# Artionyms and machine learning: auto naming of the paintings

**Anna Altynova[1], Valeria Kolycheva[1], Dmitry Grigoriev[1], Alexander Semenov[2]**
[1]Center of Econometrics and Business Analytics, Saint Petersburg State University, Saint Petersburg, Russia
[2]Herbert Wertheim College of Engineering, University of Florida, Gainesville, United States

**Article Info**

***Article history:***

Received Feb 1, 2024
Revised May 10, 2024
Accepted Jun 1, 2024

***Keywords:***

Image captioning
Machine learning in art
Text generation
Transformers
Visual art

**ABSTRACT**

Image captioning is a question of great interest in a wide range of applications. In the art market there is a particularly acute shortage of specialized machine learning methods for accelerated and at the same time in-depth study of often too specific aspects of art. One of the main difficulties is caused by ambiguous names of art works, as well as clarifying (in practice, often complicating understanding and perception) signatures of the authors to them. Although previous research has established that captioning of photos can be done with high efficacy, there is little published data about generation of captions for artistic paintings. In this research, we utilize a transformer architecture to generate an artionym for a given painting in author's manner. We describe the model and report its performance on different art styles. We assess the model performance with an expert evaluation and image captioning metrics, and then discuss their capacity to analyze art-related names.

***Corresponding Author:***

Dmitry Grigoriev
Center of Econometrics and Business Analytics, Saint Petersburg State University
7/9 Universitetskaya nab., St. Petersburg, Russia
Email: d.a.grigoriev@spbu.ru

## 1. INTRODUCTION

The act of naming, an ancient and fundamental facet of human creativity, holds a pivotal role in the domain of fine art. The term ”artionym” originates from classical Greek words ”art” plus ”onym” (name) and thus refers to the title of an artwork, that often forms the viewer's initial engagement with the piece. This paper delves into the aesthetic function of artionyms and their correlation with the artistic value of the works they denote.

Artionyms can act as triggers for the viewer's introspective analysis of a work, amplifying the artist's concept and becoming an essential component of the art object. They may even precede the art object itself, as seen in the creation of a portrait. The choice of an artionym is a conscious process, meticulously aligned with the overall ethos of the work. This paper posits that a work's imagery extends beyond the art object to its title, which, through diverse linguistic methods, draws the viewer nearer to the authentic author's concept and the emotion embedded in the work.

This paper also scrutinizes the subjective motivation of artionyms, which is often the most unforgettable element for viewers and is regularly employed in successful advertising strategies. Instances from the works of Magritte [1] serve to exemplify this point, showcasing how artionyms can challenge the visible perception of reality and sway viewers. The aesthetic, economic, and social implications of artionyms are significant in the field of fine art. The previous studies highlight the complex interplay between the visual characteristics of an artwork [2], its perceived and economic value [3], and the influence of expert and public opinions on its pricing [4]. This could shed light on how artionyms, as an integral part of the artwork, might interact with these factors and influence the viewer's perception and valuation of the artwork.

Traditional machine learning methods struggle to swiftly and confidently incorporate artistic elements, such as ambiguous painting titles and author's notations that contribute to the sophistication of art. The latest advancements at the crossroads of computer vision and natural language processing (NLP) can invigorate the study of art. We employ the transformer model [5] that receives image features from a pre-trained convolutional neural network (CNN) Resnet101. The transformer model is trained from scratch, yielding a desirable word distribution typical for artionyms. Our experiments utilize the Wikiart paintings dataset, comprising over 80,000 fine art paintings across 27 different genres, making the WikiArt dataset one of the largest currently available collections of paintings. For the evaluation of generated captions, we adopt metrics used in the field of machine translation, such as bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), metric for evaluation of translation with explicit ordering (METEOR), and consensus-based image description evaluation (CIDEr).

The ability of a metric to estimate a result can be evaluated by comparing it with a human grade of that result. We conducted an expert evaluation on a random sample of size 50 from the test set to estimate the ability of metrics to evaluate generated artionyms. Our experiments revealed that common image captioning metrics are not suitable for evaluating generated artionyms due to their intolerance to synonyms and symbolism. Our main contributions are as follows:

- We adapt a leading photographic captioning approach to artwork captioning.
- We construct a Transformer-based artionym generator capable of forming an adequate artwork caption across different style groups.
- We conduct an in-depth analysis of generated captions via expert evaluation.
- We demonstrate that common automatic metrics show little correlation with human judgement on the artionym captioning task.

## 2. RELATED WORK

In the recent years, image captioning community has been consecutively reporting progress on such benchmarks as Microsoft common objects in context (MS COCO) [6] and Flickr [7]. These datasets contain various photographs of many real-world objects and five direct inscriptions for each picture. The task of automatic caption generation for such images is usually addressed by training an end-to-end model with an encoder-decoder architecture. Most encoders in such models are convolutional neural networks pretrained on the ImageNet dataset [8], such as ResNet [9], which extract visual features from an input image. These features are then converted to a vector of a fixed length that corresponds to the shape of the embedding matrix of the language model decoder that generates text according to the features. This structure was inspired by encoder-decoder models used in machine translation, mainly composed of recurrent networks on both ends [10]. Image captioning approach was presented in [11], where a long short-term memory (LSTM) network was used as the decoder, and pretrained GoogleNet [12] as the encoder. Adjusted with the attention mechanism, as in show, attend, and tell [13], this structure represents more powerful models, as attention allows to exploit of the most significant features of a caption. After the transformer model was presented in attention is all you need [5], this architecture was expeditiously applied to the image captioning task [14], then progressed in other task-specific alternatives [15]. Transformers got rid of recurrent neural network (RNN) and their drawbacks, such as forgetfulness and slow training, accomplishing improved convergence and speed.

In image captioning, researchers usually evaluate their results with common metrics, such as BLEU [16] score, CIDEr [17], ROUGE [18], and METEOR [19]. These metrics mainly deal with n-grams comparison between generated and original caption and less frequently pay attention to synonyms and paraphrasing. So finding an appropriate metric for artwork captioning is another challenge.

There is little published data on image captioning of artworks. Research by Castellano and Vessio [20], it was noted that it is more common to address other tasks within the art domain than image captioning, such as attribute prediction and object detection. However, there is still a number of studies that deal with art names. According to Sheng and Moens [21], authors caption cultural images from Chinese art and Egyptian art image captioning datasets with an artwork type enriched encoder-decoder model. Their encoder extracts not only an input image feature vector, but also an artwork type representation, which can be a statue, stone, jar, and print.

They used captions obtained from museum object inscriptions, so the caption namespace was rather formal and of an observational kind.

Closest to our task is the study [22] where captioning was done on an original Iconclass dataset [23], mainly composed of iconographic images. The authors fine-tuned a transformer-based vision-language pre-trained model using the Iconclass dataset. They also explored the model's ability to generalize to new data on the WikiArt subset and compared their model to ones that were trained on natural images to demonstrate the impact of context. Authors also noted that standard image captioning evaluation metrics are not very informative for artwork captioning, although they were working within one specific genre. As it was said earlier, architectures of image captioning models are inspired by ones from the machine translation field, thus, roughly speaking, they learn to translate image contents into words. So the question we were trying to answer is: can a painting's idea be translated from an image representation to natural language via machine learning?

## 3. METHOD

We adopt the transformer architecture presented in [5]. Transformers effectively overcome major drawbacks of RNN-based models: they facilitate long-range dependencies, do not suffer from gradient vanishing and gradient explosion, require fewer training steps, and do not have any recurrence, thus enabling parallelization. The main mechanism of a transformer is the attention mechanism, which was initially presented in an RNN-based model [13] to help the RNN decoder choose an encoded part of the input to attend to at each time step during sequence generation. Transformers got rid of RNNs altogether and work entirely on Attention, or in this case multi-head attention.

Image captioning training stage model diagram is depicted in Figure 1. At first, image is fed to ResNet101 to extract visual features. In our implementation, features were saved during preprocessing. Next, visual features and artionym are embedded with preserving positional information, and simultaneously fed to encoder and decoder. Encoder processes and passes image information to decoder, which utilizes it along with processed text. At each timestep, output of the decoder is passed to linear layer with softmax activation to produce next token.

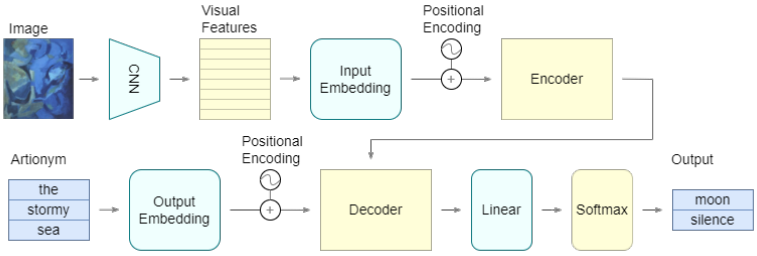

Figure 1. Artionym captioning with transformer: model diagram

### 3.1. Evaluation metrics

Automatic quality evaluation of the obtained captions is done by common image captioning metrics: BLEU score, ROUGE, METEOR, and CIDEr. This choice was also made in [22] mentioned earlier. These metrics, except for CIDEr, were first introduced for machine translation tasks, and their judging about artionyms is questionable, which we discuss in the next section.

BLEU score [16], utilizes a modified form of the precision metric:

$$BLEU = BP \cdot exp\sum_{n=1}^{N}\frac{1}{N}ln(p_n), \tag{1}$$

where BP is a brevity penalty, $N \in \{1,2,3,4\}$, $p_n$ is a modified n-gram precision. The brevity penalty is a constant depending on the lengths of the target and generated sentences, $N$ is a maximum length of the n-grams, usually from 1 to 4, and precision is modified to alleviate the effect of overrating repeated words in generated sequence.

ROUGE [18] was originally proposed as a summary scoring method. ROUGE that is, unlike BLEU, which uses modified precision, ROUGE modifies recall:

$$ROUGE - N = \frac{\sum\limits_{S \in RS} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in RS} \sum\limits_{gram_n \in S} Count(gram_n)}, \tag{2}$$

where $n$ is the length of the n-gram, $S$ is a text summary from the set of reference summaries $RS$, $gram_n$ and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. Image captioning often uses ROUGE-L, which utilizes the lengths of max matching substrings to evaluate the structure of sentences [18].

METEOR [24] is a combination of these two metrics. It exploits the harmonic mean between recall and nine precision:

$$\begin{gathered} METEOR = F_{mean} \cdot (1 - Penalty), \\ \text{where } F_{mean} = \frac{10PR}{R + 9P} \end{gathered} \tag{3}$$

and $Penalty$ here takes into account longer matches. Different modules of METEOR can also map cognates and synonyms using the WordNet database. This metric was designed to outperform the BLEU metric by addressing its shortcomings, such that it does not call recall and does not reward modifications of the same word [24].

Consensus-based image description evaluation score [17] was proposed in 2015 specifically for evaluating image captions. This metric is based on term frequency-inverse document frequency (TF-IDF): the weight of a word is proportional to the frequency of use of this word in a document and is inversely proportional to the frequency of use of a word in all documents of the collection. $CIDER_n$ score for n-grams of length $n$ is computed using the average cosine similarity between the candidate sentence and the reference sentences, which account for both precision and recall:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{||g^n(c_i)|| \cdot ||g^n(s_{ij})||}, \tag{4}$$

where $c_i$ is a candidate sentence,$S_i = \{s_{i1}, ...s_{im}\}$ is a set of reference sentences, $n$is the length of n-grams, $g^n(\cdot)$ is a vector formed by tf-idf weighting,and $|| \cdot ||$ is a magnitude. Then the scores from n-grams of varying length are combined as:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} \frac{1}{N} CIDEr_n(c_i, S_i) \tag{5}$$

## 4. DATA

Our experiments were conducted using the WikiArt dataset, a comprehensive collection of over 80,000 fine-art paintings from more than 1,000 artists spanning from the fifteenth century to modern times. The dataset, which includes 27 different styles (Table 1), is one of the largest collections of paintings currently available. A schematic description of the most commonly used and relevant art datasets is provided in [20].

The first step in our methodology involved cleaning the dataset. We removed untitled images and duplicates, resulting in a final count of 78,516 images. These images were then divided into training, validation, and testing sets, with 44,165, 14,722, and 19,629 images respectively, representing works from 1,102 artists. The images were resized to match the ResNet101 input shape of 224x224.

To further organize the dataset, we clustered the paintings into ten groups according to their style. The sequence of groups and the sequence of styles within each group were arranged in the order of their emergence in art history. The descriptions of these groups are depicted in Table 1. Each group is specified by style, number of examples in train, test, and validation sets, as well as a number of authors. Note that authors may overlap between groups. For the implementation of our model, we used the PyTorch framework. After testing various hyperparameters, we found that the Adam optimizer was most suitable, with an upper limit of the learning rate set at 1e-04. The learning rate decayed every three epochs at a rate of 0.1. We trained our model for 10 epochs using cross-entropy loss and a batch size of 16.

Table 1. WikiArt dataset partitioning

| | Styles | train | test | val | authors |
|---|---|---|---|---|---|
| Group 1 | Early Renaissance, High Renaissance, Northern Renaissance, Mannerism Late Renaissance | 3701 | 1603 | 1232 | 79 |
| Group 2 | Baroque, Rococo | 3463 | 1561 | 1243 | 94 |
| Group 3 | Ukiyo | 598 | 288 | 199 | 15 |
| Group 4 | Romanticism | 3944 | 1745 | 1281 | 105 |
| Group 5 | Realism, New Realism, Contemporary Realism | 6494 | 2845 | 2053 | 204 |
| Group 6 | Symbolism, Art Nouveau Modern | 4943 | 2176 | 1602 | 132 |
| Group 7 | Impressionism, Pointillism, Post Impressionism, Naive Art Primitivism, Fauvism | 12948 | 5783 | 4364 | 392 |
| Group 8 | Cubism, Analytical Cubism, Synthetic Cubism | 1394 | 599 | 487 | 139 |
| Group 9 | Expressionism, Abstract Expressionism, Color Field Painting, Action painting | 5384 | 2420 | 1829 | 373 |
| Group 10 | Pop Art, Minimalism | 1296 | 609 | 432 | 165 |

## 5. RESULTS AND ANALYSIS

Our experiments yielded average metric scores across the entire test corpus, as shown in Table 2. To understand the cause of the relatively low scores, we conducted an expert assessment of the quality of the generated names. This assessment aimed to determine whether the low scores were due to the n-gram nature of the metrics, the dataset, or the quality of the signatures.

Table 2. Average image captioning metrics scores on Wikiart test set

| Metric | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | CIDEr | METEOR |
|---|---|---|---|---|---|---|---|
| Value | 4.4 | 2.3 | 1.2 | 0.6 | 5.9 | 11.2 | 3.5 |

We selected five random samples from each group in the test set and solicited evaluations from a panel of 15 art experts and 17 non-experts, including 10 individuals from Amazon Mechanical Turk. The experts, who were art historians working in well-known museums, evaluated 50 auto-generated artionyms on a scale from 1 to 5. The average human scores are presented in Table 3, and the average human scores in percentages and metric results for each group are presented in Table 4. The automatic metric scores were computed for the entire group in the test set.

Our analysis revealed that experts rated the generated names five times higher than the automatic metrics. We observed that the realistic and impressionistic groups (5 and 7, respectively) received higher scores in both human and automatic evaluations. These groups also constitute a larger portion of the dataset. It's also worth noting that in our case, the automatic metrics have almost identical shapes in relation to the group train size. However, the similarity with human judgment is insufficient, as shown in Table 5.

Automatic metrics typically utilize uniformly weighted n-grams for $n$ from 1 to 4. However, most captions and reference lengths are less than 4. The mean caption length is only 2.3 with a variance of 0.2, and the mean reference length is 3.9 with a variance of 6.3. Even BLEU-1, which utilizes unigrams, does not adhere to human judgment.

Table 3. Questionnaire mean expert, non-expert and mutual human scores in percentages for corresponding style groups in WikiArt test set

| | Human | Expert | Non-expert |
|---|---|---|---|
| Group 1 | 2.3 ± 0.36 | 2.2 ± 0.71 | 2.4 ± 0.22 |
| Group 2 | 2.8 ± 0.97 | 2.7 ± 1.43 | 2.9 ± 0.74 |
| Group 3 | 2 ± 0.23 | 1.8 ± 0.57 | 2.1 ± 0.19 |
| Group 4 | 2.3 ± 0.38 | 2.2 ± 0.69 | 2.4 ± 0.25 |
| Group 5 | 3.2 ± 0.48 | 3.1 ± 0.82 | 3.3 ± 0.47 |
| Group 6 | 3.1 ± 0.71 | 3 ± 0.93 | 3.2 ± 0.58 |
| Group 7 | 3.4 ± 0.68 | 3.4 ± 0.96 | 3.3 ± 0.52 |
| Group 8 | 3 ± 0.64 | 3.1 ± 0.91 | 3 ± 0.55 |
| Group 9 | 2.6 ± 0.67 | 2.5 ± 1.56 | 2.7 ± 0.34 |
| Group 10 | 2.7 ± 0.4 | 2.8 ± 0.55 | 2.6 ± 0.3 |
| All | 2.75 ± 1.7 | 2.7 ± 1.7 | 2.8 ± 1.67 |

Table 4. Mean expert, non-expert, and mutual human scores with standard variance for style groups in questionnaire

| | Human | Experts | Non-experts | BLEU-1 | BLEU-4 | CIDEr | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|---|---|
| Group 1 | 45 | 43 | 47 | 6.3 | 0.8 | 11.8 | 8.4 | 5 |
| Group 2 | 57 | 54 | 59 | 5.1 | 0.6 | 8 | 6.8 | 3.8 |
| Group 3 | 40 | 36 | 42 | 2.1 | 0 | 3.4 | 2.8 | 1.6 |
| Group 4 | 47 | 45 | 49 | 4.5 | 0 | 12 | 6.8 | 3.7 |
| Group 5 | 65 | 63 | 67 | 5 | 0 | 12.6 | 6.6 | 4.1 |
| Group 6 | 62 | 59 | 63 | 3.2 | 0 | 8.3 | 4.9 | 2.7 |
| Group 7 | 67 | 68 | 66 | 4.9 | 0.4 | 14.6 | 6.9 | 4.2 |
| Group 8 | 61 | 61 | 61 | 2.6 | 0 | 9.8 | 4.2 | 2.6 |
| Group 9 | 52 | 50 | 54 | 2.6 | 0 | 7.7 | 3.5 | 1.9 |
| Group 10 | 54 | 56 | 53 | 1.1 | 0 | 4.8 | 1.4 | 1 |
| All | 55 | 53.6 | 56 | 4.4 | 0.6 | 11.2 | 5.9 | 3.5 |

Table 5. Pearson correlation coefficient between human scores and automatic metrics scores for generated artionyms

| | BLEU-1 | BLEU-4 | ROUGE-L | METEOR | CIDEr |
|---|---|---|---|---|---|
| Human | 0.14 | -0.7 | 0.17 | 0.23 | 0.5 |

## 6. DISCUSSION

Our findings indicate that computational metrics, particularly the BLEU metric, do not always accurately reflect the quality of the generated artionyms. This observation aligns with the findings of a previous study [22], where metrics were used to assess the quality of captions for iconographic images. One possible explanation for this is that the machine remains ”indifferent” to both the image itself and its name. In contrast, for a human observer, both the image and its name separately are significant factors in the perception of the world. To provide a broader context for our results, we compared our findings on the WikiArt dataset with those on the MS-COCO dataset [6] and the Iconclass dataset [23]. These comparisons were made with models of similar architecture [14] and a pretrained vision-language model [25]. The comparison results are presented in Table 6.

Our results indicate a significant gap between the BLEU-4 score for WikiArt and MS-COCO, with the BLEU-1 difference being approximately 18 times, ROUGE-L around 10 times, METEOR about 8 times, and CIDEr difference being approximately 11 times. This suggests that these metrics, particularly the BLEU metric, may not always accurately reflect the quality of the generated artionyms. Artionyms can appear ambiguous due to various factors such as the art style or the author’s idea. For example, the generated name ”Village Soldier” for Vincent van Gogh’s ”Orphans” seems suitable, but its automatic scores are zero due to missing matching n-grams.

We also conducted a comparison of expert and non-expert scores using the Mann-Whitney U rank test [26], with the null hypothesis being that they come from the same distribution. The test rejects the null hypothesis with a p-value of 0.498, indicating a resemblance between the two groups of estimators. To determine whether experts reached a consensus, we performed a k-sample Anderson-Darling test [27], with k equal to the number of experts. The null hypothesis that expert scores come from the same distribution can be rejected at a significance level of 0.001, indicating that experts did not come to an agreement. However, a pairwise Mann-Whitney U rank test between experts showed many similar opinions.

Table 6. Comparison of automatic metrics scores for generated captions across MS-COCO, Iconclass, and WikiArt datasets

| | MS-COCO | Iconclass | WikiArt |
|---|---|---|---|
| BLEU-1 | 80.5 | 12.8 | 4.4 |
| BLEU-4 | 38.6 | 10.0 | 0.6 |
| ROUGE-L | 58.4 | 31.9 | 5.9 |
| CIDEr | 128.3 | 172.1 | 11.2 |
| METEOR | 28.7 | 11.7 | 3.5 |

## 7. CONCLUSION

In this study, we made a concerted effort to address the existing gap in machine learning methods for artionyms captioning. We trained a transformer model on the WikiArt dataset, enabling it to predict an artionym across a wide spectrum of art styles. Our model was examined through both expert and automatic evaluations. The findings revealed that automatic metrics do not always correlate strongly with human judgment in this task. The highest correlation was observed with the CIDEr score, while the lowest was with the BLEU-4 score. Despite these metric discrepancies, human evaluation of the generated captions demonstrated a general adequacy of our model. Interestingly, while there was no consensus among the totality of expert opinions, we found that experts in pairs often shared similar views. This dichotomy underscores the complexity of professional opinions in the field, as well as the human element in the assessment of artionyms. Looking ahead, there are several potential avenues for advancing this task. One possibility is to fine-tune a more complex model from the pre-trained transformers family, which could enhance the quality of the generated artionyms. Additionally, training the model on a larger dataset of a specific genre could provide more nuanced and genre-specific artionyms. Furthermore the ambiguity of artionyms, influenced by factors such as art style or the author's idea, presents a challenge for automatic scoring metrics. Future research could explore methods to improve these metrics' ability to accurately reflect the quality of generated artionyms. In conclusion, our study represents a significant step forward in the application of machine learning to the generation of artionyms. While challenges remain, the findings provide a solid foundation for future research in this intriguing intersection of art and technology. Our hope is that these efforts will continue to push the boundaries of what is possible in the realm of artionyms captioning.

## ACKNOWLEDGEMENTS

This research was supported by a grant from the Russian Science Foundation (Project No. 20-18-00365). We acknowledge to Saint-Petersburg State University for a research project 95438429.

## REFERENCES

[1] S. Levy, "Magritte and words," *Journal of European Studies*, vol. 22, no. 4, pp. 313–321, 1992, doi: 10.1177/004724419202200402.

[2] M. Borisov, V. Kolycheva, A. Semenov, and D. Grigoriev, "The influence of color on prices of abstract paintings," in *Computational Data and Social Networks*, 2023, pp. 64–68 , doi: 10.1007/978-3-031-26303-3_6.

[3] A. Vasina, V. Bukanov, V. Kolycheva, A. Semenov, and D. Grigoriev, "The profitability of investing in fine art: an analysis of resale data from sotheby's, christie's, and phillips," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 17, pp. 6844–6852, 2023.

[4] T. Pimenova, V. Kolycheva, A. Semenov, D. Grigoryev, and A. Pimenov, "The impact of news, expert and public opinion on painting prices," in *The 12th International Conference on Computational Data and Social Networks*, 2024, pp. 426–430.

[5] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, 2017, pp. 1–11.

[6] T.-Y. Lin *et al.*, "Microsoft COCO: common objects in context," in *Computer Vision-European Conference on Computer Vision*, 2015, pp. 3686–3693 , doi: 10.1007/978-3-319-10602-1_48.

[7] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014, doi: 10.1162/tacl_a_00166.

[8] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778 , doi: 10.1109/CVPR.2016.90.

[10] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734 , doi: 10.3115/v1/d14-1179.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164 , doi: 10.1109/CVPR.2015.7298935.

[12] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[13] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, pp. 2048–2057.

[14] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–11, 2019.

[15] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10575–10584 , doi: 10.1109/CVPR42600.2020.01059.

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[17] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575 , doi: 10.1109/CVPR.2015.7299087.

[18] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, Barcelona, Spain, Association for Computational Linguistics, 2004.

[19] M. Kilickaya, A. Erdem, N. I. -Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning,” in *15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, vol. 1, pp. 199–209 , doi: 10.18653/v1/e17-1019.

[20] G. Castellano and G. Vessio, “Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview,” *Neural Computing and Applications*, vol. 33, no. 19, pp. 12263–12282, 2021, doi: 10.1007/s00521-021-05893-z.

[21] S. Sheng and M.-F. Moens, “Generating captions for images of ancient artworks,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2478–2486 , doi: 10.1145/3343031.3350972.

[22] E. Cetinic, “Iconographic image captioning for artworks,” in *ICPR International Workshops and Challenges*, 2021, pp. 502–516 , doi: 10.1007/978-3-030-68796-0_36.

[23] F. Milani and P. Fraternali, “A dataset and a convolutional model for iconography classification in Paintings,” *Journal on Computing and Cultural Heritage*, vol. 14, no. 4, pp. 1–18, 2021, doi: 10.1145/3458885.

[24] A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007, pp. 228–231.

[25] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and VQA,” in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020, pp. 13041–13049 , doi: 10.1609/aaai.v34i07.7005.

[26] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947, doi: 10.1214/aoms/1177730491.

[27] F. W. Scholz and M. A. Stephens, “K-sample Anderson-darling tests,” *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 918–924, 1987, doi: 10.2307/2288805.

## BIOGRAPHIES OF AUTHORS

**Anna Altynova** is Data Analyst at WAIW, working primarily on CV and LLM applications in geophysics. She got a bachelor degree at the Faculty of Mathematics and Computer science, Saint Petersburg State University (SPbU). Her research interests include semantic segmentation, NLP, and signal processing. She can be contacted at email: altynanke@gmail.com.

**Valeria Kolycheva** is an associate professor in the Department of Statistics, Accounting, and Audit at the Faculty of Economics, Saint Petersburg State University (SPbU). She earned her Ph.D. in Statistics from Saint Petersburg State University. Her research interests include art economics, art statistics, and art analytics. She can be contacted at email: v.kolycheva@spbu.ru.

**Dmitry Grigoriev** is an associate professor in the Department of Informatics at the Faculty of Mathematics and Computer Science, Saint Petersburg State University (SPbU). He earned his Ph.D. in Physics and Mathematics from Saint Petersburg State University. His research interests include multi-agent systems, data analysis, machine learning, and econometrics. He can be contacted at email: gridmer@mail.ru.

**Alexander Semenov** is an assistant research professor in the Department of Industrial and Systems Engineering, University of Florida. He earned his Ph.D. in Computer science from University of Jyvaskyla. His research interests include social network analysis, big data analytics, machine learning, and operations research. He can be contacted at email: asemenov@ufl.edu.