MANIC: Multi-Attention Network for **Image Captioning**

GMIK Rupasinghe, Kanishka Karunarathne, Chainbia Najet, Walid Hamdi
University of Sousse, Sousse National School of Engineers, B.P 264 Cite Erriadh, 4023 Sousse, Tunisia

Abstract—In this paper, we present a novel model Multi-Attention Network that can unify different attention models from existing literature. The proposed model significantly outperforms state-of-the-art methods on two benchmark datasets, Flickr30k and MSCOCO, in terms of SPICE, which correlates best with human judgments. Experiments shows that the refined attention mechanisms perform better than their existing versions and the merging gate contributes essentially to the overall improvement.

Index Terms—**Image captioning**

I. Introduction

Aiming to generate a description of an image in natural language form, **image captioning** is a prevalent task in both natural language processing and computer vision. The task is very challenging while pragmatic. On the one hand, it consists of three subtasks: identifying the objects in the image, analyzing their attributes and relationships, and describing them in a fluent sentence, each of which contributes to the difficulty of this task. On the other hand, it incorporates two mainstream fields in artificial intelligence, that is, natural language processing and computer vision. Furthermore, it has a wide range of applications, including text-based image retrieval, helping visually impaired people see, human-robot interaction [1], etc.

Some recent works have explored models based on the encoder-decoder frame-work and shown success in **image captioning**. According to the pivot representation, they respectively fall into the category of models based on visual information [2]–[6], and models based on conceptual information [7]–[9]. The later explicitly provides keywords extracted from the image (e.g. dog, sit, red) to the decoder instead of the image itself, which is more effective in **image** in describing the image in a comprehensive way and tend to describe merely a subregion.

In this work, we get the best of both worlds by integrating visual attention and semantic attention in order to generate captions that are both detailed and comprehensive. We propose an Adaptive Image-Keyword Merging Network as the decoder to guide the information flow from visual features and extracted keywords to generate captions. At each time step, the decoder first extracts visual information from the image. Then, it picks out the most probable keywords for the current time step. Finally, it attends differently to the visual information and the conceptual information to generate the output word. Hence, the model can adaptively merge the two kinds of information, leading to desirable results in **image captioning**. In summary, we mainly make the following contributions:

- We propose a novel model that can effectively merge visual information and semantic information to generate cohesive captions that are both detailed and comprehensive. We refine both the visual attention and the semantic attention to address their existing weaknesses. Then, we introduce the score-based merging gate that effectively integrates and balances the two kinds of information.
- The proposed model significantly outperforms state-of-the-art methods on two benchmark datasets, Flickr30k and MSCOCO, in terms of SPICE, which correlates best with human judgments. Experiments shows that the refined attention mechanisms perform better than their existing versions and the merging gate contributes essentially to the overall improvement.

II. Related Works

In recent years, a large number of neural model systems

captioning according to evaluations on benchmark datasets. However, the models based on conceptual information have a major drawback in associating the details in image, such as the position (e.g. behind, on), the state (e.g. close, open), and the attribute (e.g. color,size), because the visual words are inherently unordered in semantics. Figure 1 shows an example. For semantic attention, although the word open is provided as a visual word, due to the insufficient use of visual information, the model gets confused about what objects open should be associated with and thus excludes the word open when generating the caption. It's even likely that the model may associate the details incorrectly, which is the case for the position of the dog. In contrast, models based on visual information are more accurate in details but have difficulty

Page 2

relevant image regions at each time step. Most recently, visual features based on predicted bounding boxes [13], [14] are utilized to improve the quality of captions. The advantages are that the attention mechanism no longer needs to find the relevant generic regions but instead find relevant bounding boxes that orientating objects. These bounding boxes can then serve as a semantic guide. However, the drawback lies in predicting bounding boxes, which requires large datasets [21] and complex models [22]. Intuitively, it's also difficult to define a bounding box corresponding to keywords that describe attributes or relationships [7], [20], [23].

Semantic attention models pay attention to a set of predicted semantic concepts [7]–[9], [24]. The semantic concepts are the most frequent words in the ground-truth captions in training set, and the word-extractor can be trained using various methodologies but typically only on the given **image captioning** dataset. This kind of approach can be seen as an extension of the earlier proposed template-based slotting-filling approaches [25].

However, few works delve into the issue of combining these two kinds of attention models to take advantage of both. Firstly, due to limited expression capacity of visual features, information provided to the decoder is less comprehensive. Additionally, the extracted semantic concepts are unordered, making it hard for the decoder to portray the details of the objects precisely.

To the best of our knowledge, we are one among the first endeavors that focus on combining the visual attention and the semantic attention efficiently to offset their drawbacks and make full use of their merits. The visual attention is designed to focus on the attributes and the relationships of the objects, while the semantic attention only attends to the extracted keywords. The combination is controlled by the importance-based merging mechanism that decides at each time step which kind of information should be paid more attention to. The goal is to generate image captions that are both detailed and comprehensive.

have been proposed for **image captioning**. Neural models based on the encoder-decoder framework have drawn increasing attention in several multi-discipline tasks, such as neural image/video captioning (NIC) and visual question answering (VQA) [2], [6], [11], [12]. State-of-the-art approaches relying on neural networks [13], [14],incorporate the attention mechanism in machine translation [15] to generate coherent image captions. According to what they attend to, the models can be classified into visual attention models and semantic attention models.

Visual attention models pay attention to visual features captured by CNNs. CNNs are typically pre-trained on the image recognition task to extract general visual information [16]–[20]. The visual attention is intended to find the most

a dimension of $2048 \times 7 \times 7$, so the dimension of a_i is 2048. Following [27], the global image feature can be obtained by:

$$a_g = \frac{1}{k} \sum_{i=1}^{k} a_i \tag{2}$$

where a_g is the global image feature. For modeling convenience, we use a single layer perceptron with rectifier activation function to transform each image feature vector a_i into a new vector v_i with dimension g. For the global image feature a_g, we transform it into a new vector v_g with dimension e:

$$v_g = \mathrm{ReLU}(W_g a_g) \tag{3}$$

$$v_i = \mathrm{ReLU}(W_a a_i) \tag{4}$$

where $W_g \in R_{e \times 2048}$, $W_a \in R_{g \times 2048}$ are learnable parameters, and e denotes the dimension of word embeddings.After the calculations, we get a series of new visual feature vectors $V = \{v_1, v_2, ..., v_k\}$, $v_i \in R_g$, and the global image feature vector $v_g \in R_e$.

B. Keyword Extractor

Considering limited expression capacity of the visual features, it is hard for the traditional decoder to describe the objects in the image comprehensively. An advance in **image captioning** is to directly provide the decoder with the semantic concepts in the image, so that the decoder is equipped with an overall perspective of the image. The semantic concepts can be objects (e.g. person, car), attributes (e.g. off, electric), and relationships (e.g. using, sitting). In this work, we choose the most frequent words, referred to as keywords, from the ground-truth captions in training set to represent the semantic concepts,. The keyword extractor concludes a list of candidate keyword embeddings $T = \{w_1, w_2, ..., w_m\}$, $w_i \in R_e$ from the image. Following common practice [7], [8], we adopt the weakly-supervised approach of Multiple Instance Learning [28] to build a keyword extractor. Due to limited space, please refer to [7] for detailed explanation.

III. Approach

The proposed model comprises three parts: an image encoder, a topic extractor, and an adaptive merging decoder.

A. Image Encoder

Given an input image, the image encoder translates the image into a series of visual feature vectors $A = \{a_1, a_2, ..., a_k\}$. Each feature represents a certain aspect of the image. The visual features serve as a guide for the decoder to describe the objects in the image. ResNet152 [27], a CNN-based model, is used to obtain visual features. The output of the last convolutional layer can be calculated as below:

$$A = CNN(I) \tag{1}$$

where I is the input image. Specifically, the spatial feature outputs of the last convolutional layer of ResNet152 [27]have

Existing work relies on attribute words and relationship words to provide visual information to the decoder. However, it not only complicates the extracting procedure but also contributes little to overall performance. As such words are specific to certain objects but are provided to the decoder unorderly, the decoder is likely to associate the attributes with the objects randomly when the image contains a lot of objects. In comparison, our model has visual information as input so we expect that the decoder can refer to the image for such information about details.

C. Adaptive Image-Keyword Merging Decoder

The decoder part of the proposed adaptive image-keyword merging network is essential in the way that it translates the extracted visual features and keywords into their corresponding caption. The decoder includes two LSTMs [29], with each used for one kind of attention. Following the practice of both two LSTMs take the word embedding vector $w_{e_t} \in R_e$,

Page 3

concatenated with the global image feature vector $v_g \in R_e$, as input, which is denoted as $x_t = [w_{e_t}; v_g]$. We use $h_t \in R_{d_1}$ to indicate the current hidden state of the first LSTM, and use $h_t \in R_{d_2}$ to indicate the current hidden state of the second LSTM. At each time step, the decoder combines the current textual state (the input and the last hidden state), the attentive visual information and the attentive semantic information to generate an output word. The goal is achieved by three modules, the visual attention, the semantic attention, and the adaptive merging gate.

a) Visual Attention: The visual attention attends to relevant regions in the image based on current hidden state of the first LSTM decoder. In existing work [18], the current hidden state h_t of the first LSTM is used to computate the visual attention:

$$Z_t = \tanh(W_{Z,V} V \oplus W_{Z,h} h_t) \tag{5}$$

$$\alpha_t = \text{softmax}(Z_t w_{\alpha,Z}) \tag{6}$$

where $W_{Z,V} \in R_{k \times g}$, $W_{Z,h} \in R_{k \times d_1}$, $w_{\alpha,Z} \in R_k$ are learnable parameters.[1] We denote the matrix-vector addition as $\oplus$, which is defined by adding the vector to each column of the matrix. $\alpha_t \in R_k$ is the attentive weights of V and the attentive visual output $z_t \in R_g$ is calculated as

$$z_t = V\alpha_t \tag{7}$$

The attentive visual output z_t and the hidden state h_t are combined to attain the visual information $s_t \in R_{d_1}$:

$$s_t = \tanh(W_s(z_t + h_t)) \tag{8}$$

where $W_s \in R_{d_1 \times d_1}$ is a transfer matrix that can be learned.[2] The s_t can be seen as a relevant information from the visual perspective.

The attentive semantic output k_t and the hidden state Ch_t are combined to attain the conceptual information $q_t \in R_{d_1}$:

$$q_t = \tanh(W_q(k_t + Ch_t)) \tag{12}$$

where $W_q \in R_{d_1 \times d_2}$ are learnable parameters.[2]

c) Adaptive Merging Gate: It is not reasonable to treat s_t and q_t equally when the decoder generates different type of words. For example, when generating descriptive words (e.g., behind, red), s_t should matters more than q_t as it contains visual information that accounts for details. However, when generating object words (e.g., people, table), q_t is more important since it contains conceptual information that is responsible for a comprehensive view. The balance between s_t and q_t is dynamically adjusted in the generation process. We introduce a novel score-based merging mechanism that helps the model to adaptively adjust the balance:

$$\gamma_t = \sigma(S(q_t) - S(s_t)) \tag{13}$$

$$c_t = \gamma_t q_t + (1 - \gamma_t) s_t \tag{14}$$

where σ is the sigmoid function, $\gamma_t \in [0, 1]$ indicates how important q_t is compared to s_t, and S is the scoring function. This function evaluates the importance of the visual attention and the semantic attention. Note that Eq. 5 and Eq. 6 evaluate the importance of every $v_i \in V$, meanwhile, Eq. 9 and Eq. 10 evaluate the importance of every $w_i \in T$. As these equations have similar forms and purposes, we can combine them organically to define the scoring function S:

$$\begin{aligned} S(q_t) &= \tanh(W_{s,q} q_t + W_{s,h} h_t) \cdot w_s \\ &\quad + \tanh(b\ W_{s,q} q_t + b\quad W_{s,h}\quad Ch_t) \cdot C w_s \\ S(s_t) &= \tanh(W_{s,s} s_t + W_{s,h} h_t) \cdot w_s \end{aligned} \tag{15}$$

b).Semantic Attention: In different parts of an image caption, the keywords are concerned differently. So we introduce a second LSTM to perform semantic attention. Different from the existing work [8], for model simplicity and calculation convenience of the Adaptive Merging Gate, we take an approach similar to the visual attention. Therefore, the model attends to the keywords based on the hidden state of the second LSTM:

$$Q_t = \tanh(W_{Q,T}\, T \oplus W_{Q,h}\, Ch_t) \quad (9)$$

$$\beta_t = \text{softmax}(Q_t w_{\beta,Q}) \quad (10)$$

where $W_{Q,T} \in R_{m\times e}$, $W_{Q,h} \in R_{m\times d_2}$, $w_{\beta,Q} \in R_m$ are the parameters to be learned. $\beta_t \in R_m$ is the attentive weights of T and the attentive semantic output $k_t \in R_e$ is calculated as

$$k_t = T\beta_t \quad (11)$$

[1]For conciseness, all the bias terms of linear transformations in this paper are omitted.

[2]Since g is equal to d_1(e is equal to d_2), z_t and h_t(k_t and $\tilde{}h_t$) can be added directly. See section 4.2 for details.

$$+ \tanh(b\ W_{S,s}s_t + b\ W_{S,h}\ Ch_t) \cdot Cws \quad (16)$$

where · denotes dot product of vectors, $W_{S,q} \in R^{k\times d_1}$, $W_{S,s} \in R_{k\times d_1}$, $b\ W_{S,q} \in R_{m\times d_1}$, $b\ W_{S,s} \in R_{m\times d_1}$ are the learnable parameters, and $W_{S,h}$, ws share the weights of $W_{Z,h}$, $w_{\alpha,Z}$ from Eq. 5 and Eq. 6, and $b\ W_{S,h}$, Cws share weights with $W_{Q,h}$, $w_{\beta,Q}$ from Eq. 9 and Eq. 10, respectively.

Finally, the output word is sampled from this distribution:

$$y_t \sim p_t = \text{softmax}(W_{p,e}c_t) \quad (17)$$

where each value of $p_t \in R_{|D|}$ is a probability indicating how likely each word in vocabulary D should be the current output word. The training objectives of the whole model can be found in Section 4.

Through the above approaches, the model is encouraged to take advantage of all the available information.

IV. Training Objectives

Traditional approach of training text generation model is to minimize Cross-Entropy Loss [14], [17], [18], [30]. Given a

Page 4

target ground truth sequence $y^*_{1:T}$ and a captioning model with parameters θ, we minimize the following cross entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y^*_t | y^*_{1:t-1})) \quad (18)$$

In recent years, reinforcement learning has proven to be quite helpful in natural language generation tasks. In this case, the cross-entropy loss approach is used to pre-train the model, after pre-training our goal is to minimize this negative expectation:

$$L_R(\theta) = -E_{y_{1:T}\sim p_\theta}[r(y_{1:T})] \quad (19)$$

where r is the score function (e.g., CIDEr). Following the approach described as Self-Critical Sequence Training [31] (SCST), the gradient of this loss can be approximated:

$$\nabla_\theta L_R(\theta) \approx -(r(y^s_{1:T}) - r(y_{1:T}))\nabla_\theta \log p_\theta(y^s_{1:T}) \quad (20)$$

where $y^s_{1:T}$ is a sampled caption and $r(y_{1:T})$ defines the baseline score obtained by greedily decoding the current model. SCST explores the space of captions by sampling from the policy during training. This gradient tends to increase the probability of sampled captions that score higher than the score from the current model. Using this approach, we complete CIDEr optimization.

V. Experiment

In this section, we describe two benchmark datasets for **image captioning** and some widely-used metrics, followed by our training details and evaluation of the proposed model.

A. Datasets and Metrics

There are several datasets made up of images-caption pairs.

B. Settings

Following common practice, we use the ResNet152 model [27] pre-trained on ImageNet.[3] Then we project the extracted 2048 7 × 7 feature maps into 512 feature maps, i.e. g is 512. The word embedding size e is 256, the hidden size d_1 of the first LSTM is 512 and the hidden size d_2 of the second LSTM is 256. We replace caption words that occur less than 5 times in the training set with the generic unknown word token UNK, resulting in 9,567 words for MSCOCO and 7,649 for Flickr30k. We use the keyword extractor pre-trained by [7] for 1,000 keywords on MSCOCO. For an image, only the top 19 keywords are selected, which means m is 19. The same keywords extractor is used for Flickr30k. The caption words and the key words share the same embeddings. In training, we train the entire model with the batch size of 60. The learning rates for the CNN and two LSTMs are 1e-5 and 4e-4, respectively. We also use momentum of 0.8 and weight decay of 0.999. We use Adam [41] for parameter optimization. For fair comparison, we adopt early stop based on CIDEr within maximum 50 epochs. We apply beam-search with a beam size of 5 when sampling the caption for both COCO and Flickr30k datasets.

C. Step-Wise Instructions for Training Full Model

In order to optimize the final performance and increase the speed of convergence, we recommend the following steps to train the full model:

Step 1: Train the first LSTM (Visual Attention). Fix the weight of the CNN encoder model, set the γ value to 0 for every time step, and train 15 epochs.

Our reported results are evaluated on the popular Microsoft COCO [32] dataset, which contains 123,287 images, and the Flickr30k [33] dataset, which contains 31,000 images. Each image in both dataset is paired with 5 sentences. We report results using the widely-used publicly-available splits in the work of Karpathy and Li [6]. There are 5,000 images each in validation set and test set for MSCOCO, for Flickr30k, there are 1,000 images in each set.

The COCO captioning evaluation toolkit [32] that reports the widely-used automatic evaluation metrics SPICE, CIDEr, BLEU, METEOR, and ROUGE is used to test the model performance. SPICE [34], which is based on scene graph matching, and CIDEr [35], which is based on n-gram matching, are specifically designed to evaluate **image captioning** systems. They both incorporate the consensus of a set of references for an example. BLEU [36] and METOR [37] are originally proposed for machine translation evaluation. ROUGE [38], [39] is designed for automatic evaluation of extractive text summarization. In some related researches, the conclusion is that SPICE correlates best with human judgments with a small margin of discrepancy, and is excellently good at judging detailedness, where the other metrics show negative correlations; CIDEr and METEOR follows with no conspicuous superiority, after comes ROUGE-L, and BLEU-4, in that order [34], [35], [40].

Step 2: Train the second LSTM (Semantic Attention). Fix the weight of the CNN encoder model, set the γ value to 1 for every time step, and train 15 epochs.

Step 3: Train the Adaptive Merging Gate. Fix the weight of the CNN encoder model and two LSTMs, and train 10 epochs.

Step 4: Joint training.

D. Results

Comparison with various representative systems on MSCOCO, including the recently proposed NBT [14] and Up-Down [13], which is the state-of-the-art on the dataset in comparable settings, further demonstrates the advantages of our approach. Table I shows the result on MSCOCO, when the model is trained with Cross-Entropy Loss. As we can see, our model has desirable improvement w.r.t. comparable systems in terms of all of the metrics except BLEU-4. Especially, our model overpasses the state-of-the-art Up-Down, trained with extra data, to a comfortable extent for the SPICE metric, which is considered to correlate best with human judgments [34].

Table II shows the results on MSCOCO, when the model is trained to optimize CIDEr. Among directly comparable models, our model is arguably the best and outperforms

[3]We use the pre-trained model from torchvision(https://github.com/pytorch/vision).

MS COCO(Cross-Entropy Loss)	SPICE	CIDEr	METEOR	ROUGE-L	BLEU-4
HardAtt [16]	-	-	0.230	-	0.250
ATT-FCN [8]	-	-	0.243	-	0.304
SCA-CNN [17]	-	0.952	0.250	0.531	0.311
LSTM-A [30]	0.186	1.002	0.254	0.540	0.326
AdaAtt [18]	0.195	1.085	0.266	0.549	0.332
NBT [14]	0.201	1.072	0.271	-	0.347
Up-Down(Cross-Entropy Loss) [13]	0.203	1.135	0.270	0.564	0.362
aimNet(Cross-Entropy Loss)	0.222	1.141	0.276	0.567	0.333

TABLE I

WHEN MODEL IS TRAINED WITH CROSS-ENTROPY LOSS, THE PERFORMANCE ON THE MSCOCO KARPATHY TEST SPLIT. THE SYMBOL DENOTES USING EXTRA DATA FOR TRAINING. ESPECIALLY, OUR MODEL SUPERSEDES ALL EXISTING MODELS IN SPICE, WHICH CORRELATES THE BEST WITH HUMAN JUDGMENTS.

MS COCO(CIDEr Optimization)	SPICE	CIDEr	METEOR	ROUGE-L	BLEU-4
SCST [31]	-	1.140	0.267	0.557	0.342
Up-Down(CIDEr Optimization) [13]	0.214	1.201	0.277	0.569	0.363
aimNet(Cross-Entropy Loss)	0.222	1.141	0.276	0.567	0.333
aimNet(CIDEr Optimization)	0.212	1.191	0.274	0.571	0.364

TABLE II

WHEN MODEL IS TRAINED WITH CIDER OPTIMIZATION, THE PERFORMANCE ON THE MSCOCO KARPATHY TEST SPLIT. SYMBOLS, ARE DEFINED SIMILARLY. OUR MODEL SIGNIFICANTLY OUTPERFORMS THE CURRENT STATE-OF-THE-ART UP-DOWN IN TERMS OF SPICE.

the existing models in terms of ROUGE-L and BLEU-4. Encouragingly, even when trained with Cross-Entropy Loss, our model is also very competitive with Up-Down(CIDEr Optimization) [13], which uses much larger dataset, Visual Genome [21], where dense annotations are used to train the object detector. Despite the gaps in CIDEr and BLEU-4, our

[5] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," CoRR, vol. abs/1412.6632, 2014. [Online]. Available: http://arxiv.org/abs/1412.6632

[6] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2015, pp. 3128–3137. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298932

model performs well in terms of ROUGE-L and METEOR, and significantly outperforms the state-of-the-art in SPICE.

VI. Conclusions

We propose the adaptive image-keyword merging network (aimNet) to improve **image captioning**. Based on a mechanism that extends traditional visual attention and semantic attention, aimNet sequentially and adaptively merges the visual and the conceptual information to generate high-quality image captions. The score-based adaptive merging mechanism serves as an efficient guide of the two kinds of information flow when generating the caption. Experiments demonstrate the effectiveness of the proposed approach.

References

[1] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 1080–1089. [Online]. Available: https://doi.org/10.1109/CVPR.2017.121

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2015, pp. 3156–3164. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298935

[3] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2015, pp. 2422–2431. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298856

[4] F. Liu, C. You, X. Wu, S. Ge, S. Wang, and X. Sun, "Auto-encoding knowledge graph for unsupervised medical report generation," in NeurIPS, 2021, pp. 16 266–16 279.

[7] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2015, pp. 1473–1482. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298754

[8] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "**Image captioning** with semantic attention," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 4651–4659. [Online]. Available: https://doi.org/10.1109/CVPR.2016.503

[9] Q. Wu, C. Shen, L. Liu, A. R. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 203–212. [Online]. Available: https://doi.org/10.1109/CVPR.2016.29

[10] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," in ACL/IJCNLP (Findings), 2021.

[11] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. IEEE Computer Society, 2015, pp. 4534–4542. [Online]. Available: https://doi.org/10.1109/ICCV.2015.515

[12] H. Zhang, Z. Kyaw, S. Chang, and T. Chua, "Visual translation embedding network for visual relation detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 3107–3115. [Online]. Available: https://doi.org/10.1109/CVPR.2017.331

[13] 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 VQA," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018.

[14] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in 2018

Page 6

IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," CoRR, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[16] 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 Proceedings of the 32nd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.html

[17] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for **image captioning**," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 6298–6306. [Online]. Available: https://doi.org/10.1109/CVPR.2017.667

[18] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for **image captioning**," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 3242–3250. [Online]. Available: https://doi.org/10.1109/CVPR.2017.345

[19] F. Liu, X. Ren, X. Wu, S. Ge, W. Fan, Y. Zou, and X. Sun, "Prophet attention: Predicting attention with future attention," in NeurIPS, 2020.

[20] F. Liu, X. Ren, Y. Liu, H. Wang, and X. Sun, "simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions," in EMNLP, 2018.

[21] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using IEEE Computer Society, 2017, pp. 4904–4912. [Online]. Available: https://doi.org/10.1109/ICCV.2017.524

[31] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for **image captioning**," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 1179–1195. [Online]. Available: https://doi.org/10.1109/CVPR.2017.131

[32] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," CoRR, vol. abs/1504.00325, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

[33] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014. [Online]. Available: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229

[34] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9909. Springer, 2016, pp. 382–398.

[35] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, 2015, pp. 4566–4575. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7299087

[36] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 2002, pp. 311–318.

crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, no. 1, pp. 32–73, 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0981-7

[22] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

[23] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in NeurIPS, 2019.

[24] F. Liu, X. Ren, Z. Zhang, X. Sun, and Y. Zou, "Rethinking skip connection with layer normalization," in COLING, 2020, pp. 3586–3598.

[25] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 35, no. 12, pp. 2891–2903, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.162

[26] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation," in ACL/IJCNLP, 2021.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[28] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. MIT Press, 2006, pp. 1417–1424. [Online]. Available: http://papers.nips.cc/paper/2926-multiple-instance-boosting-for-object-detection.pdf

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[30] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting **image captioning** with attributes," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. [Online]. Available: http://www.aclweb.org/anthology/P02-1040.pdf

[37] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, June 29, 2005, J. Goldstein, A. Lavie, C. Lin, and C. R. Voss, Eds. Association for Computational Linguistics, 2005, pp. 65–72. [Online]. Available: https://aclanthology.info/papers/W05-0909/w05-0909

[38] C. Lin and E. H. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003, M. A. Hearst and M. Ostendorf, Eds. The Association for Computational Linguistics, 2003. [Online]. Available: http://aclweb.org/anthology/N/N03/N03-1020.pdf

[39] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, July, 2004, S. S. Marie-Francine Moens, Ed. Association for Computational Linguistics, 2004, pp. 74–81.

[40] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Federated learning for vision-and-language grounding problems," in AAAI, 2020.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in CVPR, 2021.