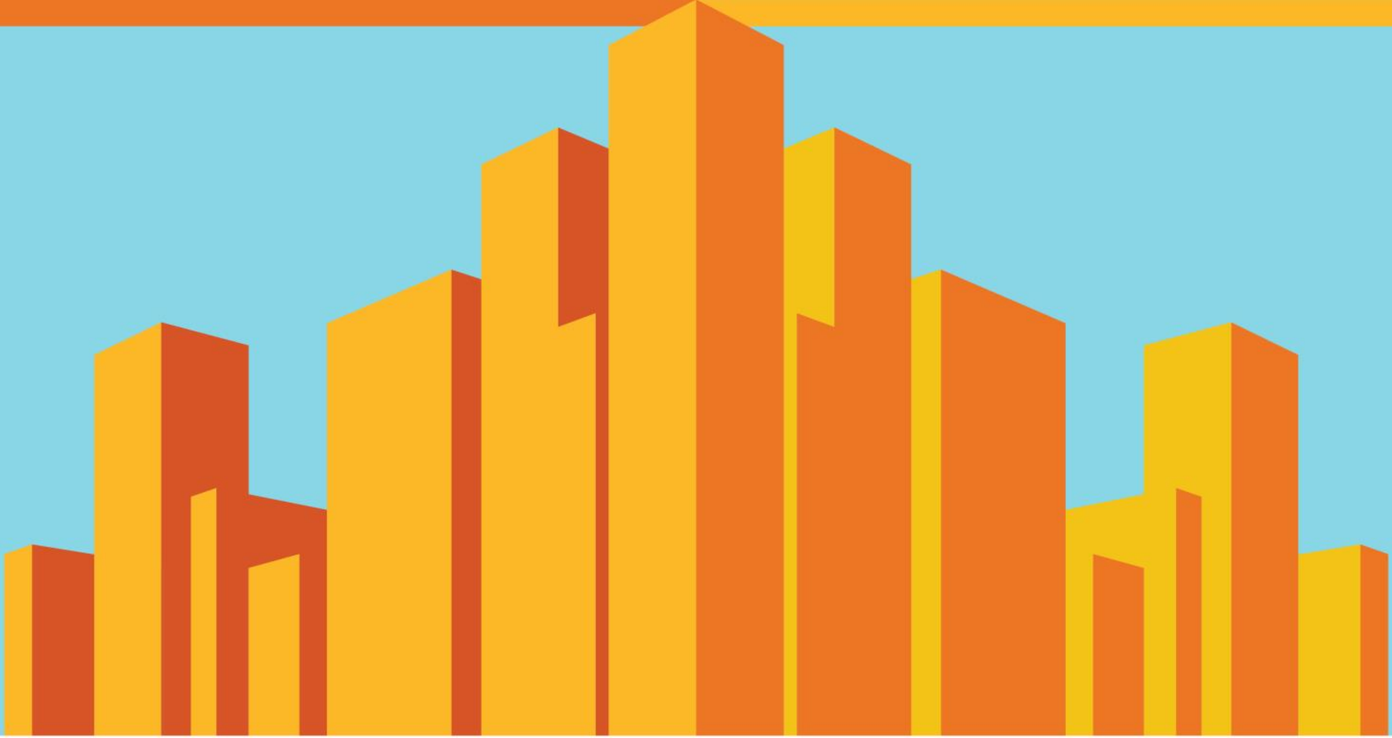


informs ANNUAL MEETING

2021 ANAHEIM, CALIFORNIA

informs ANNUAL MEETING
2021 ANAHEIM, CALIFORNIA

Artionyms and Machine Learning: Auto naming of the paintings

Anna Altynova[1], Alexander Semenov[1,2], Dmitry Grigoriev[1], Valeria Kolycheva[1];
[1]CEBA lab, Saint-Petersburg State University, Russian Federation, [2]University of Florida and St. Petersburg State University, USA.

Table of contents

Introduction

- Still little published data about captions generation for artistic paintings
- We use a deep neural network with Attention mechanism
- We evaluate the model using image captioning metrics and discuss its capacity to generate art-related names.

Backgroung work

There wasn't many studies contributed to the task of generating descriptions for artworks. Here is the closest to ours task:

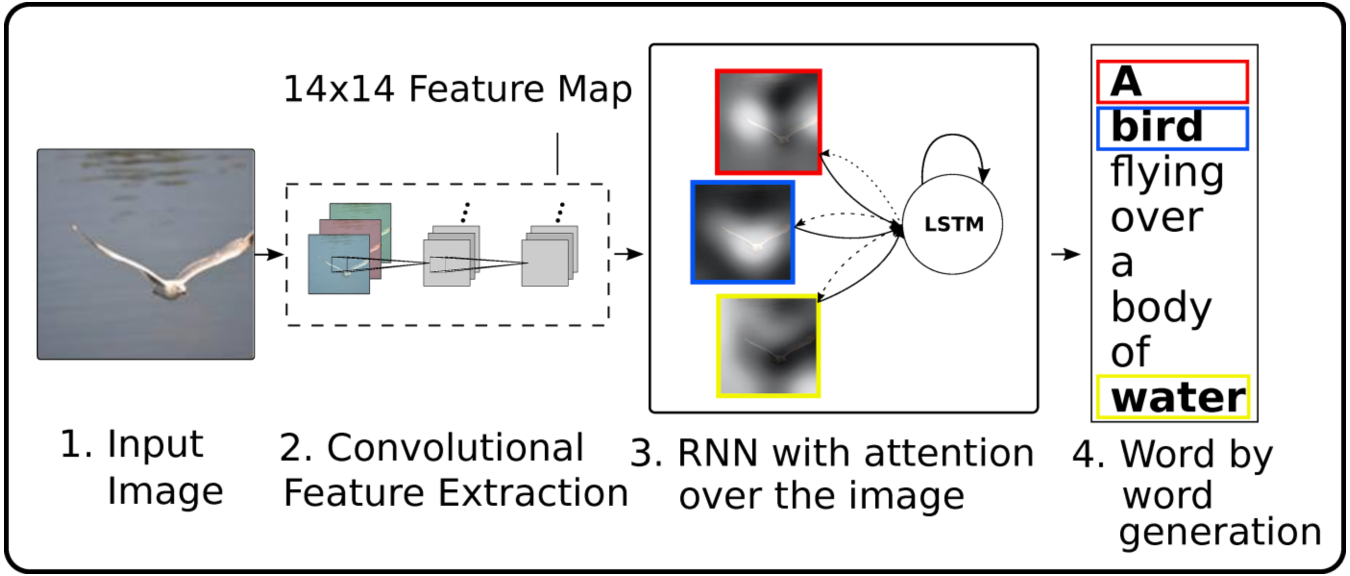
- Generating Captions for Images of Ancient Artworks[1]:
    - datasets of Ancient Egipt and Ancient Chinese artworks
    - encoder-decoder architecture for caption generation.
    - little classic paintings observation
- Iconographic Image Captioning for Artworks[2]:
    - Iconclass dataset
    - pre-trained Vision-Language Model
    - specific art style

Dataset

The Wikiart paintings dataset[3]:

- more than 80,000 fine-art paintings
- more than 1,000 artists
- fifteen century to modern times
- 27 different styles
- 45 different genres

Neural network



"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", 2016[4]

Neural network

- Encoder: pretrained InceptionResNetV2 [5]
- Decoder: controlled recurrent blocks GRU [6].

Metrics

$$\text{BLEU} = \underbrace{\text{BP}}_{\text{Brevity Penalty}} \cdot \exp\Big(\sum_{n=1}^{\underbrace{N}_{\text{4 (considers only 1 to 4-gram precisions)}}} \frac{1}{N} \log(\underbrace{p_n}_{\text{Modified n-gram precision}})\Big) \quad \boxed{\in [0,1]}$$

1. BLEU or the Bilingual Evaluation Understudy Score[7]

Metrics

$$\text{ROUGE - n} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}_{\text{match}}(\text{n - gram})}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}(\text{n - gram})}$$

2. ROUGE, or Recall-Oriented Understudy for Gisting Evaluation[8]

Metrics

$$Fmean = \frac{10PR}{R + 9P} \qquad Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)$$

$$Score = Fmean * (1 - Penalty)$$

3. METEOR [9]

Metrics

$$\mathrm{CIDEr}_n(\underbrace{c_i}_{\text{candidate sentence}}, \underbrace{S_i}_{\text{set of reference sentences}}) = \frac{1}{m}\underbrace{\sum_j}_{\text{average over references}} \underbrace{\frac{\overbrace{\boldsymbol{g}^{\boldsymbol{n}}}^{\text{TF-IDF vector (n-gram)}}(c_i)\cdot \boldsymbol{g}^{\boldsymbol{n}}(\overbrace{s_{ij}}^{\text{j-th reference}})}{\|\boldsymbol{g}^{\boldsymbol{n}}(c_i)\|\|\boldsymbol{g}^{\boldsymbol{n}}(s_{ij})\|}}_{\text{cosine similarity}}$$

$$\mathrm{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \mathrm{CIDEr}_n(c_i, S_i)$$

4. CIDEr [10], proposed in 2015 specifically for evaluating image captions.

Implementation & results

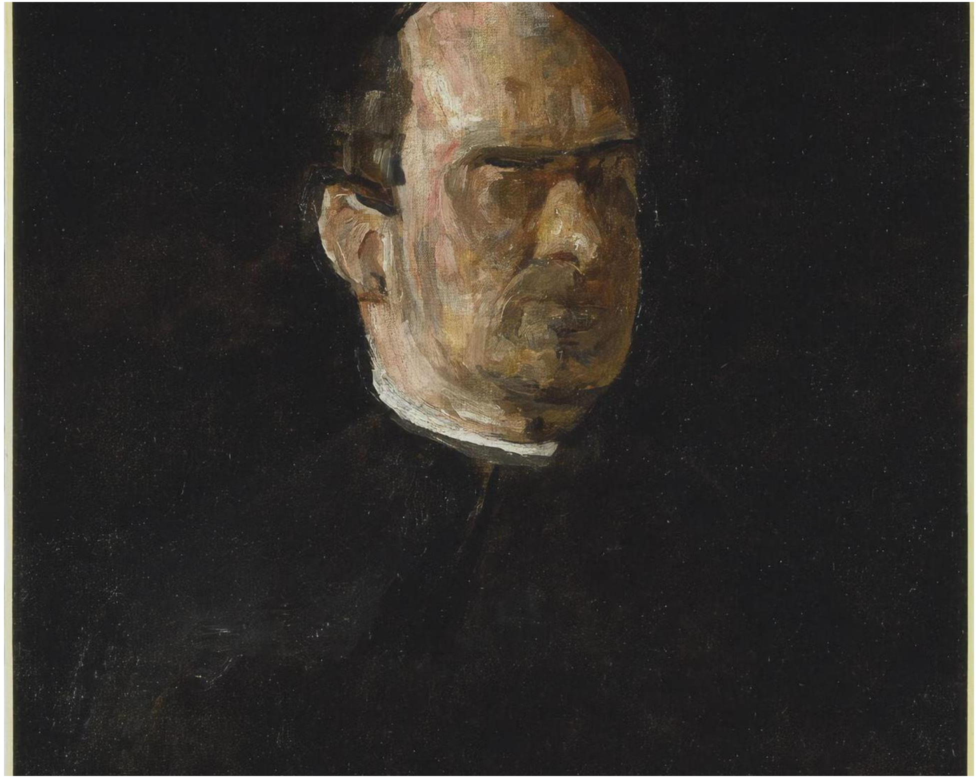
- The network was trained with Adam optimizer with the learning rate upper limit = 3e-7 during the main part of the training and the transition to the cyclical learning rate [11] on the plateau. Batch size = 32, the number of units= 512.
- The best quality of the generated names (according to the subjective human assessment on a test sample of 50 images) was obtained as a result of training the model for 35 epochs.

Metric	Test evaluation
Bleu-1	5.2
Bleu-2	1.5
Bleu-3	0.4
Bleu-4	19e-6
Rouge-L	4.2
Cider	1.9
Meteor	2.6

Examples

Original caption: forest horizont

Predicted caption: springtime



Original caption: portrait of dr edward anthony spitzka

Predicted caption: portrait of the surrender

Original caption: the beach at trouville at low tide

Predicted caption: the fields of the thames

Original caption: bridge astrakhan

Predicted caption: sad woman harbor the temple

Metrics discussion

- As it turns out, standard metrics and especially the BLEU score do not always accurately reflect the quality of the name obtained.
- For instance, generated name "Village soldier" for Vincent van Gogh's "Orphans" seems fine, but its bleu score is zero.

Metrics discussion

The same was noted in Iconographic Image Captioning for Artworks, where metrics were used to assess the quality of captions for iconographic images.

Since no measurements were previously performed on the WikiArt dataset, let us compare the results with those on the MS-COCO dataset[12] (performed on a model with a similar architecture), and on the Iconclass dataset.

Metrics discussion

Metric	MS-COCO	Iconclass
Bleu-1	73.1	12.8
Bleu-2	56.2	12.2
Bleu-3	41	11.3
Bleu-4	32.6	10.0
Rouge-L	-	31.9
Cider	87.2	172.1
Meteor	26.1	11.7

- Thus, when our results are compared with MS-COCO with the bleu-4 metric, the gap is the largest. The difference with bleu-1 ~14 times, with Meteor ~ 10, with Cider ~ 46 times.
- To understand whether such a gap is caused by a feature of the metrics, dataset, or low quality signatures, an expert assessment of the quality of the generated names is required and its subsequent comparison with machine assessment.

Literature

[1] Sheng, S., Moens, M.F.: Generating captions for images of ancient artworks. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2478– 2486 (2019)

[2] Cetinic E (2021) Iconographic image captioning for artworks. In: Del Bimbo A et al (eds) Pattern recognition. ICPR international workshops and challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12663. Springer, Cham

[3] https://www.wikiart.org.

[4] *Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio*: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:2048-2057, 2015.

[5] Christian Szegedy, S. Ioffe, Alexander Amir Alemi: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

[6] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation" InL EMNLP 2014

[7] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation*. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318

[8] Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[9] Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005*

[10] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh: CIDEr: Consensus-based Image Description Evaluation. In:CVPR 2015

[11] Leslie N. Smith: Cyclical Learning Rates for Training Neural Networks. In:WACV 2017

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár: Microsoft COCO: Common Objects in Context. In: ECCV 2014: Computer Vision – ECCV 2014 pp 740-755