Towards Image Caption Generation for Art Historical Data

Jahnvi Gupta[1], Prathmesh Madhu[2], Ronak Kosti[2], Peter Bell[3], Andreas Maier[2], and Vincent Christlein[2]

[1] National Institute of Technology, Hamirpur, India
[2] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg,
[3] Institute for Art History, Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract. One of the first steps to understanding images is by reading their captions. In the field of computer vision and natural language processing, generating captions from images has been an important and a challenging problem. In this work, we aim at understanding the pioneering work in neural caption generation, Show-and-Tell [4], with respect to art historical data. Artworks are characterized by various artistic styles, attributes, and motives; along with great diversity of artists creating these artworks and the different periods in history when these were created. This makes it very challenging to build models that are agnostic to all such variations.

In this work, we propose an art history based image captioning dataset of 4000 images across 9 iconographies (*annunciation, adoration, baptism, still-life, nativity, virgin and child, rape, tower of babel and noli me tangere*) along with a description for each image consisting of one or more paragraphs. Inspired by [4], we fine-tune the captioning models on our new data. We conduct a series of ablation studies with various embedding sizes and state-of-the-art (SOTA) architectures as backbone like ResNet, ResNext and EfficientNet along with the proposed InceptionNet [4]. We use these SOTA image classification models to evaluate their ability of generating features for captioning. For quantitative evaluation and comparison, we use the standard 4 BLEU metrics [2]. This value indicates how similar the candidate text is to the reference texts. Higher BLEU scores represent higher similarity between predicted and reference text. In addition to the quantitative comparison, a qualitative discussion on the predicted captions is also presented.

Results & Observations

1. **Table 1**: Based on Flickr8k dataset, we observe that the performance decreases with increase in embedding sizes for both the architectures, via the BLEU scores for the last (30^{th}) epoch of training. Empirically, we observe that an embedding size of 512 gives the best result. The performances of Inception and ResNext101 are comparable, and as such either can be chosen for the task.

Table 1. Comparison of different embedding sizes across networks

	Model	**Inception**			**ResNext101**		
Dataset	**Embedding Size**	512	1024	2048	512	1024	2048
Flickr8k	BLEU1	**58.76**	57.76	53.80	**58.56**	58.51	54.82
	BLEU2	**39.22**	38.32	34.56	**39.22**	39.15	35.68
	BLEU3	**25.29**	24.72	21.60	**25.36**	25.22	22.48
	BLEU4	**16.92**	16.83	14.08	**16.77**	16.59	14.71

Table 2. Comparison of results for across models with fixed embedding size 512

	Model — InceptionV3	**— ResNext101**	**— EfficientNetB0**
Flickr8k	BLEU1	59.57	59.24	**60.03**
	BLEU2	40.00	40.26	**40.98**
	BLEU3	25.94	26.51	**27.08**
	BLEU4	17.36	17.96	**18.20**
Flickr30k	BLEU1	57.07	57.53	**57.99**
	BLEU2	37.00	**37.66**	37.33
	BLEU3	23.23	**23.94**	23.64
	BLEU4	15.58	**16.26**	16.06
Art History	BLEU1	20.69	20.12	**22.47**
	BLEU2	10.72	10.42	**12.02**
	BLEU3	06.56	06.46	**07.95**
	BLEU4	04.94	04.78	**06.27**

2. **Table 2**: *EfficientNet*, trained for classification, also captures good features for image captioning outperforming all other models, across all BLEU scores.
3. Although the BLEU scores for our Art Historical datasets are low, the qualitative results are motivating enough to continue working towards more refined approaches [3][1].
4. **Qualitative Results**

   Figure 1: We can see in the figure that there is diversity in the descriptions predicted by both models where each description can be considered to be quite appropriate.

   Figure 2: EfficientNet is able to correctly describe the given image whereas Inception successfully recognizes the sculpture in the image but incorrectly considers it to be a sculpture of the virgin.

Fig. 1. Still-life iconography example image from test set. **Original caption** : This still-life of plums and apples, a blue and white bowl of Strawberries, grapes, gooseberries, a Roemer of wine anda pewter vessel on a table. **Show-and-Tell** : this picture shows a still life with a roemer a silver glit columbine cup a jar a porcelain dish with an orange and peaches. **EfficientNet** : *the painting represents a still life with peaches plums and grapes in a blue and white wan li porcelain dish with an orange and peaches.*

Fig. 2. Rape iconography example image from validation set. **Original caption** : The Rape of the Sabines marked the climax of Giambologna's career as an official Medici sculptor. This great marble was unveiled in the Loggia dei Lanzi in January 1583 in place of Donatello's Judith. **Show-and-Tell** : the virgin is clothed in a drapery drawn out in very narrow folds with sharp ridges the sculpture comes as an exception from a time. **EfficientNet** : *the rape of the sabines marked the climax of giambologna's career as an official medici sculptor this great marble was unveiled in the loggia.*

Conclusion

We curated a novel captioning dataset for art historical images consisting of 9 iconographies. We fine-tuned the SOTA models on these datasets using [4] as our base pipeline. These models are able to generate decent descriptions also in the case of art history images. The BLEU scores for the art history dataset are low in comparison to the Flickr images. Considering the art historical dataset used, one of the future works, apart from using modern approaches ([3][1]), would be to clean the data and to see the performance of the model on different art history datasets with shorter captions. Another interesting further work would be the use of the generated captions for retrieval and being inclusive for vision impaired community with respect to the context and contents of artworks.

Keywords: neural captioning · art history · understanding iconographies

References

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
2. 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
4. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.