Automatic artworks captioning

Data efficient training through style transfer techniques

R. Barile[a]

[a]*Dipartimento di Informatica – Università di Bari*

November, 2022

Abstract

Automatic artwork captioning consists in the generation of sentences describing the content of the artworks. In this domain, available annotated data are very few and definitely not sufficient for the training of deep architectures. In this paper, we explore possible improvements of an existing approach based on virtual-real semantic alignment training process in which to provide sufficient training data, a virtual artwork captioning dataset is generated by applying style transfer to a large-scale photographic image captioning dataset and maintaining their annotations. In particular we tried to extend existing work adopting a different style transfer technique in the generation of the virtual artworks dataset.

1. Introduction

With the need for cultural preservation and art appreciation, artworks have been made digitalized and easily available in online museums. An important aspect, that is often missing in such online museums, is the artwork textual description that can be helpful for several aspects, such as:

- artworks content-based searching
- early childhood education for art
- visually impaired assistance in art appreciation

For this reason we explore the image captioning task in relation to artworks. Image captioning aims at automatically generating language to accurately describe images.
Large scale multi-modal data, in particular images and language, is required for the task of image captioning. MSCOCO dataset [25], Flicker 30 K [38], and Visual Genome [24] are datasets suitable for this task, but they contain only photographic images, so they are not directly useful for domain specific image captioning. Multi-modal data specific for artworks is not of satisfactory quality because often the descriptions include casual phrases, affective sentences or historical background.
The idea is to combine the two worlds obtaining an approach able to apply the clean and precise data available for photographic images to a domain specific task, in order to follow this direction we explore style transfer techniques. Style transfer [13] can combine a "content" image and a "target" image obtaining a third image with the semantic content of the former and the style of the latter, in our case it can reduce the visual domain gap between photographic images and artworks. Besides, the effectiveness of style transfer was already shown in the object detection task, in particular in [23] the use of style transfer improves the detection of people in artworks. Inspired by these, we generate virtual artworks by applying a style transfer procedure using photographic images as contents and artworks as targets. In order to use the generated virtual artworks as training data we employ the semantic alignment training proposed in [26]. We try to extend this work with the adoption of another set of virtual artworks obtained following the same approach, but with a newer style transfer approach, called StyleFlow [7]. We compare the results obtained in the two different setups.
This experiment did not show any significant improvement, in the conclusive sections some personal opinions and possible explanations of the unsatisfactory results will be provided.
The rest of this paper is arranged as follows. Section 2 reviews the related work of this paper. Section 3 presents the details of the proposed method. Section 4 describes the datasets. Section 5 shows the results and analysis. Section 6 shows the conclusion and future work.

2. Related works

2.1. Vision-and-language tasks of paintings

Current solutions to the tasks involving multi-modal data, e.g. artwork captioning, are based on interpreting the task as searching [3–5, 11, 32], answering [12, 31], or generating [12, 31]. Those tasks

reside in the intersection of several fields, including computer vision and natural language processing. Compared with other tasks in the artwork domain such as artwork classification [8, 9, 18, 33, 34], artwork object detection [23, 39] and influence modeling among artworks [20], the vision-and-language task need annotated data for both vision, language, and their paired information. Therefore, some works use image text cross-searching [3–5, 11, 32] or VQA [12, 31] to tackle the above problems. However, the image-sentence cross-searching task can only output existing sentences, resulting in limited description ability and poor diversity. The VQA task in painting needs sufficient annotated data in the training and inference processes but provides limited information, e.g., only a word or phrase as the answer. Compared with generating the descriptions directly, those tasks can only output existing sentences or phrases, thus having limited application value and generalization ability. Therefore, this paper aims at generating fluent content descriptions for painting directly rather than making it an image-sentence cross-searching task or VQA task.

2.2. Object detection

Object detection models [14, 15] can predict coordinates and dimensions of bounding boxes for the main objects recognized in an input image, a object category (e.g. person, dog, car) is then assigned to such bounding boxes. Object features can be extracted from the intermediate layers in an object detection model, object features represent a better image description because they are more precise and fine-grained with respect to plain CNN features. In state-of-the-art image captioning methods, like [6], object-level features are employed. Among the object detection models, Faster R-CNN is frequently used as a base framework to evaluate extra network components [29]. Following the work by [26], we use Faster R-CNN [27] model as our feature extractor since their semantic alignment loss procedure can be directly applied to this setup.

2.3. Image captioning

The machine learning task of image captioning consists in the generation of a sentence describing the content of an analyzed image. The classical approach is neural image captioning (NIC) model [36]; it uses a pre-trained convolutional neural network (CNN) to extract plain image features and a long short term memory network (LSTM) [17] to generate the sentence word by word based on the image features.

Main image captioning models consist, like NIC, of a feature extractor and a sentence generator [28, 37]. Follow-up studies attempt to optimize the two parts respectively to improve the image captioning performance. For the feature extractor, a fundamental model is the pre-trained CNN network.

As previously stated researchers [2, 6] begin to use object detection features rather than the pre-trained CNN features to feed the sentence generator.

For the sentence generator, transformers architecture [35] and its variants [6] were introduced to design more advanced structures of image captioning, achieving better performance than traditional structures.

In this paper, following [26], we utilize the meshed-memory transformer image captioning model [6] as sentence generator which use the object detector [27] as feature extractor.

2.4. Style transfer

The style transfer [21] can represent the content of a source image using the style of a target image.

The classical style transfer [13] uses the CNN features and their Gram Matrix to represent the image content and image style, respectively, and iteratively optimize the pixel of a noise image until getting a satisfactory style transfer result. This style transfer process is based on iterative optimization, so its generation process is slow. To accelerate the generation process, researchers [22] train an end-to-end network to generate the style transfer result through a fast feed-forward pass. However, this improved style transfer method can only refer to a fixed group of style images, so researchers [19] further utilize an adaptive instance normalization layer to construct an arbitrary style transfer model, which can refer to arbitrary images as reference style and generate the resulting image in real-time. For time efficiency and style diversity, we use the arbitrary style transfer method [19] to generate our virtual artworks dataset. In addition we also explore StyleFlow [7] which consists of invertible normalizing flows and a novel Style-Aware Normalization (SAN) module, we use it to generate another virtual artworks dataset for training.

3. Methodology

To tackle the challenges of lacking training data and the difficulty of abstract expressions in artworks, we adopt the virtual real semantic alignment training designed in [26] and summarized in the following:

Step 1: To employ the sufficient data in photographic image captioning dataset, a photographic image captioning model, which consists of an object feature extractor and a caption generator, is

pre-trained. This module act as a basic reference for the artwork caption model in the later steps.

Step 2: we generate two virtual artworks captioning dataset through style transfer. Specifically, we use the images in the photographic dataset as source images and the artworks as target images. We maintain the annotations, including object bounding boxes, object labels, and image captions of the photographic dataset.

Step 3: We fine-tune an artwork feature extractor using our generated virtual artworks dataset. Since the virtual paintings share the same annotations and semantics with their corresponding photographic images, we start from the object-level feature extractor of photographic images and fine-tune it through the semantic alignment loss of [26]

3.1. Pre-training image captioning model

To pretrain the image captioning model we use the photographic image dataset. Like previous works [2, 6], the first step is the training of the feature extractor, then its parameters are freezed and used to train the caption generator.

3.1.1. Pre-training feature extractor

The feature extractor is based on the framework of Faster RCNN object detection model [27]. Before the extraction process, a convolution neural network (CNN) backbone is used to extract the image features:

$$\mathcal{F}_x = f_{CNN}(x)$$

where x is the input image and the image features consist of a set of features at n CNN layers: $\mathcal{F}_x = \{\mathcal{F}_1, \mathcal{F}_1, \ldots, \mathcal{F}_n\}$. Faster R-CNN model predicts object labels and bounding boxes using its detection network $f_{detect}(\cdot)$. The CNN backbone and detection modules are trained by the detection loss:

$$\mathcal{L}_{detect} = f_{detect}(\mathcal{F}_x, y)$$

where y is the annotation of input image x. The detection loss $\mathcal{L}_{detect}$ consists of classification loss $\mathcal{L}_{cls}^{R-CNN}$ and regression loss $\mathcal{L}_{reg}^{R-CNN}$, and classification loss$\mathcal{L}_{cls}^{RPN}$ and regression loss $\mathcal{L}_{reg}^{RPN}$ of region proposal network (RPN):

$$\mathcal{L}_{detect} = \frac{\mathcal{L}_{cls}^{R-CNN} + \mathcal{L}_{reg}^{R-CNN} + \mathcal{L}_{cls}^{RPN} + \mathcal{L}_{reg}^{R-CNN}}{4}$$

After training the detection model, the object features can be obtained by pooling the network intermediate features according to the object bounding boxes.

3.1.2. Pre-training caption generator

We use a meshed-memory transformer method [6] to generate captions based on the object features. The caption generator is composed of:

- a feature encoder
- a sentence decoder

consists of a feature encoder and a sentence decoder.

Starting from the object features Q, the feature encoder encodes them into an hidden representation that we will call $\mathcal{X}$: $\mathcal{X} = f_{encoder}(\mathcal{L})$ Then, the sentence decoder outputs the word probabilities at every time step i:

$$p_i = f_{decoder}(\mathcal{X}, \mathcal{S}_{i-1}) \;\; i = 1, 2, \ldots, N$$

where $\mathcal{S}$ is the ground truth sentence formed by n words and $\mathcal{S}_0$ is the starting token. p contains the output probabilities of words in each position and $p = \{p_1, p_2, \ldots, p_n\}$.
The feature encoder and sentence decoder are trained by optimizing the cross-entropy loss following [36]:

$$\mathcal{L}_{XE} = -\sum_{t=1}^{N} logp_t(\mathcal{S}_t)$$

At the inference stage, the sentence decoder outputs the word probabilities based on the generated word at the former time step rather than the ground truth word. We also use the beam search to improve the variety of generated descriptions.

3.2. Generating virtual painting dataset

To overcome the data-hungry problem in artwork captioning task, we aim to generate a dataset as training data that have a smaller domain gap to paintings than the photographic datasets.
To provide sufficient training data for artworks captioning, the generated dataset should satisfy the following requirements: 1) Its images should have similar styles with artworks. 2) The images should be annotated with sentences that can be used for training the image captioning model. 3) It should have a large quantity of data to satisfy the needs of data quantity for training the image captioning model. We use the style transfer [13] as the generation method to meet the above requirements.
In particular, we use

- a real-time arbitrary style transfer method [19] to generate virtual paintings, which use an image as the style reference and another image as the content reference to generate the final output image. This style transfer method [19] calculates the features by the

input style image and content image and decodes the features into the style transfer result:

$$x_v = decoder(AdaIN(f(x_r), f(x_p)))$$

$$y_v := y_r$$

where we use the MSCOCO image as content image x_r, and the artwork image from WikiArt painting image dataset [30] as style image x_v. y_v and y_r are the labels of x_v and x_r, respectively. $AdaIN(\cdot)$ calculates the adaptive instance normalization based on the style and content images and $decoder(\cdot)$ generates the resulting image based on the instance normalization result.

- StyleFlow [7] which consists of invertible normalizing flows and a novel Style-Aware Normalization (SAN) module to achieve content-fixed image-to-image translation. Following previous works, a pretrained and fixed VGG-19 encoder E_{VGG} is utilized in StyleFlow for feature extraction and loss computation. As the model is invertible, the forward pass is denoted as E, and the backward pass as E^{-1}. Taking a source domain X and a target domain Y as an example. The forward pass of StyleFlow maps images from source domain into deep features, i.e. $E : X \mapsto F_X$, where F_X is source feature space. The pre-trained VGG encoder would map the target domain into a shared style space, i.e. $E_{VGG} : Y \mapsto F_S$. Taking F_X and F_S as inputs, SAN module performs content-fixed feature transformation to obtain features in the stylized space $\hat{F}_X$. And the backward pass of StyleFlow generates translated images by mapping the stylized features back to the image space, i.e. $E^{-1} : \hat{F}_X \mapsto \hat{X}$, where $\hat{X}$ is assumed to share the style properties of Y while retain the content information of X.

3.3. Training painting feature extractor

Our feature extractor is based on the Faster R-CNN [27] object detection model. Specifically, we obtain the object features by pooling the network intermediate features according to the object bounding boxes. Therefore, we can obtain the artwork feature extractor following the approach in [26]. The generated virtual artwork dataset alleviates the problem of insufficient available training data, however, the abstract expressions of virtual artworks make it hard to train the object detection model only relying on the original detection loss proposed by [27]. As the generated virtual painting image x_v has the same content with its corresponding photographic image x_r, they should have similar feature representations extracted by fully trained feature extractors. Therefore, the adopted semantic alignment loss L_{align} has the objective of improving the consistency between the virtual artwork features and the corresponding photographic image features. The semantic alignment loss L_{align} calculates the average of the mean-squared errors between the image features of the photographic image and the virtual artwork, and we combine it with the original detection loss to train the artwork feature extractor by the following steps:
First, we obtain the photographic image features:

$$\mathcal{F}^r = f^r_{CNN}(x_r), \quad \mathcal{F}^r = \{\mathcal{F}^r_1, \mathcal{F}^r_2, \ldots, \mathcal{F}^r_n\}$$

where the module $f^r_{CNN}(\cdot)$ is the CNN backbone for the photographic image, which is pre-trained on the MSCOCO dataset. The module $f^r_{CNN}(\cdot)$ is frozen when training the virtual artworks feature extractor. Second, we obtain the virtual artworks features:

$$\mathcal{F}^v = f^v_{CNN}(x_v), \quad \mathcal{F}^v = \{\mathcal{F}^v_1, \mathcal{F}^v_2, \ldots, \mathcal{F}^v_n\}$$

where the module $f^v_{CNN}(\cdot)$ is the CNN backbone for the virtual painting. For $f^v_{CNN}(\cdot)$ we employ the transfer learning process rather than training a new model from random initialization. Specifically, the module $f^v_{CNN}(\cdot)$ is initialized from the pre-trained $f^r_{CNN}(\cdot)$ in the first step. Third, we obtain the semantic alignment loss $L_a lign$ by calculating the average of the mean-squared error between the photographic image features and the virtual painting features:

$$L_{align} = \frac{1}{n} \sum_{j=0}^{n} \sum_{i=0}^{m_i} \frac{(\mathcal{F}^v_{ji} - \mathcal{F}^r_{ji})^2}{m_i}$$

where $\mathcal{F}^v_{ji}$ and $\mathcal{F}^r_{ji}$ are the i-th element of the j-th level feature vector of the virtual painting features $\mathcal{F}^v$ and the photographic image features $\mathcal{F}^r$, respectively. m_i is the number of elements in the feature vector $\mathcal{F}^v_i$. Fourth, we obtain the detection loss on virtual paintings following the pre-training process in Section 3.1.1:

$$L^v_{detect} = f^v_{detect}(\mathcal{F}^v, y_v)$$

where y_v is the annotations of the virtual painting image x_v. As the detection network $f^r_{detect}(\cdot)$ is fully trained by the MSCOCO dataset and the virtual artwork dataset has the same object label distribution with the MSCOCO dataset, we copy the parameters of the detection network $f^r_{detect}(\cdot)$ the detection network $f^r_{detect}(\cdot)$ and freeze $f^r_{detect}(\cdot)$ when training artwork feature extractor. Finally, we train the virtual artwork CNN backbone $f^v_{CNN}(\cdot)$ by minimizing the sum of the proposed semantic

alignment loss and the original detection loss to train $f^v_{CNN}(\cdot)$

$$L_{total} = \alpha L_{align} + L^v_{detect}$$

where α is the semantic alignment loss ratio.

4. Datasets and Implementation details

4.1. Datasets

This paper involves four datasets to train the painting captioning model and three datasets for evaluation. In the training process, the involved datasets are MSCOCO photographic image dataset [25], WikiArt artworks dataset [30], and the two virtual artwork datasets that we generated. In the evaluation process, we use three artwork captioning dataset: OLA [1], SemArt [10] and ArtPedia [32].

MSCOCO photographic image captioning dataset: The MSCOCO dataset is a photographic image dataset that has both object detection annotations and image captioning annotations. This dataset contains 118287 images and each image is annotated with object labels, object bounding boxes, and 5 sentences describing the image content.

WikiArt artworks image dataset: WikiArt is a website that contains many artworks annotated with information including style, genre, author, and artwork size. Wikiart is growing with time in the number of images and annotations it makes available. This dataset is used for the training of the style transfer models.

Virtual painting captioning dataset - AdaIN: This dataset is generated by the style transfer model AdaIn to satisfy our need for painting captioning. On the one hand, it was a derived MSCOCO dataset with styles transferred from WikiArt dataset, thus has a smaller domain gap to the paintings than the MSCOCO dataset. On the other hand, it shares both the objects and captioning annotations of MSCOCO dataset.

Virtual painting captioning dataset - styleflow: This dataset is generated with the same method and rationale of the previous one, the only difference is the style transfer technique, in this case StyleFlow is adopted.

Objective Language for Art (OLA) painting captioning dataset: The OLA dataset [1] is a publicly available painting captioning dataset, which contains 5000 paintings and each painting is annotated with one sentence. The annotated sentences are required to contain only objective (not affective) descriptions of the artwork. We use the OLA Dataset as evaluation dataset.

SemArt: SemArt is a multi-modal dataset for semantic art understanding. SemArt is a collection of fine-art painting images in which each image is associated to a number of attributes and a textual artistic comment, such as those that appear in art catalogues or museum collections. It contains 21,384 samples that provides artistic comments along with fine-art paintings and their attributes for studying semantic art understanding.

ArtPedia: contains paintings and textual sentences describing both the visual content of the paintings and other contextual information. Thus, the annotators also identified which sentences actually describe the visual content of a given image. As we also use the WikiArt dataset to train the style transfer model and generate our virtual artworks datasets, the images in the OLA, SemArt and ArtPedia datasets will not be used in the style transfer process. Compared with photographic image captioning datasets, the OLA, SemArt and ArtPedia datasets are small-scale but effective enough for evaluation purposes.

4.2. Implementation details

The feature extractor is based on the Faster R-CNN model [27] with FPN + ResNet101 [16] as the backbone. This model is available pretrained for object detection on MS COCO for 37 epochs.

The caption generator is based on the meshed-memory transform image captioning model [6]. We train the model on the object features with batch size 100 under the cross-entropy loss. We use the learning rate scheduling strategy the same as [40], together with a learning rate warm-up of 10000 iterations. We train the model with early stopping with a patient of 5. The max feature size is 50, and the memory slot is 40.

Generating virtual painting dataset - AdaIn: We train the arbitrary style transfer model [19] using COCO as content images and wikiart as target images. We used as content image size $S_c = (512, 512)$ and style image size $S_s = (512, 512)$. In the inference phase for each content image we use a random style image selected from Wikiart.

Generating virtual painting dataset - StyleFlow: Also StyleFlow [7] is trained using COCO as content images and wikiart as target images. The inference phase is sampling based, the style code is sampled from learned latent distribution.

Training painting feature extractor: Different from the pretraining process, we employ the generated virtual artworks to train the painting feature extractor. The network is optimized by the stochastic gradient descent optimizer with the learning rate of 0.005, weight decay of 0.0005, and momentum of 0.9, where we drop the learning rate to 10% for every 3 epochs. We also utilize a learning rate warm-up by pulling it linearly from 0 to 0.005

during the first 5000 iterations. We set the semantic alignment loss ratio as 1.0. To guarantee the same semantic meaning between the virtual painting and the photographic image, we disable the random flipping operation in training the painting feature extractor. We freeze the detection network of Faster R-CNN to maintain their detection ability already trained in the pre-training stage.

5. Results

In this unsupervised artwork captioning scenario, there are no annotated paintings for training. Our model employs the generated virtual artworks datasets (using AdaIn and StyleFlow) to train the artworks feature extractor, and it is combined with the caption generator pre-trained on the MSCOCO dataset to form the artwork captioning model. As baseline model in this scenario we use the pretrained captioning model on the MSCOCO dataset. We report the results of the baseline and of the models trained with the two different virtual dataset in tables 1, 2, 3. In all datasets we notice that the best performance is obtained using the baseline model. We think that the performance decrease is due to the semantic alignment loss, in particular in the existing literature this loss is proposed as sum, instead of average, of mean squared errors, but trying this setup the training process does not converge due to really high loss values. In addition, in the referred works is not specified what are the elements of the feature vector obtained as output of the backbone, since we used a FPN as backbone we assumed that the features vector is composed of the different levels of the pyramid network. This two choices are our assumptions because no indication are provided in the literature, investigating more in this directions may lead to better results.

6. Conclusion

To deal with the data-hungry problem and abstract expressions in artwork captioning task, we experimented with two virtual artworks generation methods and a virtual-real semantic alignment training process proposed in [26] to build a artworks captioning model. The proposed method is evaluated on three public painting captioning datasets. The experiment results are unsatisfactory, they showed a decrease in performances with respect to the baseline, the reason may be in the way we implemented the loss because such implementation required to make assumption on aspects not reported in literature.
In the future more representative features can be exploited by employing the photographic datasets with more fine-grained annotations, e.g., the instance segmentations. At last, the designed framework can provide a possible strategy for the description generation in other domains, e.g., pathological description generation of the medical images [40], for the interpretation of diagnostic models or to provide a diagnostic reference for junior doctors.

References

[1] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas, *Artemis: affective language for visual art*, 2021, 10.48550/ARXIV.2101.07396.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, ≪Bottom-up and top-down attention for image captioning and visual question answering≫, in 2018 ieee/cvf conference on computer vision and pattern recognition (2018), pp. 6077–6086, 10.1109/CVPR.2018.00636.

[3] L. Baraldi, M. Cornia, C. Grana, and R. Cucchiara, ≪Aligning text and document illustrations: towards visually explainable digital humanities≫, in 2018 24th international conference on pattern recognition (icpr) (2018), pp. 1097–1102, 10.1109/ICPR.2018.8545064.

[4] C. Bartz, N. Jain, and R. Krestel, ≪Automatic matching of paintings and descriptions in arthistoric archives using multimodal analysis≫, English, in Proceedings of the 1st international workshop on artificial intelligence for historical image enrichment and access (May 2020), pp. 23–28, ISBN: 979-10-95546-63-4.

[5] A. Carraggi, M. Cornia, L. Baraldi, and R. Cucchiara, ≪Visual-semantic alignment across domains using a semi-supervised approach≫, in Proceedings of the european conference on computer vision (eccv) workshops (2018), pp. 0–0.

[6] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, ≪Meshed-memory transformer for image captioning≫, in Proceedings of the ieee/cvf conference on computer vision and pattern recognition (2020), pp. 10578–10587.

[7] W. Fan, J. Chen, J. Ma, J. Hou, and S. Yi, *Styleflow for content-fixed image to image translation*, 2022, 10.48550/ARXIV.2207.01909.

[8] N. Garcia, B. Renoust, and Y. Nakashima, ≪Context-aware embeddings for automatic art analysis≫, in Proceedings of the 2019 on international conference on multimedia retrieval, ICMR '19 (2019), pp. 25–33, ISBN: 9781450367653, 10.1145/3323873.3325028.

Table 1: Artwork captioning performances on OLA dataset

Method	METEOR	ROUGE	CIDEr	BLEU1	BLEU2	BLEU3	BLEU4
Baseline	0.058	0.202	0.099	0.173	0.073	0.031	0.015
AdaIn style transfer	0.057	0.200	0.095	0.169	0.072	0.030	0.015
StyleFlow style transfer	0.054	0.197	0.075	0.166	0.068	0.025	0.011

Table 2: Artwork captioning performances on SemArt dataset

Method	METEOR	ROUGE	CIDEr
Baseline	0.011	0.067	0.0016
AdaIn	0.011	0.065	0.0013
StyleFlow	0.010	0.063	0.0011

Table 3: Artwork captioning performances on ArtPedia dataset

Method	METEOR	ROUGE	CIDEr
Baseline	0.037	0.152	0.0123
AdaIn	0.035	0.149	0.0121
StyleFlow	0.034	0.145	0.008

[9]N. Garcia, B. Renoust, and Y. Nakashima, «Contextnet: representation and exploration for painting classification and retrieval in context», International Journal of Multimedia Information Retrieval **9**, 10.1007/s13735-019-00189-4 (2020) 10.1007/s13735-019-00189-4.

[10]N. Garcia and G. Vogiatzis, *How to read paintings: semantic art understanding with multi-modal retrieval*, 2018, 10.48550/ARXIV.1810.09617.

[11]N. Garcia and G. Vogiatzis, «How to read paintings: semantic art understanding with multi-modal retrieval», in Proceedings of the european conference on computer vision (eccv) workshops (2018), pp. 0–0.

[12]N. Garcia, C. Ye, Z. Liu, Q. Hu, M. Otani, C. Chu, Y. Nakashima, and T. Mitamura, «A dataset and baselines for visual question answering on art», in Computer vision – eccv 2020 workshops, edited by A. Bartoli and A. Fusiello (2020), pp. 92–108, ISBN: 978-3-030-66096-3.

[13]L. A. Gatys, A. S. Ecker, and M. Bethge, «Image style transfer using convolutional neural networks», in 2016 ieee conference on computer vision and pattern recognition (cvpr) (2016), pp. 2414–2423, 10.1109/CVPR.2016.265.

[14]R. Girshick, *Fast r-cnn*, 2015, 10.48550/ARXIV.1504.08083.

[15]R. Girshick, J. Donahue, T. Darrell, and J. Malik, «Rich feature hierarchies for accurate object detection and semantic segmentation», in Proceedings of the ieee conference on computer vision and pattern recognition (2014), pp. 580–587.

[16]K. He, X. Zhang, S. Ren, and J. Sun, «Deep residual learning for image recognition», in Proceedings of the ieee conference on computer vision and pattern recognition (2016), pp. 770–778.

[17]S. Hochreiter and J. Schmidhuber, «Long Short-Term Memory», Neural Computation **9**, 1735–1780, ISSN: 0899-7667 (1997) 10.1162/neco.1997.9.8.1735.

[18]X. Huang, S.-h. Zhong, and Z. Xiao, «Fine-art painting classification via two-channel deep residual network», in Pcm (2017).

[19]X. Huang and S. Belongie, «Arbitrary style transfer in real-time with adaptive instance normalization», in Proceedings of the ieee international conference on computer vision (2017), pp. 1501–1510.

[20]N. Huckle, N. Garcia, and Y. Nakashima, «Demographic influences on contemporary art with unsupervised style embeddings», in Computer vision – eccv 2020 workshops, edited by A. Bartoli and A. Fusiello (2020), pp. 126–142, ISBN: 978-3-030-66096-3.

[21]Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, «Neural style transfer: a review», IEEE Transactions on Visualization and Computer Graphics **26**, 3365–3385 (2020) 10.1109/TVCG.2019.2921336.

[22]J. Johnson, A. Alahi, and L. Fei-Fei, «Perceptual losses for real-time style transfer and super-resolution», in Computer vision – eccv 2016, edited by B. Leibe, J. Matas, N. Sebe, and M. Welling (2016), pp. 694–711, ISBN: 978-3-319-46475-6.

[23]D. Kadish, S. Risi, and A. S. Løvlie, *Improving object detection in art images using only style transfer*, 2021, 10.48550/ARXIV.2102.06529.

[24]R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and F.-F. Li, *Visual genome: connecting language and vision using crowdsourced dense image annotations*, 2016, 10.48550/ARXIV.1602.07332.

[25]T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, «Microsoft coco: common objects in context», in Computer vision – eccv 2014, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (2014), pp. 740–755, ISBN: 978-3-319-10602-1.

[26]Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, «Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training», Neurocomputing **490**, 163–180, ISSN: 0925-2312 (2022) https://doi.org/10.1016/j.neucom.2022.01.068.

[27]S. Ren, K. He, R. Girshick, and J. Sun, «Faster r-cnn. towards real-time object detection with region proposal networks. ieee transactions on pattern analysis and machine intelligence», (2017).

[28]S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, «Self-critical sequence training for image captioning», in Proceedings of the ieee conference on computer vision and pattern recognition (2017), pp. 7008–7024.

[29]K. Saito, Y. Ushiku, T. Harada, and K. Saenko, «Strong-weak distribution alignment for adaptive object detection», in Proceedings of the ieee/cvf conference on computer vision and pattern recognition (2019), pp. 6956–6965.

[30]B. Saleh and A. Elgammal, «A unified framework for painting classification», in 2015 ieee international conference on data mining workshop (icdmw) (2015), pp. 1254–1261, 10.1109/ICDMW.2015.93.

[31]S. Sheng, L. Van Gool, and M.-F. Moens, «A dataset for multimodal question answering in the cultural heritage domain», in Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH) (Dec. 2016), pp. 10–17.

[32]M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara, «Artpedia: a new visual-semantic dataset with visual and contextual sentences in the artistic domain», in (Sept. 2019), ISBN: 978-3-030-30644-1, 10.1007/978-3-030-30645-8_66.

[33]W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, «Ceci n'est pas une pipe: a deep convolutional network for fine-art paintings classification», in 2016 ieee international conference on image processing (icip) (2016), pp. 3703–3707, 10.1109/ICIP.2016.7533051.

[34]C. B. E. Vaigh, N. Garcia, B. Renoust, C. Chu, Y. Nakashima, and H. Nagahara, *Gcnboost: artwork classification by label propagation through a knowledge graph*, 2021, 10.48550/ARXIV.2105.11852.

[35]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2017, 10.48550/ARXIV.1706.03762.

[36]O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, «Show and tell: a neural image caption generator», in Proceedings of the ieee conference on computer vision and pattern recognition (2015), pp. 3156–3164.

[37]K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, «Show, attend and tell: neural image caption generation with visual attention», in International conference on machine learning (PMLR, 2015), pp. 2048–2057.

[38]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 **2**, 67–78, ISSN: 2307-387X (2014) 10.1162/tacl_a_00166.

[39]N.-A. Ypsilantis, «Instance-level recognition for artworks», (2022).

[40]Z. Zhang, D. Wang, and Y. Guo, «Fretting friction and wear behaviors of spiral wound gasket (swg) sealing surface», Tribology International **133**, 236–245, ISSN: 0301-679X (2019) https://doi.org/10.1016/j.triboint.2019.01.017.