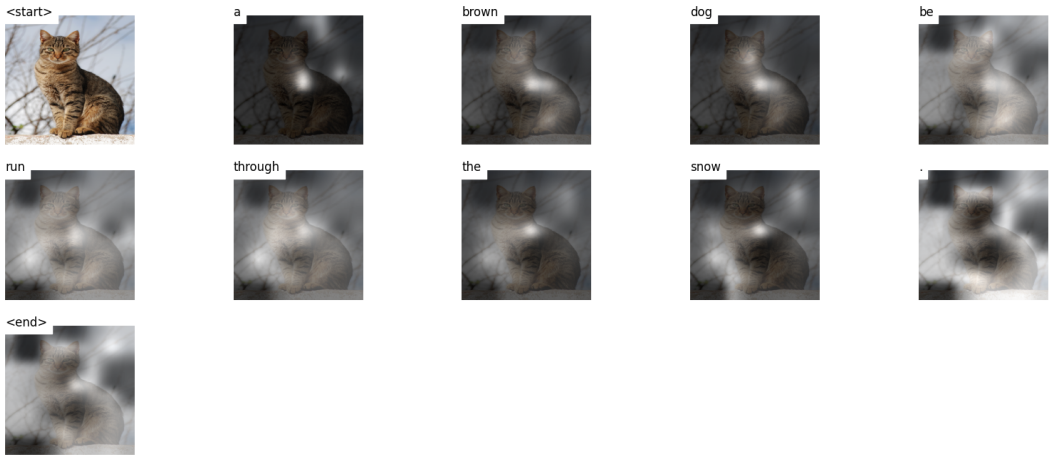
(c) The Decoder architechture for AoANet.

Figure 3: Major AoA modules

A.3 Sample translations in SA&T

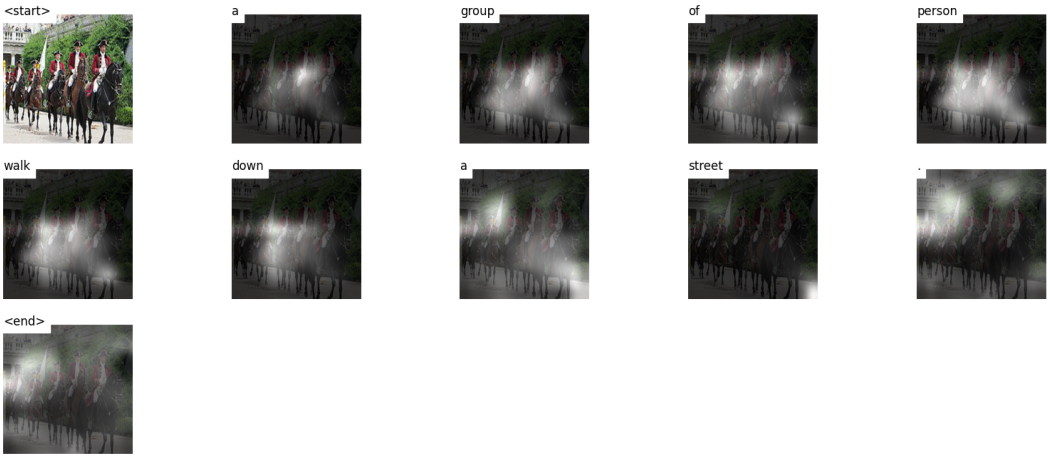


(a) Bahdanau

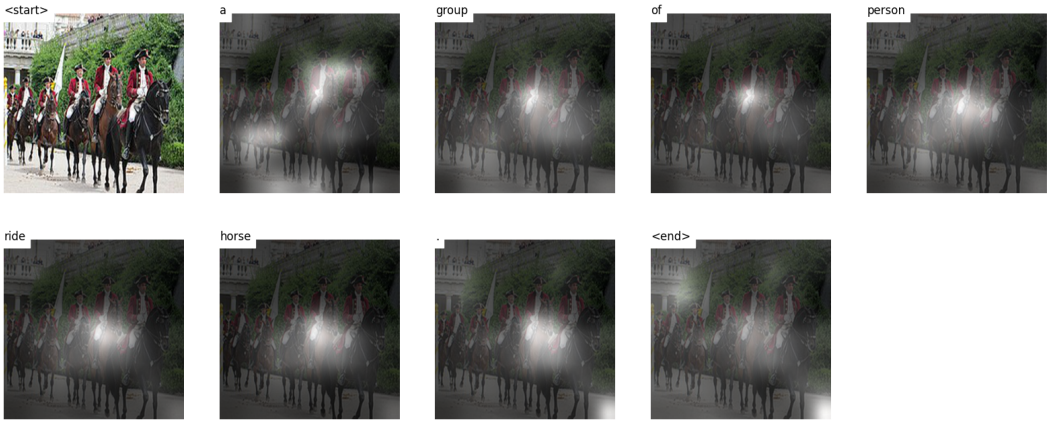


(b) Scaled-dot

Figure 4: A cat sitting in snow. From `https://en.wikipedia.org/wiki/Tabby_cat`

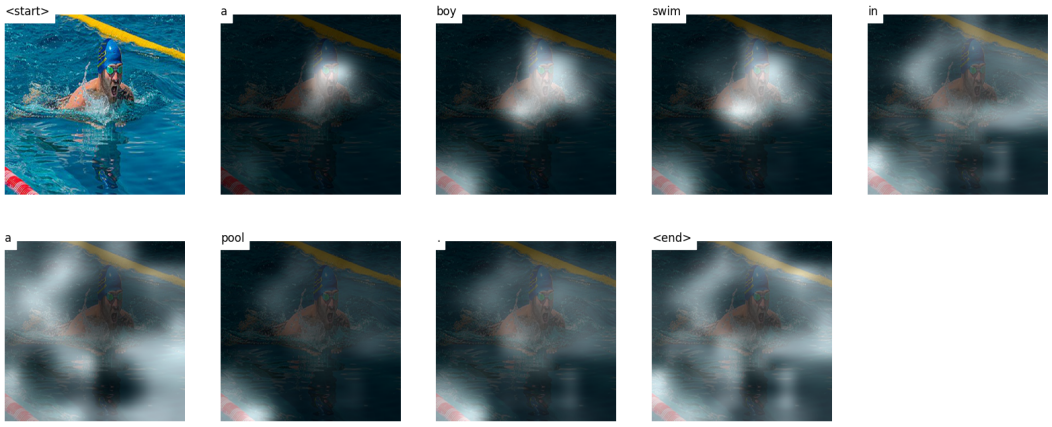


(a) Bahdanau

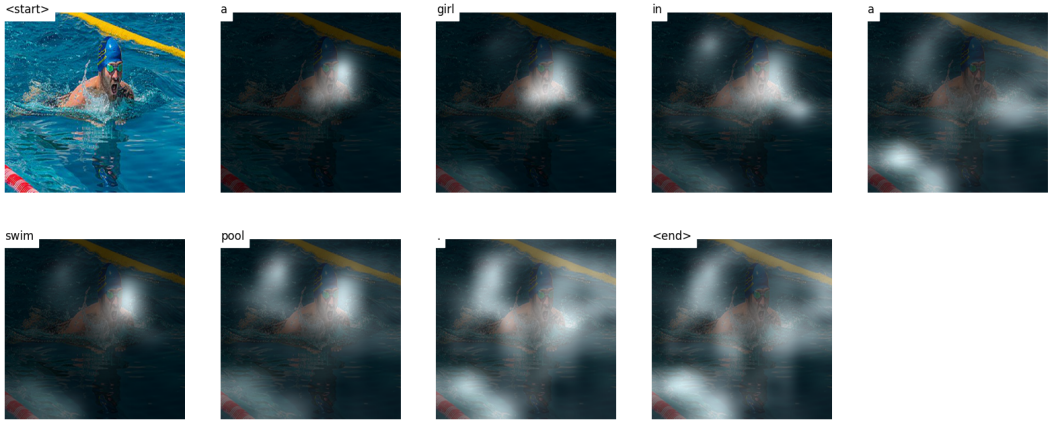


(b) Scaled-dot

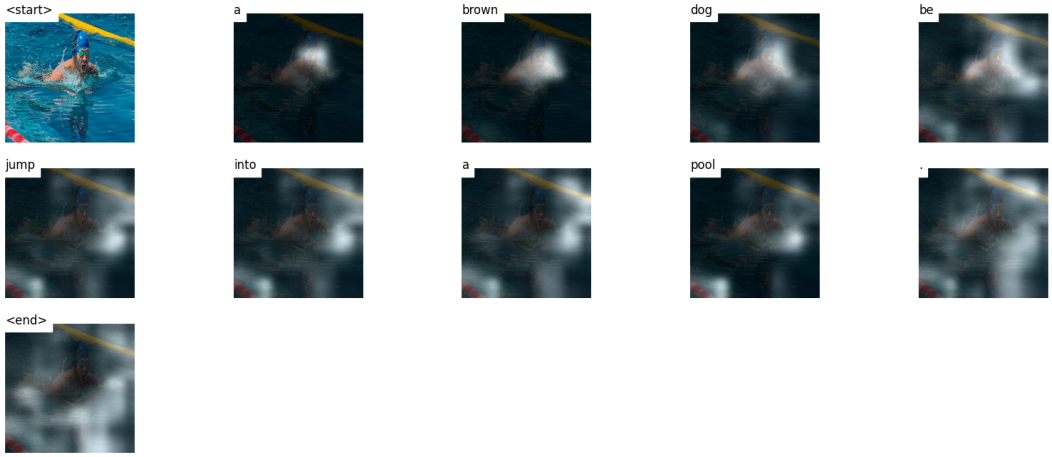
Figure 5: A group of people riding horse. From `https://en.wikipedia.org/wiki/Equestrianism`



(a) Original image



(b) Same image under noise level 20



(c) Same image under noise level 50

Figure 6: A person swimming in the pool. Different levels of noise added. From `https://en.wikipedia.org/wiki/Swimming`