Sentimentalized Image Captioning using Encoder-Decoder Networks with Attention

Peyman Bateni [*1] **Ignacio J. Iturralde Ch.** [*1]

Abstract

This paper seeks to elucidate on the sentimentalization of image captioning through the use of Deep Neural Networks. We propose two sets of models, first focused on improved visual sentiment classification using Adjective-Noun Pairs and the latter for sentimentalized image captioning. An Attention model was used to seek improvement and exploration on the model's understanding of adjectives within adjective-noun pairs. Our evaluated results show significant gains over state-of-the-art Softmax ANP classification by using Attentional Autoencoders and showcase notable findings in use of attention with respect to visual sentiment analysis/captioning. Additionally, we propose two novel approaches for sentimentalized image captioning using caption updates and noun-forcing respectively.

1. Introduction

Image captioning models take an image as input and produce a sentence or a paragraph describing the image. Captions are descriptive in noting the objects in the image, their relations to each other, actions taking place, and more depending on the task at hand. However, existing image captioning models are limited in their use of adjectives, lack emotions and don't convey sentiment. Sentimentalized captions are richer text descriptions and would allow image captioning models to describe scenes more like humans would.

On the other hand, all previous work in image captioning (that we are aware off) rely on large caption-associated image datasets, a luxury that's not present in sentimental visual captioning. Thus, existing Encoder-Decoder models, otherwise known as autoencoders, are unable to address the task at hand due to the lack of sentimentally captioned datasets. In addition, the majority of the past work in visual sentiment analysis focus on image sentiment classification, often using binary or scaled labels for predicting positivity/negativity, a level of granularity that is too shallow for sentementalized captioning.

In this paper, we propose two sets of models focusing on improving more fine-grained adjective-based sentimental classification, updating captions, using attention in identifying sentimental cues, and using both pre-training and post-evaluation caption updates to produce sentimentalized image captions.

2. Related Work

2.1. Image Captioning

In 2015, Vinyals et al. proposed Neural Image Caption, otherwise known as NIC, leveraging a CNN-encoder to create vector representation of the input image that is then used as the initial hidden state to the LSTM-decoder which produces the caption (Vinyals et al., 2014). Soon after, Xu et. al showcased the attention mechanism, improving performance by adding last-word dependent image-based context vectors to the LSTM decoder, effectively allowing the network to "look" around the image as it forms the caption (Xu et al., 2015). More recent works such as "Neural Baby Talk" by Lu et al. improve image captioning through use of classified object labels using Faster-RCNN+ResNet (Lu et al., 2018). However, to our knowledge, all major works in this area rely on datasets of image-caption pairs with minimal research in cross-dataset tasks and training.

2.2. Visual Sentiment Analysis

Visual sentiment analysis has received relatively less attention that its NLP counterpart. With that being said, research in visual sentiment analysis has seen a noteworthy rise recently. Jindal et al. proposed deep CNNs with domain specific fine tuning to improve visual sentiment classification accuracy using 7 labels ranging from depressed to excited (Jindal & Singh, 2015). Borth et al. have published the Visual Sentiment Ontology (VSO), leveraging 1200 Adjective-Noun Pairs to create the SentiBank features involving more fine-grained features (Borth et al., 2013). VSO has been used extensively as part of our study and more details are provided in the later sections. Cao et al.

[*]Equal contribution [1]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada. Correspondence to: Peyman Bateni <peyman.bateni@alumni.ubc.ca>, Ignacio J. Iturralde Ch. <i.iturralde@alumni.ubc.com>.

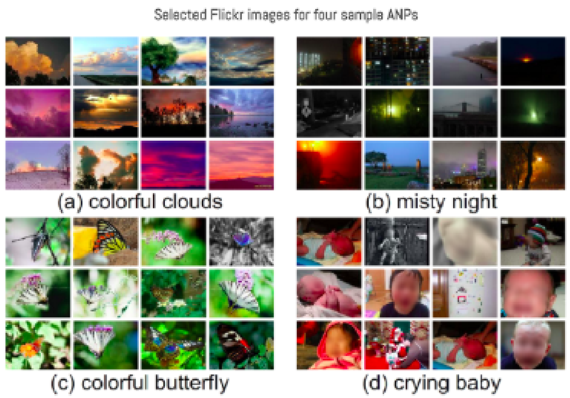

Figure 1. Sample Flickr images from VSO for demonstration.

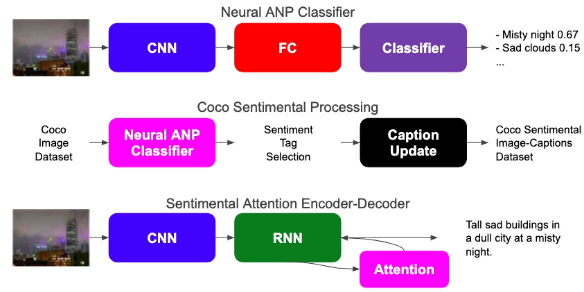

Figure 2. Overview of our data focused approach consisting of three steps in order shown: Step 1. training a Neural ANP Classifier, Step 2. evaluating ANP tags on COCO images and updating captions, and Step 3. training a new attention-based Autoencoder for sentimentalized visual captioning using the updated dataset.

showcase the Visual Sentiment Topic Model (VSTM) that uses VSO features on images of the same topic, enhancing tertiary sentiment classification (Cao et al., 2014). Wang et al. present unsupervised training in closing the semantic gap between low-level visual features in the image and the higher level over sentiment Wang2015. Yuan et. al propose Sentribute, using Mid-Level visual features to capture the visual sentiment built upon the lower level features. Domain transferred deep CNNs are evaluated and trained progressively by You, Lu et al. to the address the issue of unlabeled visual data from social media (Yuan et al., 2013). Pang et al. provide an extensive study into opinion mining and sentiment analysis as a whole outside of Computer Vision (Pang & Lee, 2008), and Machajdik et al. exploit theoretical and empirical finding in Psychology and Art Theory to combine low-level features for affective image classification (Machajdik & Hanbury, 2010). Lastly, Hu et al. use multimodal analysis to improve sentiment classification on social media posts, improving performance by considering both visual and textual sentiment (Hu & Flaxman, 2018).

3. Datasets

3.1. Microsoft COCO

Microsoft's Common Object in Common (Lin et al., 2014) dataset of images, labels, boundaries and captions is arguably the most widely used dataset in visual captioning. As of writing the dataset provides over 200K labelled images with five captions per image, 1.5M object instances with existing libraries of pre-trained models.

3.2. Columbia VSO

The Columbia Visual Sentiment Ontology (Borth et al., 2013) was created through collection and labelling of over 0.5M images from the social media platform, Flickr, using 1,200 Adjective-Noun Pairs. ANPs are used as more fine-grained sentimental tags and originally, 3244 tags were evaluated with 2044 being eliminated to limited frequency. The current dataset, as available through Columbia University, has 487,264 images clustered into 1,554 ANP tags. Please visit Figure 1 for examples of VSO images/tags.

4. Designs

In our study, we evaluated two major designs: one focused on producing ANP tags for COCO images, updating the captions and training an Autoencoder for visual captioning, and the other focused on attention-based visual sentiment retrieval with the goal of cross-training to obtain combined sentimentalized captions.

4.1. Data Focused Design

As described in Figure 2, we employed a three stage design consisting of two models, our Neural ANP Classifier and Captioning Autoencoder, and our update script used for effective sentimentalization of COCO captions.

4.1.1. Neural ANP Classifier

The Neural ANP Classifier uses a ResNet101 with modification on the final layer to extend to 1,553 outputs, one per ANP tag present in our training set. We experimented with various hyper-parameters including losses and pretrained weights. Details are provided later as part of our evaluation.

4.1.2. Caption Updates

Using our Neural ANP Classifier, we obtained top-20 ANP tags with respect to each COCO image, adding adjectives to the captions only if the corresponding noun is present and breaking ANP tags associated with the same noun using their respective Softmax probabilities. No synonyms were employed in updating captions.

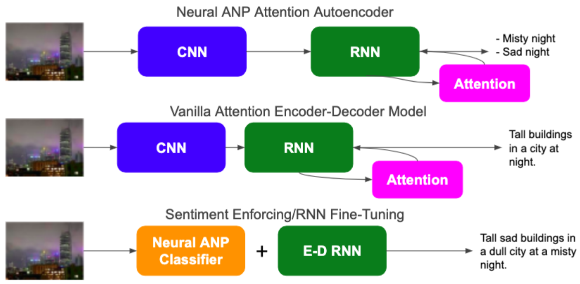

Figure 3. Overview of our model focused approach: Step 1. training a Neural ANP Attentional Autoencoder, Step 2. training a vanilla COCO Attentional Autonencoder for simple captioning, and Step 3. fine-tuning the COCO Attentional Autonencoder with VSO tags using Attention masks produced by both models.

4.1.3. Captioning Autoencoder

We adopted an attention-based Encoder-Decoder architecture from with modifications to use ResNet101 for consistency.

4.2. Model Focused Design

In our model focused design, we deviated from updating the dataset and instead focused on using attention to gain greater insights into visual sentiment analysis and explore cross-training between COCO and Columbia VSO on the same model. Similar to our first design, our model focused approach also consists of 3 steps as noted in Figure 3.

4.2.1. Neural ANP Attentional Autoeconder

Consistent to the Softmax classifier, we employ a fine-tuned ResNet101 for encoding images. A subsequent decoder consisting of an RNN and the attention unit are seeded with the start of the sentence token ([SOS]), producing the adjective, the noun and the end of sentence token ([EOS]) in order. The VSO tags are broken into arrays of one-hot encoded words for training.

4.2.2. Vanilla COCO Attentional Autoeconder

The vanilla COCO Attentional Autoeconder uses the same ResNet101, RNN, Attention architecture universally used in this paper. Captions come directly from the COCO dataset.

4.2.3. Fine-Tuning using Attention Masks

Our initial design for fine-tuning using attention masks, as outlined in Figure 5, relied on evaluating both the Neural ANP Attentional Autoencoder and the vanilla COCO Captioner on the images, using the attention masks as a metric for whether the adjective-mask must be deployed on a certain noun. This was prepared with the goal of using the Euclidean distance between the attention masks as a coefficient such that the backpropagated gradients are weighted based on whether the vanilla COCO Captioner is also looking at the same area.

Figure 4. An example of a sentimentalized image caption.

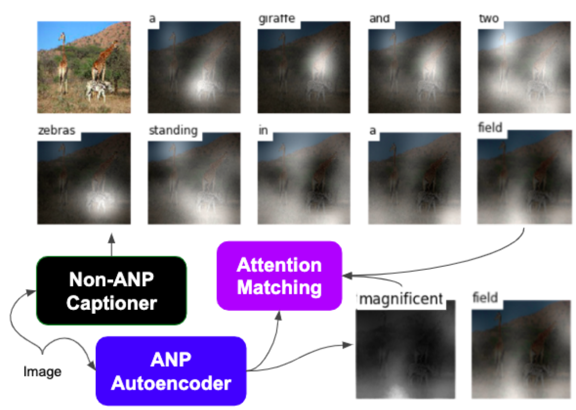

Figure 5. Initially anticipated approach to cross-training across COCO and VSO using attention matching to do weighted gradient backpropagation.

Consider the case in Figure 4. The intention behind our cross-training was to formulate gradient weights using attention masks as distance metrics to heavily enforce RNN gradients from the Neural ANP Attentional Autoencoder at places such as ”cake” and ”bears” where the attention masks are very similar but disregard ANP gradients on other nouns present such as ”table”.

With that being said, our findings on sentimentalized attention regions refuted our initial design, instead pivoting to noun forcing where evaluated results from both models are combined during inference based on the noun produced by the ANP Autoeconder, disregarding the attention masks all together. More expansion on our findings, noun-forcing and

Loss	Initialization	Epochs	Validation Accuracy			
			Top1	Top5	Top10	Top20
MultiLabelSoftMargin	Basic Initialization	10	<0.01	<0.01	0.015	0.027
NLLLoss	Basic Initialization	10	0.03	0.106	0.145	0.212
NLLLoss	Basic Initialization	20	0.022	0.07	0.109	0.165
NLLLoss	Basic Initialization	30	0.022	0.068	0.106	0.161
NLLLoss	Basic Initialization	40	0.024	0.072	0.111	0.164
NLLLoss	Pre-trained ImageNet Weights	10	**0.082**	**0.195**	**0.265**	**0.351**
NLLLoss	Pre-trained ImageNet Weights	20	0.072	0.174	0.241	0.319

Table 1. Classification accuracies for the Softmax ANP Classifier with respect to various hyperparamater settings used.

additional details are discussed as part of evaluation.

5. Evaluation

We evaluated our two designs independently, experimenting with various hyper-parameters and design choices, and selecting the most optimal settings at each step to be used in the subsequent steps. The VSO dataset was randomly sampled into 70% (336,630 images) for training, 15% (72,352 images) for validation and 15% (72,340 images) for testing. The sampling was completed after the elimination of the 5,942 small/corrupted images to maintain consistent distributions across all three sets. Note that of the 1,554 labels available, images for ”Outdoor Training” were removed totalling 1,553 labels as after sampling the three sets, no image of such label was present in the train set.

As for the COCO dataset, unfortunately the provided test set of 40,670 images does not come with captions and is thus not useful as a test set for the task at hand. Additionally, after completing caption updates, details of which are provided later in the paper, there was concerns regarding the inconsistent distribution of the COCO provided validation set compared to the train set. Combined with the fact that the provided validation set only consists of 5,000 images, a relatively small validation set compared to the 118,287 images in the train, there was motivation to form a separate validation set internally. Thus, 20,000 train images were randomly sampled to form the ”internal” validation set with 98,287 remaining images used for training. All training was completed on AWS p3.2 instances, totalling 720 GPU hours.

5.1. Design 1 - Neural ANP Classifier

In our first model, we used Multi-Label Softmargin Loss to avoid normalizing probabilities with the goal of allowing the classifier to predict more labels independently when we evaluate on COCO images. However, as demonstrated in Table 1, this was unsuccessful with less than 1% single label classification accuracy which only extends to 2.7% on Top20 accuracy, often leading to the same set of few labels being predicted for all images.

Figure 6.
Left) Actual Label: fresh water, Predicated Top5 Labels: natural spring (0.628), healing water (0.242), holy water (0.050), warm water (0.0261), natural wonder (0.019),
Right) Actual Label: cold drink, Predicted Top10 Labels: favorite city (0.557), expensive hotel (0.226), stunning building (0.158), famous hotel (0.036), heavy clouds (0.004).

Switching to NLLLoss with a Log Softmax final layer for classification, improved performance to 3% single-label accuracy and 21.2% top20 accuracy. Further improvements were obtained by initializing the model with pre-trained ResNet101 ImageNet weights available through PyTorch, boosting single-label accuracy to 8.2% and Top20 classification to 35.1% accuracy which would be more reasonable for updating captions.

Additionally, it was found that the model is relatively quick to converge with learning rate set to 1×10^-4 with the tendency to overfit with more epochs of training as both the vanilla and the pre-trained initializations performed best at 10 epochs. The learning rate was selected after 1 epoch experimentation with convergence using learning rates ranging between 1×10^-3 and 1×10^-5.

As shown in Figure 6, it was found that softmax classification using ANP tags produces tags that are often reasonable but not the same as the actual tag. Furthermore, the classifier has a tendency to focus more on correctly classifying the noun as oppose to the adjective as with the left picture.

In addition, through examining of 100 randomly sampled validation images, it was observed that in certain cases,

Set	Train	Internal Validation	COCO Validation
Number of Images	98,287	20,000	5000
Number of Updates	249,454	50,525	0

Table 2. Caption update statistic across different sets. Note that all sets come from the COCO dataset. The internal validation set is a separate set manually formed from the COCO training where as the COCO validation was directly provided by the dataset itself.

Model	Loss	BLEU2 Score	Validation Accuracy			
			Top1	Top5	Top10	Top20
Softmax ANP Classifier	NLLLoss	N/A	0.082	0.195	0.265	0.351
ANP Attentional Autoencoder	Cross Entropy	**0.357**	**0.512**	**0.701**	**0.779**	**0.844**

Table 3. Classification accuracies for the ANP Attentional Autoencoder compared to the best performing Softmax ANP Classifier.

Figure 7. Example of a caption update for a picture where the ANP tag, "travelling birds" was found in the Top20 tags after evaluation using the Softmax ANP Classifier.

the classified tags were objectively more relevant than the provided tag. For instance, the right image in Figure 6 is labelled "cold drink" by the dataset while predicted labels such as "expensive hotel" or "stunning building" are more appropriate.

Considering this fact, the final model was trained with NLLLoss and pre-trained ImageNet ResNet101 weights for 10 epochs. The Softmax ANP Classifier was evaluated on the test set and achieved 8.1%, 19.3%, 26.4% and 35.1% on each of the TopX accuracy metrics respectively and was selected for further use in the next steps.

5.2. Design 1 - Caption Updates

Using the selected ANP Classifier, caption updates were performed using Top20 evaluated tags on COCO images as outlined previously (see Figure 7 for example). As noted in Table 2, total of 299,979 updates were applied in form of adjective additions to nouns with a split roughly proportional to the number of images in the train and the internal validation set.

The external validation set provided by COCO received 0 updates, and after examining 100 randomly samples images and their tags and captions, there seems to be no algorithmic problem. The COCO validation set's caption are more conceptual which may draw questions on how good of a validation set it would be for the task at hand. Therefore, the internal validation set was chosen to fulfill the role of validation in our study.

5.3. Design 1 - Attentional Autoeoncder for Visual Captioning using Updated Captions

An attention-based Autoencoder adopted implementation was trained using the new captions. Earlier experiments using fewer caption updates resulted in zero-adjective captioning. However, the 300% increase in caption updates using the reported approach, improved results.

100 randomly sampled validation images and results from the final model were evaluated qualitatively. As shown by the images on the left and the right in Figure 8, the network is able to identify ANPs effectively, especially with respect to colour. However, often an incorrect adjective classification, as in the case of "clean car" in Figure 8, leads to subsequent miscaptioning arising from the incorrect adjective input.

Zero-adjective examples still persisted, but only accounting for roughly half of the captions, an anticipated figure given that we applied 299,979 total updates to 591,435 captions, with an average of 0.507 updates per caption.

It was also interesting to see that often ANPs were achieved that were otherwise rarely seen in the training data. An example would be the image on the right in Figure 8, with the network focusing on a colour-based adjective for the dog. Overall, the captioning Attentional Autoencoder achieved a BLEU2 score of 0.0794.

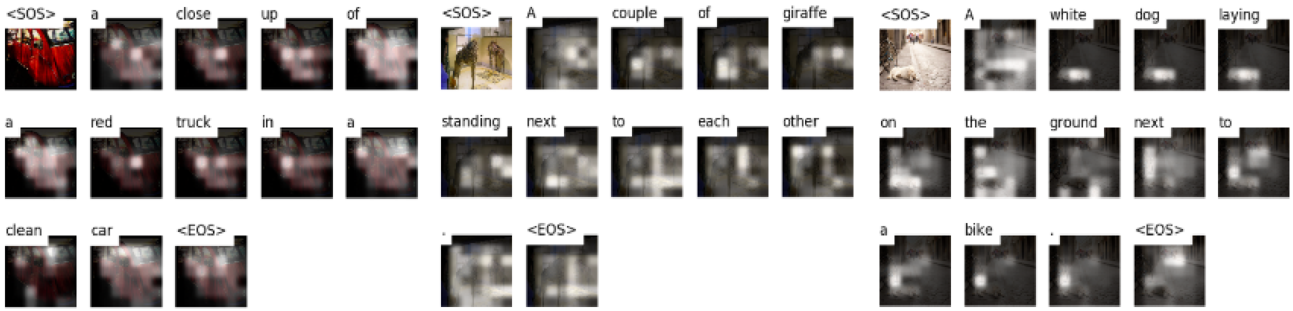

Figure 8. left) example of adjective captioning successfully at ”red car” but unsuccessfully at ”clean car”, center) zero adjective captioning, right) unseen adjective pair association used in the caption.

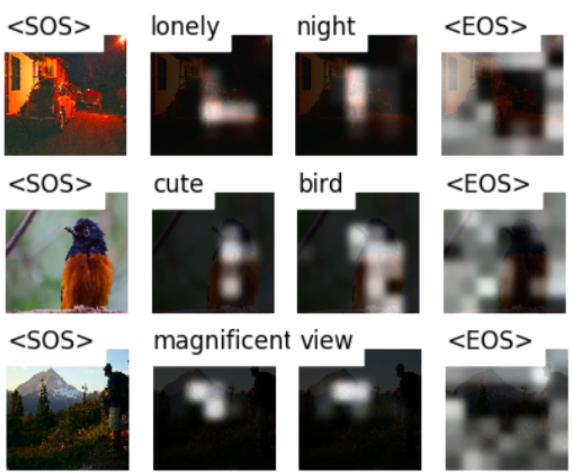

Figure 9. Attention masks from evaluated validation images using the ANP Attentional Autoencoder.

5.4. Design 2 - Neural ANP Attentional Autoencoder

Our findings through attention masks obtained using updated caption further motivated the study of attention in classifying ANPs. The Neural ANP Attentional Autoencoder was trained and evaluated using the same VSO training and validation sets. We used an adoptive learning rate, reducing every time 2 epochs of training resulted in no improvements and trained for 20 epochsin total. The architecture was adopted from the same attention-based Autoeconder implementation universally used across this paper.

As shown in Table 3, the Neural ANP Attentional Autoencoder achieved between $2.5\times - 6\times$ improvements across the TopX classification metrics, resulting in a 51.2% single-label classification and an 84.4% Top20 accuracy.

Qualitative evaluated on sampled validation images resulted in important findings with respect to ANP classification using attention. As shown in Figure 9, the attention marks are often looking at places very different to that of the actual noun. For instance, with respect to the first image, the network uses cues from the seen when determining the ”lonely” sentiment, at times areas completely different from what it uses to classifying the ”night” itself. Same holds with respect to ”magnificent view” in Figure 9.

On the other hand, with less conceptual sentiments focused on more identifiable nouns such the ”bird”, the attention mask focuses on more fine-grained details to identify the adjective ”cute”, focusing on a smaller span of the image compared to when it's evaluating the noun, in this case looking at the ”bird” itself.

It's important to note that the Attentional Autoencoder is tasked with determining the ANP in a sequential manner, first predicting the adjective followed then by the noun. This dynamic way of evaluation allows the network to better predict the adjective, where as the Softmax ANP Classifier often has trouble accurately predicting both the adjective and the noun at the same time. This finding extends further by suggesting more heuristic visual definition for adjectives, often disjoint from the noun subjected, although additional future experiments are needed to explore this idea more extensively.

5.5. Design 2 - Cross Model Sentimentalized Captioning with Noun-Forcing

Our findings with respect to attention-based ANP classification resulted in refuting our cross-training approach with respect to attention. While we were expecting different attention masks used for adjective and noun classification, the contract was found to be much more expansive, often focusing on entirely different areas or very fine-grained levels of detail. This, thus, resulted in the use of any adjective-first attention matching to be problematic as most vector-based distance metrics would not allow adequate matching of the masks. Use of noun attention masks would not allow gra-

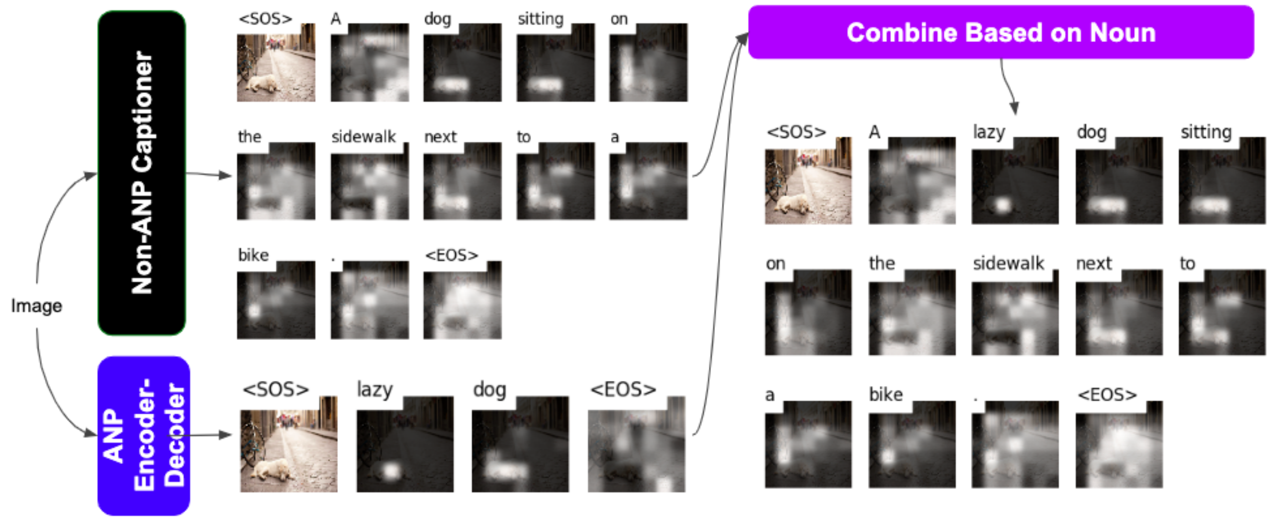

Figure 10. Overview of noun-forced combining of results. Note that the attention masks are for demonstration purposes only and the combining happens solely based on nouns.

dient propagation with respect to the adjectives due to the ordering of the prediction. However, modifications to use noun-first approaches show potential exploration in the future.

As a result, our alternative solution using noun-forcing was evaluated instead. As shown in Figure 10, the noun-forcing approach uses a non-ANP Attentional Autoencoder trained on original COCO captions to produce a description of the input image. In parallel, the input image is evaluated using the trained ANP Attentional Autoencoder, where the results are afterward combined based on the nouns produced by each model.

This approach tackles the zero-adjective captioning issue through more explicit enforcing of the use of ANPs. However, as shown in Figure 11, there are limitations in noun-forcing where the model may mistakenly treat a "train track" is a similar way to a "train". This motivates future work to tackle this issue using attention masks, although our models only use non-attention noun-forcing.

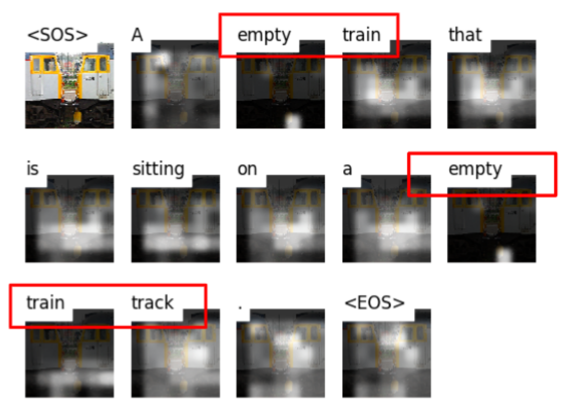

Figure 11. An example of where noun-forcing fails, and where attention masks could have potentially helped.

6. Conclusion

We propose two sets of models for sentimentalized image captioning, each employing visual sentiment classification using ANPs and Attentional Autoencoders for captioning. Our first set uses a Softmax ANP Classifier with optimal hyperparameters, caption updates and a classical Attentional Autoencoder for captioning. Our second model uses an Attentional ANP Autoencoder to boost ANP classification accuracy by $2.5\times - 6\times$ using attention masks and dynamic classification of adjectives and nouns separately. It then uses noun-forcing to combine results from the ANP Autoencoder and a vanilla COCO Attention-Based Encoder-Decoder to produce sentimentalized captions. Our work also demonstrates the importance of attention in visual sentiment analysis with findings in relative attention masks for sentiment with respect to their associated nouns.

In the future, we hope to extend this work by exploring Noun-Adjective Pair (NAP) Autoencoders using noun-first classification motivated by achieving more fine-grained networks where a noun-seed could be used for identifying specific adjectives throughout the photo as oppose to more overall single-label classifications. In addition, we hope to use noun-first NAP classification to successfully perform attention matching, using what we hope to be similar attention masks on the nouns this time with forward post-noun adjective predictions that can propagate gradients. Other future extensions including exploring ensemble models and deeper neural networks may also hold great potential in producing sentimentalized visual captioning models.

7. Source Code

To recreate our experiments please visit our Github Repository. The COCO dataset is available publicly and the VSO dataset can be obtained via direct request through the Columbia University.

8. Acknowledgements

The source code from which we adopted our Attention-based Encoder-Decoder models can be found here and we would like to thank the authors for making it publicly available. The AWS instances for training were provided by Beam Engage. We would like to thank Professor Shih-Fu Chang for providing us access to Columbia VSO, and the Machine Learning community at UBC for their support.

References

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pp. 223–232, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502282. URL http://doi.acm.org/10.1145/2502081.2502282.

Cao, D., Ji, R., Lin, D., and Li, S. Visual sentiment topic model based microblog image sentiment analysis. *Multimedia Tools and Applications*, 11 2014. doi: 10.1007/s11042-014-2337-z.

Hu, A. and Flaxman, S. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 350–358, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3219853. URL http://doi.acm.org/10.1145/3219819.3219853.

Jindal, S. and Singh, S. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *2015 International Conference on Information Processing (ICIP)*, pp. 447–451, Dec 2015. doi: 10.1109/INFOP.2015.7489424.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Lu, J., Yang, J., Batra, D., and Parikh, D. Neural baby talk. *CoRR*, abs/1803.09845, 2018. URL http://arxiv.org/abs/1803.09845.

Machajdik, J. and Hanbury, A. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 83–92, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1873965. URL http://doi.acm.org/10.1145/1873951.1873965.

Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL http://dx.doi.org/10.1561/1500000011.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. URL http://arxiv.org/abs/1411.4555.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL http://arxiv.org/abs/1502.03044.

Yuan, J., Mcdonough, S., You, Q., and Luo, J. Sentribute: Image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pp. 10:1–10:8, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2332-1. doi: 10.1145/2502069.2502079. URL http://doi.acm.org/10.1145/2502069.2502079.