Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Revisiting image captioning via maximum discrepancy competition

Boyang Wan[a], Wenhui Jiang[a], Yu-Ming Fang[a,*], Minwei Zhu[a], Qin Li[a], Yang Liu[b]

[a] Jiangxi University of Finance and Economics, Nanchang, China
[b] SANY Heavy Industry Co., Ltd., Beijing, China

ARTICLE INFO

Article history:
Received 19 January 2021
Revised 23 August 2021
Accepted 29 September 2021
Available online 1 October 2021

Keywords:
Image captioning
Model comparison
Attention mechanism

ABSTRACT

Image captioning is a hot research topic bridging computer vision and natural language processing during the past several decades. It has achieved great progress with the help of large-scale datasets and deep learning techniques. Though the variety of image captioning models (ICMs), the performance of ICMs have got stuck in a bottleneck judging from the publicly published results. Considering the marginal performance gains brought by recent ICMs, we raise the following question: "what about the performances of the recent ICMs achieve on in-the-wild images? To clarify this question, we compare existing ICMs by evaluating their generalization ability. Specifically, we propose a novel method based on maximum discrepancy competition to diagnose existing ICMs. Firstly, we establish a new test set containing only informative images selected by adopting maximum discrepancy competition on the existing ICMs, from an arbitrary large-scale raw image set. Secondly, a small-scale and low-cost subjective annotation experiment is conducted on the new test set. Thirdly, we rank the generalization ability of the existing ICMs by comparing their performances on the new test set. Finally, the keys of different ICMs are demonstrated based on a detailed analysis of experimental results. Our analysis yields several interesting findings, including that 1) Using simultaneously low- and high-level object features may be an effective tool to boost the generalization ability for the Transformer based ICMs. 2) Self-attention mechanism may provide better modelling ability for inter- and intra-modal data than other attention-based mechanisms. 3) Constructing an ICM with a multistage language decoder may be a promising way to improve its performance.

1. Introduction

Image captioning is a fundamental but significant research topic for visual information expression in cross-media area [1]. In the past five years, driven by off-the-shelf large-scale datasets like Flickr30K [2] and COCO [3,4] and considerable efforts have been put to design novel model architectures [5–9] and efficient training strategies [10–12], deep learning techniques have greatly contributed to the rapid progress of image captioning.

Generally, the comparison of ICMs is implemented to compare their scores in terms of various metrics, including BLEU [13], CIDEr [14], METEOR [15], ROUGE-L [16] and SPICE [17]. An ICM with higher scores on large-scale datasets is considered better. However, recent ICMs, such as NG-SAN [18], X-LAN [8] and MM [7], present only marginal improvements in conventional evaluation metrics, as is clearly illustrated from Fig. 1. For example, the difference between their performance is not larger than 0.6% in terms of CIDEr. Considering the tiny performance boost on 0 finite test set (*i.e.*, 5000 samples in COCO), we wonder if the improvement is consistent when evaluated on a larger-scale independent test set.

Intuitively, creating a new dataset from the natural world with more diverse samples, is the most direct method to compare different ICMs. However, the expensive time and labour cost of a large-scale subjective annotation experiment are unaffordable. In this paper, instead of building large-scale subjective annotations, we compare existing ICMs by ranking their generalization ability on a new small-scale test set, which contains informative images automatically collected from the natural world and only requires an affordable small-scale subjective experiment.

Specifically, we propose a model comparison method by adopting a maximum discrepancy competition methodology. Our motivation is that ICMs can be effectively compared when they generate different captions and always behave differently on informative images. Thus, we can automatically collect small-scale informative images based on the discrepancy of the captions generated by different ICMs. To measure the discrepancy between every two predictions of two difference compared ICMs on an image, we propose a new similarity measure named NGSM (N-Gram-based Similarity Metric), which gets permutation-invariant discrepancy be-

[*] Corresponding author.
E-mail addresses: 2201810057@stu.jxufe.edu.cn (B. Wan), jiang1st@bupt.edu.cn (W. Jiang), fa0001ng@e.ntu.edu.sg (Y.-M. Fang), 413740229@qq.com (M. Zhu), liqin4948@163.com (Q. Li).

https://doi.org/10.1016/j.patcog.2021.108358

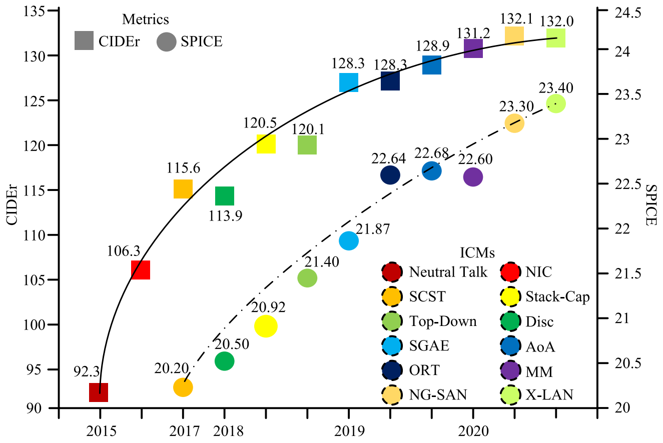


Fig. 1. Illustration of performances of state-of-the-art ICMs on COCO Karpathy test split. The solid and dashed lines show the development trajectories of the state-of-the-art ICMs on SPICE and CIDEr respectively. The listed twelve ICMs include Neural Talk [19], NIC [1], SCST [10], Stack-Cap [20], Top-Down [6], Disc [11], SGAE [21], AoA [19], ORT [9], MM [7], NG-SAN [18] and X-LAN [8].

tween two captions. The small-scale test set only requires limited efforts to perform a low-cost subjective annotation experiment. Finally, We achieve model comparison on ICM pairs and get global ranking results by using pairwise competing results. By analyzing the competing results, we give some summative conclusions about the competing ICMs and point out some potential ways to construct a robust ICM.

In summary, we would like to make the following technical contributions:

- We propose a new model comparison method without an unaffordable large-scale subjective annotation experiment.
- A new similarity function named NGSM actives as a semantic distance measure to model discrepancy of captions. With this NGSM, the informative images can be selected effectively from an arbitrary large-scale raw image dataset.
- We demonstrate quantitative results of the generalization ability of the competing ICMs and provide detailed analysis about the key factor of improving the generalization ability of ICMs.

The rest of the paper is organized as follows. In Section 2, we first review some related works on image captioning methods and datasets and then briefly introduce some model comparison works. Section 3 introduces our model comparison method in detail. Exhaustive experiments are introduced in Section 4. Finally, Section 5 concludes the paper.

2. Related work

2.1. Image captioning methods

Most of ICMs follow the two-state paradigm, including visual information extractor and language generator. Both the early extractors and generators are either hand-crafted or rule-based. Visual information is extracted by using support vector machine [22] or conditional random fields [23,24] and the visual information is fed to templates [23,25,26], tree-based model [27], n-gram based sequences model [22] for generating sentences. Obviously, the performance of these ICMs highly depends on their visual information extractors, and these extractors can not capture the relationship between different objects.

Recently, the encoder-decoder models dominate the field of image captioning, and most of them are trained in a typical end-to-end manner. The encoder of multimodal recurrent neural network (m-RNN) [28] is a convolutional neural network (CNN) [29], and the decoder is a modified RNN that fuses the feature of each time-step word and visual information of the whole image to generate the next word. In [1,19], the proposed ICM is composed of a more powerful CNN, which is Region-CNN [30] or GoogLeNet [31], and their long-short term memory (LSTM) network. Obviously, different words are activated from certain image regions, but this correspondence is not modeled in [1] and [19]. To address this problem, the visual attention mechanism, implemented as an attention objection function, is introduced in [32], and this model has better interpretability. Furthermore, to enhance image captioning with salient mechanism, [33] proposed a global-local attention model to extract global and local features of an image simultaneously. [34] proposed a context-aware attention mechanism implemented as a Graph Neural Network (GNN) to implicitly model the relationship among regions of interest in an image.

To optimize ICMs by using image captioning metrics directly, [33] proposed a Reinforcement Learning (RL) method named Self-Critical Sequence Training (SCST), which uses image captioning metrics as reward functions. To make an ICM generates discriminative captions, [11] proposed a new reward function to disambiguate image-caption matches.

To improve image captioning via bottom-up attention mechanism, [6] used a Faster R-CNN model, which is pre-trained on Visual Gnome dataset [35], to extract object-level visual information named bottom-up feature as implementation of bottom-up attention mechanism. The bottom-up feature is widely used as representation of visual information in image captioning. [21] proposed a scene graph auto-encoder (SGAE) to guide image caption by modeling the language inductive bias. [36] proposed a novel self-attention module to explicitly attention in both encoder and decoder of their ICM. [8] design a 2^{nd} attention module by introducing the Bilinear Pooling to model interactions between singlemodal and multimodal inputs. [7] used multi-level encoder to learn prior knowledge and propose a fully attentive decoder without using LSTM.

Latterly, there are some image captioning sub-tasks, including novel image captioning [37] and controllable image captioning [38]. In this paper, we focus on the ICMs of the standard image captioning task.

2.2. Attention mechanism

The attention mechanism is a classic and popular research area, and attention based models are widely used in various tasks. Itti et al. [39] creatively proposed a visual attention system, which consists of three head-crafted feature extractors and generates the saliency map for images. The saliency map is a quantitative metric of visual attention. Further, Fang et al. [40] proposed a new visual attention model based on Gestalt theory to obtain the saliency map of video frames. In [41], to utilize both long-term and short-term cues of the video, a deep-learning based model contains a 2D CNN, a 3D CNN and a ConvLSTM is established for video saliency detection task. As proof of the attention mechanism, several attention based models achieve good results on various computer vision tasks, including image classification [42,43], object detection [43] and semantic segmentation [44]. For the NLP area, [45] first introduced the attention mechanism into Neural Machine Translation (NMT) task. A soft-attention based LSTM is proposed to alleviate the bottleneck caused by using a fixed-length vector. Recently, self-attention based models, including Transformer [46], Bert [47] and so on, dominate the off-the-shelf NLP tasks. As a cross-area task of CV and NLP, the image captioning is dominated by attention based models. As a pioneer, [32] first proposed an ICM including an attentive weight to measure the relative importance of each element in visual features for generating words. After, with the seminal work "Bottom-Up and Top-Down", [6] imple-

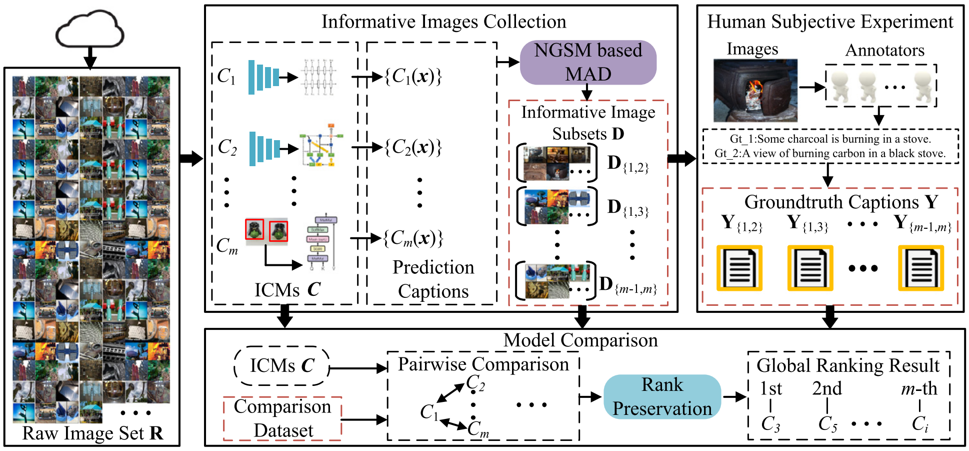


Fig. 2. Illustration of the procedure of our model comparison method which contains three components: informative images collection, human subjective experiment and model comparison. Comparison dataset consist of informative image subsets and their corresponding groundtruth captions.

mented the bottom-up and top-down attention as a pretrained object detection model and a two-layer LSTM, respectively. Lately, introducing the self-attention mechanism is a prevailing way in image captioning [7–9,36]. Among them, [9] focused on fusing geometric and visual features by self-attention model. After shortly, [36] and [8] proposed two high-order interaction self-attention modules. As the other line, [7] proposed a novel Transformer-based ICM based on the full-attentive paradigm. We analyse the correlative of attention mechanisms and generalization ability for the competing ICMs in this paper.

2.3. Model comparison

Model comparison is a long-term issue in the machine learning community. A common paradigm of model comparison methods is collecting a number of samples for a specified task, constructing human annotations as ground-truth, and selecting the model that outputs the most accurate predictions. The slight gap between the performances is that the generalization ability comparison of different models raises a request to build a new large-scale dataset for image caption.

In this work, we propose a new model comparison method for image captioning by adopting the MAximum Discrepancy competition methodology (MAD) [48]. Ma et al. [49] proposed a method base on MAD for model comparison in several tasks, including image quality, image aesthetics, and streaming video quality-of-experience. In addition, [50] investigated the generalizability of deep-learning based image classifiers by using a MAD-based method. However, little work base on MAD has been dedicated to performing model comparison in image captioning.

3. Proposed method

3.1. Method overview

The proposed framework is illustrated in Fig. 2. It mainly consists of three parts: 1) collecting a small scale but informative dataset, 2) performing human subjective experiments to get groundtruth captions of the informative dataset, and 3) effective performance comparison based on comparison dataset. To collect an informative dataset, we start with a large-scale real-world image database, then carefully pick out the samples that maximize the discrepancy among predicted captions from the compared ICMs. With these extreme samples, the divergence between different ICMs is largely amplified. Thus, model comparison becomes easier. We then perform a subjective annotation experiment on the collected image set. As the representative image set is small-scale, subjective annotation becomes affordable. To perform a thorough model comparison, we first conduct a one-vs-one local competition between every model pair on the informative dataset and then aggregate the local competition scores into a global result through rank aggregation. We will explain the details of each part in the following subsections.

3.2. Informative images collection

As shown in Fig. 2, our method first build a large-scale raw image set denoted by **R**. Then, as a direct model comparison on **R** is impractical due to lacking caption annotations, we attempt to collect a small scale representative image set **D** which requires only limited subjective annotations. To this end, we select the image samples that maximize the divergence between the predictions from different ICMs.

Assume C_i and C_j are two ICMs to be compared, we quantify the discrepancy between the outputs of C_i and C_j. The most straightforward way is to use one of the image captioning metrics, e.g., BLEU, METEOR, CIDEr, as the similarity function. However, the similarity kernels involved are inherently asymmetric. Therefore, these metrics are not optimal for judging caption similarities. To address this critical issue, we propose a N-Gram-based Similarity Metric (NGSM) for measuring caption similarities. It is formulated as follows,

$$\mathrm{NGSM}_n(\boldsymbol{s}_1,\boldsymbol{s}_2)=\frac{\mathrm{CI}_n(\boldsymbol{s}_1,\boldsymbol{s}_2)}{\mathrm{U}_n(\boldsymbol{s}_1)+\mathrm{U}_n(\boldsymbol{s}_2)-\mathrm{CI}_n(\boldsymbol{s}_1,\boldsymbol{s}_2)} \tag{1}$$

$$\mathrm{CI}_n(\boldsymbol{s}_1,\boldsymbol{s}_2)=\sum_{W_n\in\boldsymbol{s}_1} min(\mathrm{O}_{\boldsymbol{s}_1}(W_n),\mathrm{O}_{\boldsymbol{s}_2}(W_n)) \tag{2}$$

$$\mathrm{U}_n(\boldsymbol{s})=\sum_{W_n\in\boldsymbol{s}}\mathrm{O}_{\boldsymbol{s}}(W_n) \tag{3}$$

$$\mathrm{S}(C_i(\boldsymbol{x}),C_j(\boldsymbol{x}))=\sqrt[n]{\prod \mathrm{NGSM}_n(C_i(\boldsymbol{x}),C_j(\boldsymbol{x}))} \tag{4}$$

where $\boldsymbol{s}_*$ is captioning sentence, W_n is n-gram, $\mathrm{O}_{\boldsymbol{s}}(W_n)$ is count of n-gram W_n in sentence $\boldsymbol{s}$, $\mathrm{U}_n(\boldsymbol{s})$ is the sum of counts of each n-gram W_n in sentence $\boldsymbol{s}$ and $\mathrm{CI}_n(\boldsymbol{s}_1,\boldsymbol{s}_2)$ is the sum of counts of each n-gram W_n existing in both sentence $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. A smaller output of S indicates larger discrepancy. The NGSM is a permutation invariant function, *i.e.*,

$$\mathrm{S}(C_i(\boldsymbol{x}),C_j(\boldsymbol{x}))=\mathrm{S}(C_j(\boldsymbol{x}),C_i(\boldsymbol{x})) \tag{5}$$

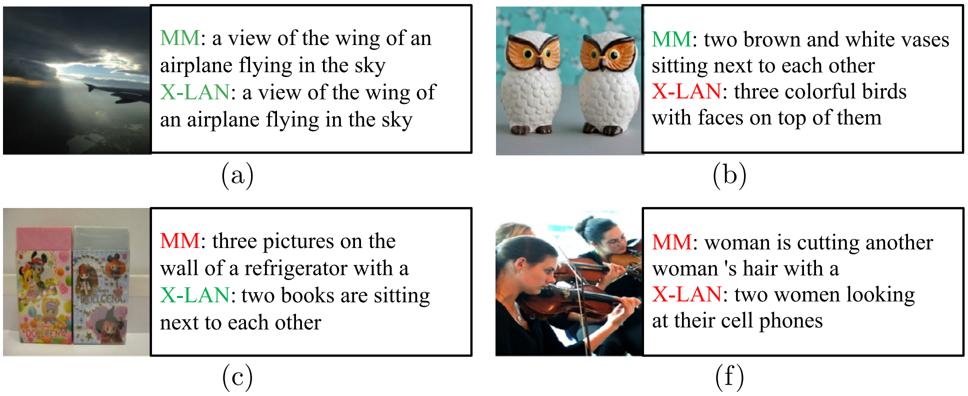

MM: a view of the wing of an airplane flying in the sky
X-LAN: a view of the wing of an airplane flying in the sky
(a)
MM: two brown and white vases sitting next to each other
X-LAN: three colorful birds with faces on top of them
(b)
MM: three pictures on the wall of a refrigerator with a
X-LAN: two books are sitting next to each other
(c)
MM: woman is cutting another woman 's hair with a
X-LAN: two women looking at their cell phones
(f)

Fig. 3. An illustration of informative images selected by our method. Each sub-fig demonstrates a selected image and the captions generated by MM and X-LAN for this image. The caption with green word model is right, and the caption with red word model is wrong. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

We then apply C_i, and C_j to all images in $\mathbf{R}$, and compare the predictions from both models according to Eqs. (1)–(5). The samples in $\mathbf{R}$ can be divided into three groups.

- Both C_i and C_j generate correct captions. This could happen when the test sample is too simple that existing methods easily capture the core image content and make precise predictions accordingly. When easy samples dominate $\mathbf{R}$, the representative samples will drown in the test set. As a result, the image captioning models would present over-optimistic performance, which is no good for effective model comparison. Therefore, these samples will be discarded in our model.
- Neither C_i nor C_j makes precise predictions. These samples have little effect on the model comparison between C_i and C_j and can be considered as uninformative images. However, these hard samples carry valuable information on the common weaknesses of existing ICMs. Therefore, it is worth deeply analyzing the functionality of existing ICMs on these samples.
- Either C_i or C_j generates correct caption. We consider that C_i and C_j show clear discrepancy on these samples, thus analysis and evaluation of these images can effectively point out strengths and weaknesses of C_i and C_j. These samples are informative for model comparison. Collecting these samples is the primary focus of our work.

With the proposed similarity function, the informative images obtained for C_i and C_j can be formulated as follow,

$$\mathbf{D}_{\{i,j\}} = \arg\min_{\boldsymbol{x}} \mathrm{S}(C_i(\boldsymbol{x}), C_j(\boldsymbol{x})) \tag{6}$$

Fig. 3 illustrates four examples. The captions are generated from MM [7] and X-LAN [8], respectively. Specifically, Fig. 3(a) belongs to the first case mentioned above. Figs. 3(b) and (c) belong to the second case, and Fig. 3(d) belong to the last case. It is clear that Fig. 3(a) and (d) carry little information to demonstrate which model is superior in terms of captioning accuracy since both models present identical accuracy. In contrast, (b) and (c) are more representative since both models present different characteristics, making model comparison feasible.

In summary, we construct $\mathbf{D}$ by picking out samples from $\mathbf{R}$ that maximize the discrepancy among predicted captions from C_i nor C_j. Towards this goal, we establish a pairwise test set $\mathbf{D}_{\{i,j\}}$ by selecting k images from $\mathbf{R}$ with k smallest similarities computed by Eq. (4). Given m models to be compared, We establish $(m(m-1))/2$ pairwise test sets for $(m(m-1))/2$ ICMs pairs. The test set $\mathbf{D} = \bigcup \mathbf{D}_{\{i,j\}}$ is constructed by integrating every pairwise test sets, and contains $k(m(m-1))/2$ test samples in total.

3.3. Subjective annotation experiment

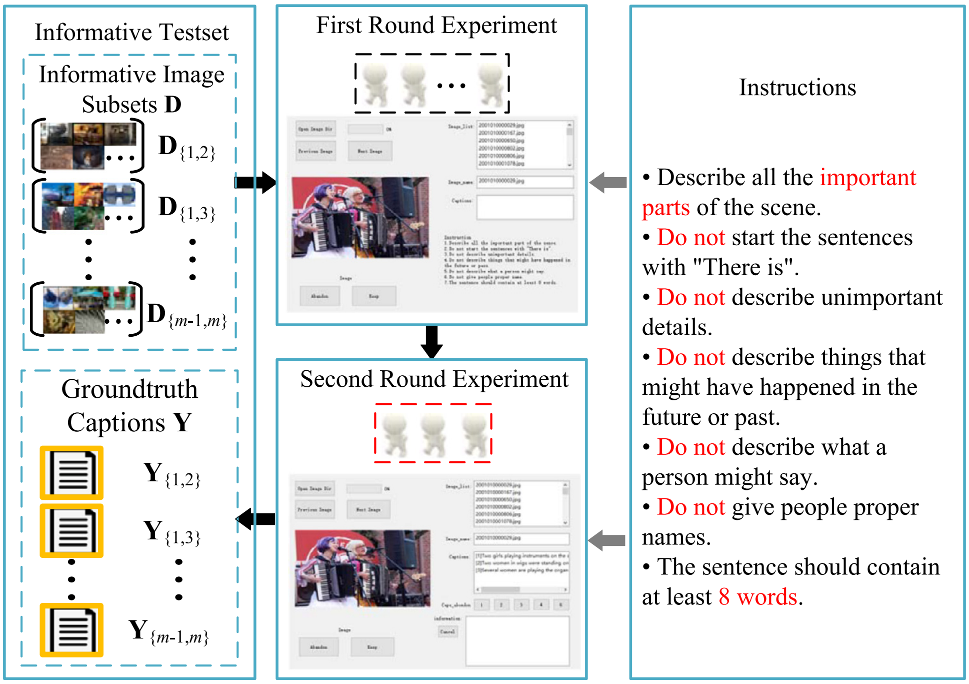
We further perform a subjective annotation experiment on $\mathbf{D}$ to collect the ground truth for image captions. As the size of $\mathbf{D}$ is relatively small, annotating the gold standard captions for $\mathbf{D}$ becomes feasible. To ensure the quality of subjective annotations, we follow the instructions proposed in [4] and construct a two-round human subjective experiment, as shown in Fig. 4. Specifically, we invite fifteen postgraduate students to take part in our first round subjective annotation experiment. Every annotator has participated in a simulated subjective annotation experiment on a separate dataset to learn annotation principles. Each image in $\mathbf{D}$ is annotated by at least three different annotators. Next, we hold a second round experiment to manually filter the annotations generated in the first round of experiment. Specifically, three postgraduate students are invited to be inspectors to check the annotations using the principle of majority, *i.e.*, an annotation will be discarded if it is considered wrong by two or more inspectors. An informative is proposed, which contains the annotations retained after the second round experiment and their corresponding images.

3.4. Performance comparison

After collecting ground truth captions for $\mathbf{D}$, we first perform comparison on pairwise ICMs, then aggregate all pairwise competition results to get the final global ranks for all ICMs. According to Kilickaya et al. [51], the CIDEr and SPICE has higher correlation with human consensus scores than other metrics, therefore we use CIDEr and SPICE when evaluating all competing ICMs on $\mathbf{D}$. Specifically, we calculate the metric scores of C_i and C_j on $\mathbf{D}_{\{i,j\}}$, respectively, and denote the corresponding scores as $p_{ij}^{\mathbf{e}}$ and $p_{ji}^{\mathbf{e}}$, where $\mathbf{e}$ indicates type of metric. Then we get the pairwise matrix $\boldsymbol{P}^{\mathbf{e}}$ by aggregating all $p_{ij}^{\mathbf{e}}$ and $p_{ji}^{\mathbf{e}}$. Based on the matrix $\boldsymbol{P}^{\mathbf{e}}$, we get the pairwise dominance matrix $\boldsymbol{F}^{\mathbf{e}}$ with $f_{ij}^{\mathbf{e}} = p_{ij}^{\mathbf{e}}/p_{ji}^{\mathbf{e}}$, where $f_{ij}^{\mathbf{e}}$ indicates the pairwise dominance degree of C_i over C_j on metric $\mathbf{e}$. The global rank of the m ICMs on the metric $\mathbf{e}$ is computed by using $\boldsymbol{F}^{\mathbf{e}}$ and rank method proposed in [52], which is denoted as $\boldsymbol{q}^{\mathbf{e}} \in \mathbb{R}^m$, and can be formulated as follow,

$$\boldsymbol{q}^{\mathbf{e}} = \lim_{t \to \infty} \frac{1}{t} \sum_{\alpha=1}^{t} \frac{(\boldsymbol{F}^{\mathbf{e}})^{\alpha} \mathbf{1}}{\mathbf{1}^{\mathrm{T}} (\boldsymbol{F}^{\mathbf{e}})^{\alpha} \mathbf{1}} \tag{7}$$

where $\mathbf{1}$ is an m-dimension vector of all ones. $\{q_i^{\mathbf{e}} \subset \boldsymbol{q}^{\mathbf{e}} | i = 1, 2, \dots m\}$ indicates the performance of C_i in model comparison among m ICMs on metric $\mathbf{e}$. The higher $q_i^{\mathbf{e}}$ is equivalent to better performance in model comparison. The procedure of our model comparison is summarized in Algorithm 1.



- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Fig. 4. Illustration of our human subjective experiment consists of two round experiments.

```
Algorithm 1: The proposed model comparison method.
Input: A raw image set $\mathbf{R}$, a group of ICMs $\mathbf{C}$=$\{C_b\}_{b=1}^{m}$, a
        similarity function BEST $S$.
Output: Six global ranking vectors $\boldsymbol{q}^{\mathbf{e}} \in \mathbb{R}^{m}$
$\mathbf{D} \leftarrow \varnothing, \boldsymbol{F}^{\mathbf{e}} \leftarrow \varnothing$.for $b = 1; b \leq m$ do
    Generate captions $\{C_b(\boldsymbol{x})\}_{b=1}^{m}, \boldsymbol{x} \in \mathbf{R}\}$;
end
for $i = 1; i \leq m - 1$ do
    for $j = i + 1; j \leq m$ do
        Compute the similarities using BEST;Establish the
        pairwise test set $\mathbf{D}_{\{i,j\}}$ for $C_i$ and $C_j$by selecting $k$
        images from $\mathbf{R}$ with $k$ smallest similarities computed
        by Eq. 4;$\mathbf{D} \leftarrow \mathbf{D} \bigcup \mathbf{D}_{\{i,j\}}$;
    end
end
Annotate the image caption groundtruth for $\mathbf{D}$;for
$i = 1; i \leq m - 1$ do
    for $j = i + 1; j \leq m$ do
        Compute the $\boldsymbol{p}^{\mathbf{e}}_{ij}$ by using metric function $\mathbf{e}$
    end
end
Compute the pairwise dominance matrix $\boldsymbol{F}^{\mathbf{e}}$ with $f^{\mathbf{e}}_{ij} = p^{\mathbf{e}}_{ij}/p^{\mathbf{e}}_{ji}$
for metric $informative$;Compute the global ranking vector $\boldsymbol{q}^{\mathbf{e}}$
for metric $\mathbf{e}$ using Eq. 7;return $\boldsymbol{q}^{\mathbf{e}}$.
```

```
Algorithm 2: Adding a new ICM into our competition.
Input: A raw image set $\mathbf{R}$, a new ICM $C_{m+1}$, the pairwise
        dominance matrix $\boldsymbol{F}^{\mathbf{e}} \in \mathbb{R}^{m\times m}$ for $\mathbf{C}$=$\{C_b\}_{b=1}^{m}$,a
        similarity function BEST $S$.
Output: Six global ranking vectors $\boldsymbol{q}^{\mathbf{e}} \in \mathbb{R}^{m+1}$
$\mathbf{D} \leftarrow \varnothing$.Generate captions $\{C(\boldsymbol{x})\}_{m+1}, \boldsymbol{x} \in \mathbf{R}\}$;for
$i = 1, j = m + 1; i \leq m$ do
    Compute the similarities using BEST;Establish the pairwise
    test set $\mathbf{D}_{\{i,m+1\}}$ for $C_i$ and $C_{m+1}$by selecting $k$ images from
    $\mathbf{R}$ with $k$ smallest similarities computed by Eq. 4;$\mathbf{D} \leftarrow \mathbf{D}$
    $\bigcup$ $\mathbf{D}_{\{i,m+1\}}$;
end
Annotate the image caption groundtruth for $\mathbf{D}$;for $\mathbf{e}$ in
metrics do
    $\boldsymbol{F}^{\mathbf{e}} \leftarrow \begin{pmatrix} \boldsymbol{F}^{\mathbf{e}} & 0 \\ 0^{\mathrm{T}} & 1 \end{pmatrix} \in \mathbb{R}^{(m+1\times m+1)}$for $i = 1, j = m + 1; i \leq m$
    do
        Compute the $\boldsymbol{p}^{\mathbf{e}}_{ij}$ by using metric function $\mathbf{e}$;Compute
        the $\boldsymbol{p}^{\mathbf{e}}_{ji}$ by using metric function $\mathbf{e}$;
    end
    Compute the pairwise dominance matrix $\boldsymbol{F}^{\mathbf{e}}$ with
    $f^{\mathbf{e}}_{ij} = p^{\mathbf{e}}_{ij}/p^{\mathbf{e}}_{ji}$;Compute the global ranking vectors $\boldsymbol{q}^{\mathbf{e}}$ for
    metric $\mathbf{e}$ using Eq. 7;return $\boldsymbol{q}^{\mathbf{e}}$.
end
```

In addition, if a new ICM takes part in model comparison with the m ICMs. We should only perform a subjective annotation experiment on mk image selecting base on the MAD between the new ICM and the m ICMs, and extend the dimensions of $\boldsymbol{P}^{\mathbf{e}}$, $\boldsymbol{F}^{\mathbf{e}}$ and $\boldsymbol{q}^{\mathbf{e}}$ from $\mathbb{R}^{(m\times m)}, \mathbb{R}^{(m\times m)}, \mathbb{R}^{m}$ to $\mathbb{R}^{(m+1\times m+1)}$, $\mathbb{R}^{(m+1\times m+1)}$, $\mathbb{R}^{m+1}$. Consequently, the global ranks of these $m+1$ ICMs can be obtained by using $\boldsymbol{F}^{\mathbf{e}}$ with $m+1$ dimension and Eq. (7). The procedure of adding a new ICM to our model comparison is summarized in Algorithm 2.

4. Experiments

4.1. Experiment setups

Construction of the Raw Image set R.

In the COCO image captioning dataset, the number of test samples is 5000. In order to efficiently evaluate the generalization ability of ICMs, we follow [53] to select the same 200 image classes. The distribution of **R** is illustrated in Fig. 5. Specifically, we crawl between 1000 and 2600 images directly from the Internet for each

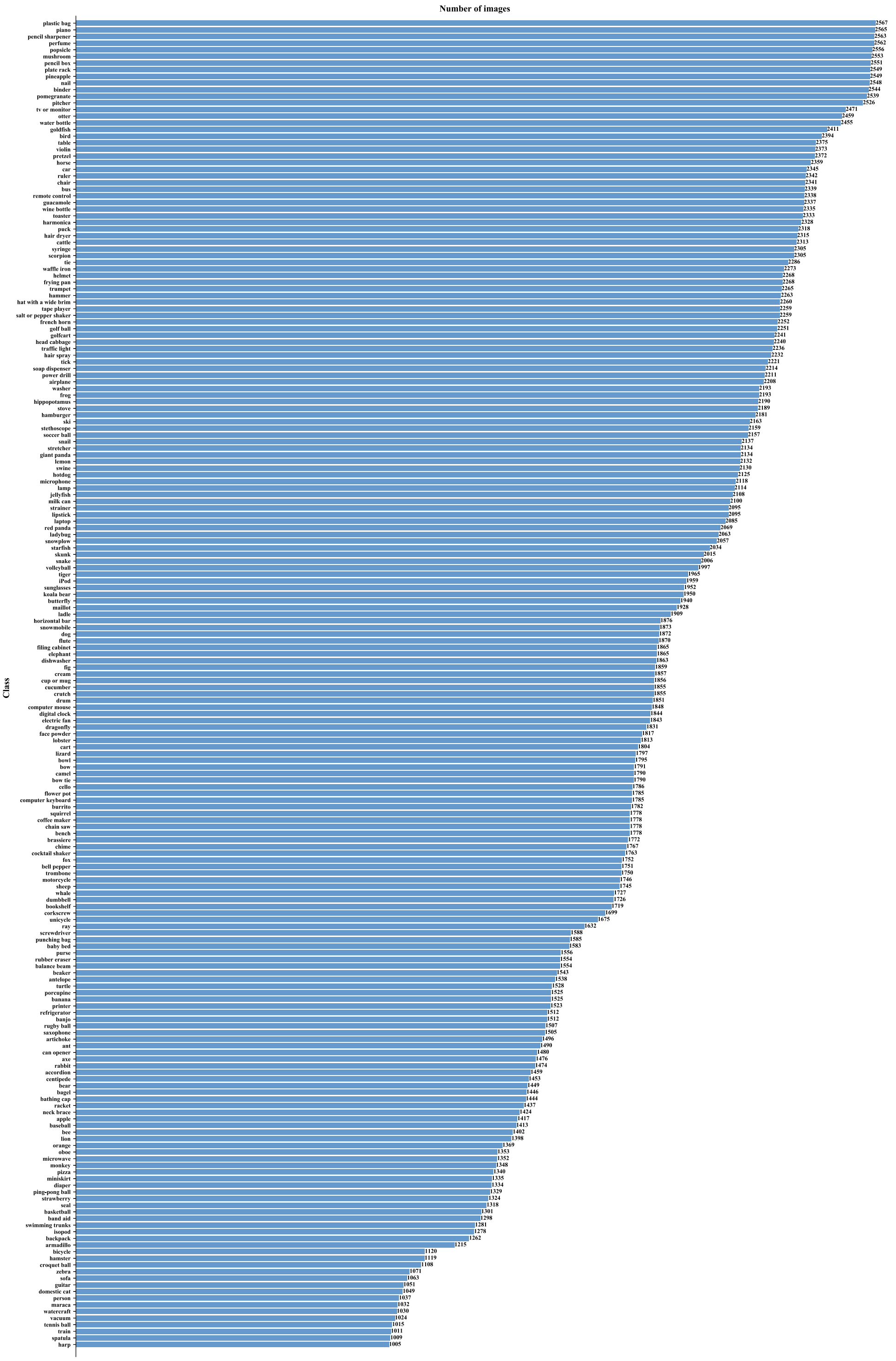


Fig. 5. Distribution of raw images set **R**.

class, and it is worth noticing that we do not set any constraints or post-processing. Consequently, the new image set **R** contains more than 370,000 images, which is at least seventy times larger than the COCO test set.

Competing ICMs.

In this work, we select 9 representative ICMs, including SCST [10], Stack-Cap [20], Top-Down [6], Disc [11], ORT [36], SGAE [21], AoA [36], MM [7] and X-LAN [8], which are published from 2017 to 2020. For a fair comparison, we directly use the publicly available code of each model. All these models trained on COCO use their default data pre-processing and post-processing settings. It is worth noting that performances of the competing ICMs are achieved by using pre-trained weights except Disc. As the pre-trained model provided from public code is not trained using the karpathy splits, we retrain a Disc model with COCO karpathy splits.

Among these models, SCST establishes a standard training paradigm of follow-up ICMs. MM and X-LAN are recently pro-

Table 1
The details about visual information encoder, language decoder and attention mechanism of the m competing ICMs, where BU, VE, LD, AMVE and AMLD are short for bottom-up attention, visual information extractor, language decoder, attention mechanism of visual information encoder and language decoder, respectively.

Models	VE	LD	AMVE	AMLD
X-LAN [8]	Object-Level	LSTM-based	Self-Attention	Self-Attention
MM [7]	Object-Level	Transformer-based	Self-Attention	Self-Attention
AoA [36]	Object-Level	LSTM-based	Self-Attention	Self-Attention
ORT [9]	Object-Level	Transformer-based	Self-Attention	Self-Attention
SGAE [21]	Object-Level	LSTM-based	Graph-based	Soft-Attention
Top-Down [6]	Object-Level	LSTM-based	-	Soft-Attention
Disc [11]	Object-Level	LSTM-based	-	Soft-Attention
Stack-Cap [20]	Image-Level	LSTM-based	-	Soft-Attention
SCST [10]	Image-Level	LSTM-based	-	-

Table 2
The statistics of subjective annotations. ‘Refs’ represents the number of annotated sentences per image and ‘Words’ denotes the number of tokens per annotated sentences. ‘Max Refs’ and ‘Min Refs’ are the maximum and minimum number of annotations of each image, respectively. ‘Max Words’ and ‘Min Words’ represent the maximum and minimum number of words in the annotate sentence among all annotations, respectively.

Refs	Words	Max Refs	Min Refs	Max Words	Min Words
2.75	8.42	4	2	37	6

posed models and achieve similar performances (all automatic evaluation metrics) on the COCO test set. Also, the contributions of objects spatial relationship (*i.e.*, ORT), RL learning (*i.e.*, Disc) and Scene Graph (*i.e.*, SGAE) are worth being investigated. We add Top-Down to the model comparison to see the influence of bottom-up and top-down attention on generalizability. In addition, we investigate the generalizability of the modified self-attention ICMs (*i.e.*, AoA). The visual information encoder, language decoder, attention mechanism used in visual information encoder and language decoder of the m competing ICMs are listed in Table 1.

Subjective Annotations.

In our subjective annotation experiment, we set k=70. Consider the overlap of samples between different $\mathbf{D}_{\{i,j\}}$, there are $\frac{(70\times9\times8)}{2}=2520$ need subjective annotation at most. if an image $\boldsymbol{x}\in\mathbf{D}_{\{i,j\}}$ is considered indescribable by two or more annotators, the image will be replaced by the $(k+1)$th smallest similarity sample for C_i and C_j. The statistics of subjective annotations are listed in Table 2.

4.2. Experimental results

Overall Performance.

Fig. 6 shows performances of the competing ICMs on the COCO test set and $\mathbf{D}$, and it should be noticed that the performances of ath ICM are achieved by using all images from $\left\{\mathbf{D}_{\{a,j\}}|j=1,2,\ldots,m,j\neq a\right\}$. Comparing performances of the competing ICMs on the COCO test set, they achieve lower and more discriminative CIDEr and SPICE scores on $\mathbf{D}$. Specifically, comparing their CIDEr scores, which at least 113.9, on the COCO test set, the competing ICMs achieve at most 26.7 CIDEr score. For SPICE, we observe at least drops 2/3 on $\mathbf{D}$ compare to performances of the competing ICMs achieve on the COCO test set. Moreover, SGAE outperforms ORT by one time on $\mathbf{D}$ while obtains similar results as ORT on the COCO test set. Thus, it indicates that competing ICMs may suffer from overfitting and may differ in generalization ability even when obtaining close performances on the COCO test set. This supports the necessity of a generalization ability test for validating ICMs performances among the model comparison for image captioning.

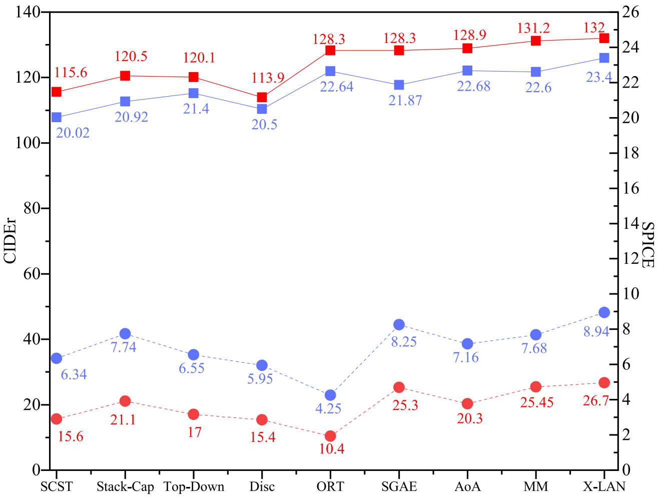


Fig. 6. Comparison of competing ICMs performances between COCO test set and **D**. Red and blue squares indicate CIDEr and SPICE scores on the COCO test set, respectively. Red and blue dots indicate CIDEr and SPICE scores on **D**, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Pairwise Comparison Results.

Figs. 7 and 8 are visualization maps of pairwise dominance matrices $\boldsymbol{F}^{\text{CIDEr}}$ and $\boldsymbol{F}^{\text{SPICE}}$, respectively. The darker color indicates a higher pairwise dominance degree, and the light color responds to a lower pairwise dominance degree in these maps. From Figs. 7 and 8, it is interesting to find the ORT gains lower pairwise dominance degrees in all pairwise comparisons. The self-attention based competing ICMs, such as AoA, X-LAN and MM, demonstrate dominant ability on this head-to-head comparison.

Global Ranking Results.

In this section, we analyze the global ranking results of the competing ICMs. The global ranking scores are shown in Fig. 9. We can observe the consensus between CIDEr and SPICE ranking results. We also list global ranking results in Table 3 and Table 4. By analyzing the ranking results in Tables 3 and 4, we summarize several interesting finds of our model comparison.

Firstly, we can observe ORT presents a poor performance in our model comparison, although its performances on par with those of AoA, MM on the COCO test set. Considering that AoA, MM and ORT have the similar self-attention based visual informative encoder and ORT and MM have the similar Transformer based language decoder. The poor performance may be caused by the original Transformer based connectivity between visual information encoder and language decoder used in ORT. This indicates using only high-level object features to model inter-modal relationship may

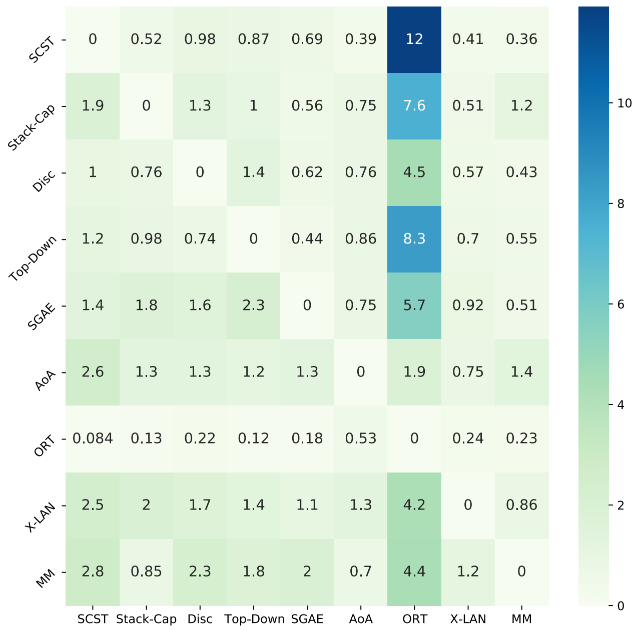


Fig. 7. Pairwise dominance matrix on CIDEr.

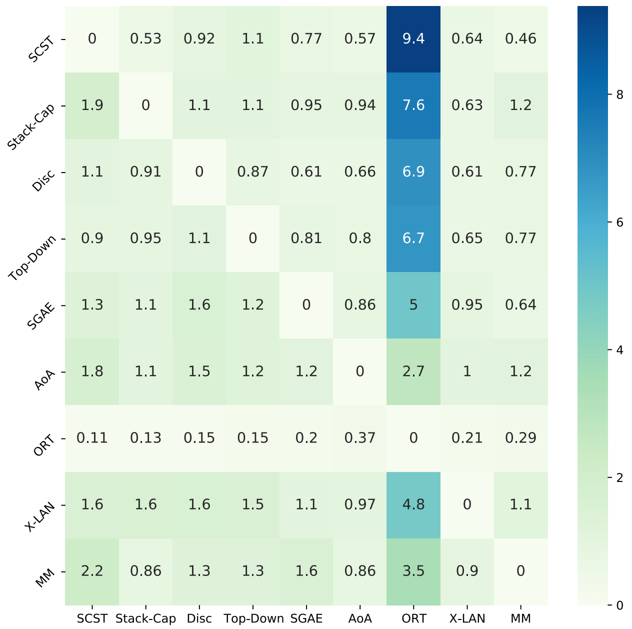


Fig. 8. Pairwise dominance matrix on SPICE.

Table 3
Global ranking results on CIDEr, where C is short for CIDEr scores. A smaller rank indicates better performance. The Δ indicates difference between ranks in performance comparisons on COCO Karpathy test split and ranks in our model comparison.

Models	COCO C	C Rank	Global Rank	Δ Rank
X-LAN [8]	132.0	1	2	-1
MM [7]	131.2	2	1	+1
AoA [36]	128.9	3	3	0
SGAE [21]	128.3	4	4	0
ORT [9]	128.3	5	9	-4
Stack-Cap [20]	120.5	6	5	+1
Top-Down [6]	120.1	7	6	+1
SCST [10]	115.6	8	8	0
Disc [11]	113.9	9	7	+2

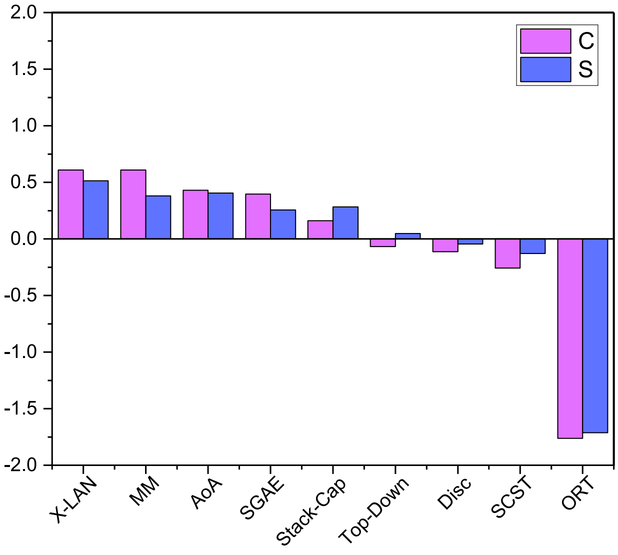


Fig. 9. Global ranking results of the competing ICMs.

Table 4
Global ranking results on SPICE, where S is short for SPICE scores. A smaller rank indicates better performance. The Δ indicates difference between ranks in performance comparisons on COCO Karpathy test split and ranks in our model comparison.

Models	COCO S	S Rank	Global Rank	Δ Rank
X-LAN [8]	23.40	1	1	0
AoA [36]	22.68	2	2	0
ORT [9]	22.64	3	9	-6
MM [7]	22.60	4	3	+1
SGAE [21]	21.87	5	5	0
Top-Down [6]	21.40	6	6	0
Stack-Cap [20]	20.92	7	4	+3
Disc [11]	20.50	8	7	+1
SCST [10]	20.02	9	8	+1

reduce the generalization ability of an ICM trained existing image captioning datasets. More analysis is provided in Section 4.3.

Secondly, the self-attention mechanism based ICMs, such as MM, X-LAN and AoA, still outperforms other attention mechanism based ICMs in our model comparison. This verifies the modelling ability of self-attention based modules for inter-modal and intra-modal data.

Finally, since the object-level visual informative encoder is considered can provide stronger support than the image-level visual informative encoder for the generalization ability of ICMs, and SCST does not maintain its superiority on the CIDEr metric in our comparison. Nevertheless, it is interesting to find Stack-Cap, which employs an image-level visual informative encoder, outperforms Top-Down and SGAE. The Stack-Cap may benefit from its coarse-to-fine language decoder. This may indicate the multistage structure is a promising way to improve ICM performance.

Impact of k.

In this section, we investigate the influence of the hyper-parameter k in our model comparison method. By using the top-70 ranking results as references, we calculate SRCC value between top-k ranking results, where $k = \{5, 6, \dots, k\}$ and top-70 ranking results on SPICE and CIDEr. As illustrated in Fig. 10, the SRCC values on two metrics are larger than 0.90 when $k > 46$. Based on the consistent results among all metrics, we can consider the comparison results are stable. This is strong support for us to set the k to 70 because increase the k will not change the ranking results.

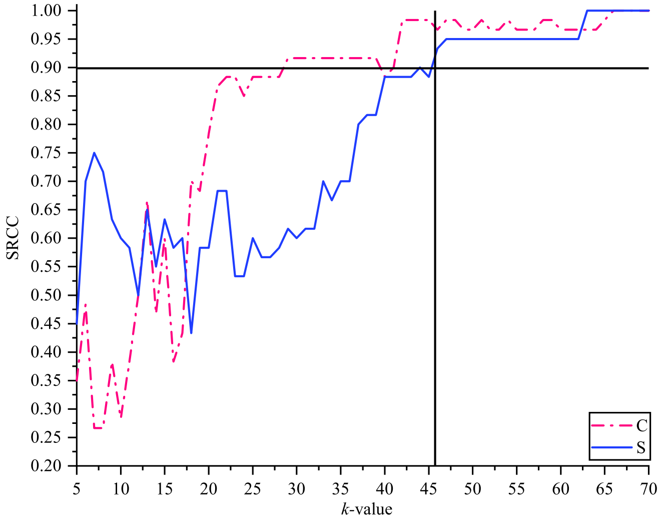


Fig. 10. The SRCC values between top-70 ranking and other top-k rankings with $k = \{5, 6, \ldots, 69\}$ on different metrics, where C and S are short for CIDEr and SPICE scores.

4.3. Discussion

Impact of the Connectivity for Transformer based ICMs. To better demonstrate the impact of the multi-scale visual features, we analyze duplicated captions, which contains at least one duplicated word generated by the competing ICMs. We consider at least one word appearing three or more times in a row in a caption to be the duplicated caption. An image with a duplicated caption can easily be selected as an informative image since the significant difference between duplicated and regular captions. Some informative images and duplicated captions are shown in Fig. 11. An ICM gains poor evolution scores when it generates duplicated caption for an image. We find that the ICMs intend to generate the duplicated caption when images are containing multiple similar targets or one target with multiple similar parts. As shown in Fig. 12, ORT produces duplicated captions with at least six times the probability of the other competing ICMs. Its caption contains more than five times the classes of duplicated words than other competing ICMs. Actually, ORT proposes a modified Transformer encoder for modelling geometry relationships of detected objects and uses the decoder of Transformer as a language decoder. To study the duplicated captions of ORT, we retrain a baseline model, denoted as ORT$^{\$}$, without the geometry feature using the official opening code. As illustrated in Fig. 12, ORT produces about 900 fewer samples of duplicated captions than those generated by ORT$^{\$}$ in raw images set **R**. It indicates that the geometry features can alleviate duplicated caption for ORT. Compares to ORT, MM, which proposes a modified Transformer based ICMs containing mesh-like connectivity between visual information encoder and language decoder, achieves the best performance about the duplicated caption in the

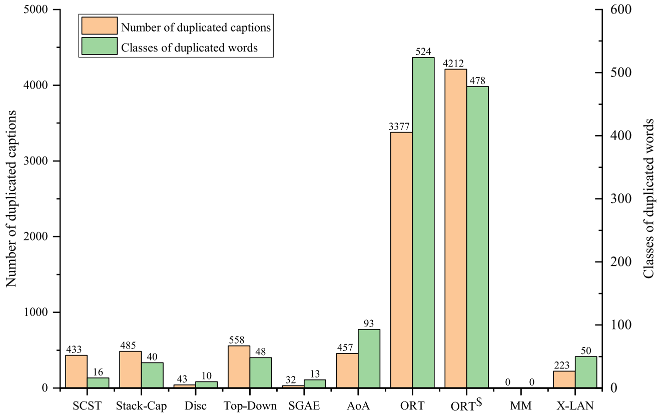
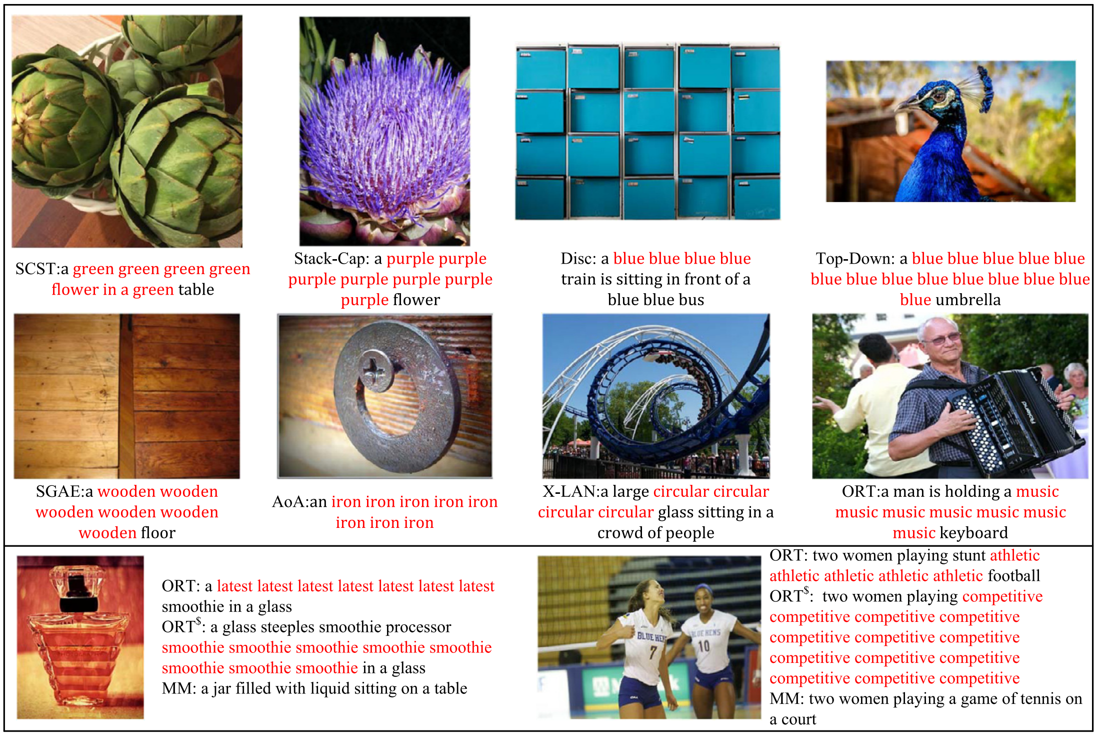


Fig. 12. Statistics of duplicated caption for the competing ICMs on raw images set **R**.



SCST:a green green green green flower in a green table

Stack-Cap: a purple purple purple purple purple purple purple flower

Disc: a blue blue blue blue train is sitting in front of a blue blue bus

Top-Down: a blue blue blue blue blue blue blue blue blue blue blue blue blue umbrella

SGAE:a wooden wooden wooden wooden wooden wooden floor

AoA:an iron iron iron iron iron iron iron iron

X-LAN:a large circular circular circular circular glass sitting in a crowd of people

ORT:a man is holding a music music music music music music music keyboard

ORT: a latest latest latest latest latest latest latest smoothie in a glass
ORT$^{\$}$: a glass steeples smoothie processor smoothie smoothie smoothie smoothie smoothie smoothie smoothie in a glass
MM: a jar filled with liquid sitting on a table

ORT: two women playing stunt athletic athletic athletic athletic athletic football
ORT$^{\$}$: two women playing competitive competitive competitive competitive competitive competitive competitive competitive competitive competitive competitive competitive
MM: two women playing a game of tennis on a court

Fig. 11. Samples of duplicated captions and images from **D**.

competing ICMs. The mesh-like connectivity allows each decoding layer of MM to simultaneously model inter-modal relationships between high-level or low-level object features and word features. Thus, modelling multiple inter-modal relationships may alleviate the problem of duplicated caption, as shown in Fig. 11.

Analysis of Attention Mechanism in the Competing ICMs. Considering the competing ICMs based on a two-stream paradigm containing a visual information encoder and a language decoder, we analyze the effect of attention mechanisms of visual information encoders and language decoders on generalization ability separately. For the visual information encoder, the image-level extractor without attention is proven incompetent in our model comparison. The bottom-up attention based extractor implemented a Faster RCNN becomes the dominant component as it can provide feature representation of salient and contextual regions. Actually, the object features detected by the bottom-up attention based extractor can be seen as a long sequence of visual representations. The self-attention visual information encoders are good at modelling global dependencies of long sequences. For language decoders of the competing ICMs, the self-attention mechanism are implemented as submodules in LSTM based decoders [8,36] and Transformer based decoders [7,9]. The self-attention attention submodules explicitly model the inter-modal relationship between each word embedding and each object feature. In addition to the inter-modal relationship modelling submodule, the Transformer based decoders contain the intro-modal relationship submodules that fed the word embedding. Compared to soft-attention based language decoders, the self-attention ones can provide better word-object and words relationship modelling. In addition, multistage decoder [20] achieves better performance than other soft-attention based decoders. It may be benefited from the multi-model structure of the multistage decoder. Different models in the multistage decoder focus on different regions of the image, and the accuracy of the focus increases with the number of sub-networks. Specifically, based on the "coarse" decoder, the "fine" decoder is easier to pays more attention to image space than the "coarse" one ignores. Therefore, it makes the multistage decoder generate more detailed captions.

4.4. Future direction

Based on the analysis presented, we can provide three promising directions for the image captioning task.

Novel Visual Information Extractor. Visual information extractor is a fundamental and decisive component for image captioning. Compared to the image-level extractor implemented by a pretrained CNN, the object-level extractor (i.e., a Faster R-CNN pretrained on Visual Genome) obtains remarkable performance in standard image caption evaluation and our model comparison. Nonetheless, the ICMs are currently still suffering from the erroneous results of the object-level extractor. Therefore, inspired by the success of Vision Transformer models [54] in CV tasks, including object detection, it is worth proposing a Vision Transformer based visual information extractor.

Novel ICM Architecture. The different performances between MM and ORT in our modal comparison demonstrates the importance of model architecture for ICMs. Motivated by the success of Neural Architecture Search (NAS) [55] in CV and NLP tasks. It is worth to study to proposed a new ICM based on NAS.

Datasets Bias Resistant Metric. As shown in Fig. 6, the competing ICMs achieve much more poorly evaluation performances on D than the standard test set of the COCO caption dataset. In principle, datasets bias in human subjective annotations is one of the majority causes. In this case, designing a bias resistant evaluation metric is a necessary and promising research direction.

5. Conclusion

In this paper, we conduct a quantitative comparison of the main image captioning models on performance in real-world scenarios. We construct a representative dataset by crawling test images over the Internet only with the keyword constraint. By doing so, the quantity and content diversity of the images can be greatly guaranteed. Furthermore, we adapt maximum discrepancy competition to the compared ICMs on the newly constructed test dataset. It automatically picks out a small-scale informative images that highlight the characteristics and differences of the existing captioning models. Compared to the standard paradigm that testifies captioning models on a large-scale dataset with enormous annotation efforts, our method only requires a small-scale and low-cost subjective experiment to collect the annotations, making model comparison convenient and highly effective. We believe our study is complementary to other relevant studies and reveals promising directions for developing image captioning in the future. Based on the analysis of our model comparison results, we argue that obtaining discriminative visual representations, designing robust ICM architectures, and proposing bias-resistant metrics are promising research directions.

Declaration of Competing Interest

The authors declare that this paper is original and has no conflict of interest.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0109301, in part by the Jiangxi Natural Science Foundation of China under Grant 20202ACB202007, in part by the National Natural Science Foundation of China under Grant 62161013, in part by the Key Research and Development Program of Jiangxi Province under Grant 20203BBE53033, and in part by the Foundation of Jiangxi Provincial Department of Education under Grant GJJ200535.

References

[1] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2016) 652–663.
[2] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. Zitnick, Microsoft COCO captions: data collection and evaluation server, arXiv preprint arXiv:1504.00325(2015).
[5] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognit. 90 (2019) 285–296.
[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
[7] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.
[8] Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980.
[9] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, in: Advances in Neural Information Processing Systems, 2019, pp. 11137–11147.
[10] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, 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.
[11] R. Luo, B. Price, S. Cohen, G. Shakhnarovich, Discriminability objective for training descriptive captions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6964–6974.

[12] J. Chen, Q. Jin, Better captioning with sequence-level exploration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10890–10899.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[14] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[15] M. Denkowski, A. Lavie, Meteor universal: language specific translation evaluation for any target language, in: Proceedings of the Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[16] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.

[17] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: semantic propositional image caption evaluation, in: European Conference on Computer Vision, 2016, pp. 382–398.

[18] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, H. Lu, Normalized and geometry-aware self--attention network for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10327–10336.

[19] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[20] J. Gu, J. Cai, G. Wang, T. Chen, Stack-captioning: Coarse-to-fine learning for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[21] X. Yang, K. Tang, H. Zhang, J. Cai, Auto-encoding scene graphs for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685–10694.

[22] S. Li, G. Kulkarni, T. Berg, A. Berg, Y. Choi, Composing simple image descriptions using web-scale n-grams, in: Proceedings of the Conference on Computational Natural Language Learning, 2011, pp. 220–228.

[23] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: generating sentences from images, in: European Conference on Computer Vision, 2010, pp. 15–29.

[24] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2891–2903.

[25] V. Ordonez, G. Kulkarni, T. Berg, Im2Text: describing images using 1 million captioned photographs, Adv. Neural Inf. Process. Syst. 24 (2011) 1143–1151.

[26] A. Gupta, Y. Verma, C. Jawahar, Choosing linguistics over vision to describe images, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2012.

[27] P. Kuznetsova, V. Ordonez, T.L. Berg, Y. Choi, TreeTalk: composition and compression of trees for image descriptions, Trans. Assoc. Comput.Linguist. 2 (2014) 351–362.

[28] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks (m-RNN), in: International Conference on Learning Representations, 2015.

[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[30] R. Girshick, J. Donahue, T. Darrell, 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.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[33] L. Li, S. Tang, L. Deng, Y. Zhang, Q. Tian, Image caption with global-local attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 4133–4139.

[34] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, Learning visual relationship and context-aware attention for image captioning, Pattern Recognit. 98 (2020) 107075.

[35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.

[36] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4634–4643.

[37] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, P. Anderson, nocaps: novel object captioning at scale, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8948–8957.

[38] S. Chen, Q. Jin, P. Wang, Q. Wu, Say as you wish: fine-grained control of image caption generation with abstract scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9962–9971.

[39] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[40] Y. Fang, X. Zhang, F. Yuan, N. Imamoglu, H. Liu, Video saliency detection by gestalt theory, Pattern Recognit. 96 (2019) 106987.

[41] Y. Fang, C. Zhang, X. Min, H. Huang, Y. Yi, G. Zhai, C.-W. Lin, DevsNet: deep video saliency network using short-term and long-term cues, Pattern Recognit. 103 (2020) 107294.

[42] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[43] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: European Conference on Computer Vision, 2018, pp. 3–19.

[44] A.G. Roy, N. Navab, C. Wachinger, Concurrent spatial and channel squeeze & excitation in fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 421–429.

[45] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[47] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[48] Z. Wang, E.P. Simoncelli, Maximum differentiation (mad) competition: amethodology for comparing computational models of perceptual quantities, J. Vis. 8 (12) (2008) 8.

[49] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, L. Zhang, Group maximum differentiation competition: model comparison with few samples, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 851–864.

[50] H. Wang, T. Chen, Z. Wang, K. Ma, I am going MAD: maximum discrepancy competition for comparing classifiers adaptively, in: International Conference on Learning Representations, 2020.

[51] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, E. Erdem, Re-evaluating automatic metrics for image captioning, in: Proceedings of the European Chapter of the Association for Computational Linguistics, 2017, pp. 199–209.

[52] T.L. Saaty, L.G. Vargas, Inconsistency and rank preservation, J. Math. Psychol. 28 (2) (1984) 205–214.

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, pp. 1254–1259.

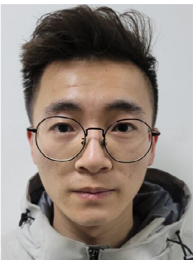
[55] T. Elsken, J.H. Metzen, F. Hutter, Neural architecture search: a survey, J. Mach. Learn. Res. 20 (1) (2019) 1997–2017.

Boyang Wan received his B.E. degree in Electronic and Information Engineering from Polytechnic Institute of Jiangxi Science and Technology Normal University in 2012, and the M.S. degree in signal and information processing with the school of communications and electronics, Jiangxi Science and Technology Normal University, Nanchang, China, in 2018. He is pursuing the Ph.D. degree in Jiangxi University of Finance and Economics, Nanchang, China. His research interests include image processing, image captioning and pattern recognition.

Wenhui Jiang Wenhui Jiang received the B.E. from Nanchang University, Nanchang, China, the M.S. and the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China. He was a senior engineer with Alibaba Damo Academy from 2017 to 2019, and a visiting student at University of California, Santa Barbara from 2015 to 2016. He is a lecturer with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His current research interests include large-scale image understanding and cross-media analysis.

Yuming Fang received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, computer vision, and 3D image/video processing. He serves as an Associate Editor for IEEE TMM. He serves on the Editorial Board of Signal Processing: Image Communication

Minwei Zhu received the B.E. degree in Broadcasting Television Engineering from Communication University of China, Beijing, China, in 2017. He is pursing a master's degree in Jiangxi University of Finance and Economics. His research interests including image processing, image captioning.



Qin Li received his B.E. degree in Computer Science from Jiangxi University of Finance and economics in 2019. He is studying for a master's degree in Jiangxi University of Finance and economics. His research interests including image processing, image captioning.

Yang Liu received the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was a visiting scholar at National University of Singapore and City University of Hong Kong, and is currently the director of Intelligent Research Institute (IRI), SANY Heavy Industry Co., Ltd. He has worked in Huawei 2012 Lab and Alibaba Damo Acadamy, focusing on computer vision, machine learning and pattern recognition.