doi:10.3772/j.issn.1006-6748.2020.02.011

Improved image captioning with subword units training and transformer[①]

Cai Qiang(蔡　强)[②], Li Jing, Li Haisheng, Zuo Min
(School of Computer and Information Engineering, Beijing Techology and Business University, Beijing 100048, P. R. China)
(Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing 100048, P. R. China)
(National Engineering Laboratory for Agri-Product Quality Traceability, Beijing 100048, P. R. China)

Abstract

Image captioning models typically operate with a fixed vocabulary, but captioning is an open-vocabulary problem. Existing work addresses the image captioning of out-of-vocabulary words by labeling it as unknown in a dictionary. In addition, recurrent neural network (RNN) and its variants used in the caption task have become a bottleneck for their generation quality and training time cost. To address these 2 essential problems, a simpler but more effective approach is proposed for generating open-vocabulary caption, long short-term memory (LSTM) unit is replaced with transformer as decoder for better caption quality and less training time. The effectiveness of different word segmentation vocabulary and generation improvement of transformer over LSTM is discussed and it is proved that the improved models achieve state-of-the-art performance for the MSCOCO2014 image captioning tasks over a back-off dictionary baseline model.

Key words: image captioning, transformer, byte pair encoding (BPE), reinforcement learning

0 Introduction

Problems combining image and language understanding like image captioning continue to inspire considerable researches at the boundary of computer vision and natural language processing. In these tasks, it is reasonable to perform some fine-grained visual processing, or even multiple steps of reasoning to create high quality outputs. As a result, visual attention mechanisms have been widely adopted in image captioning[1-4]. These mechanisms improve image captioning performance by extracting salient and useful image features.

However, this problem can also be addressed from a language perspective. Image captioning is more than an image processing problem and a fine-grained method for generating high-quality captions is proposed in this paper. Image captioning has recently shown impressive results[2] by backing off words with a frequency below 5. The training vocabulary of neural models is usually limited in 10 000 – 30 000 words on MSCOCO[5] image captioning training data, but caption generation is an open-vocabulary problem, and especially for images with massive visual parts, image captioning models require a mechanism that generates more detailed and informative words.

For previous word-level caption models, the generation of out-of-vocabulary words is impossible and these models generate some common words with fixed sentence form. It is observed that such methods make assumptions that often do not hold true in a practical scene. For instance, there is not always a 1-to-1 correspondence between training image and corresponding up to 5 captions in that not all descriptive information is involved in the captions. In addition, word-level models are unable to generate captions unseen before.

In this work, image captioning models that train on the level of subword units[6] is investigated. The goal is to build a model which can handle open-vocabulary problem in the encoder-decoder network itself. The model is able to make the captions generation model more fine-grained and achieve better accuracy for the translation of rare words than back-off dictionaries. It is showed that the neural networks are able to learn rare descriptive words from subword representations in experimental analysis.

To make the image captioning process simpler, transformer[7], instead of recurrent neural network (RNN) or its variants is used as decoder part. Transformer, as a backbone architecture, has been applied to a large amount of natural language processing

[①] Supported by the National Natural Science Foundation of China (No. 61877002) and Beijing Municipal Commission of Education (PXM2019_014213_000007).

[②] To whom correspondence should be addressed. E-mail: li10141110@gmail.com

Received on Apr. 29, 2019

tasks[8,9]. Transformer is a novel neural network architecture based on a self-attention mechanism proposed by Google that has been proved particularly well suited for generation tasks, such as machine translation and text-to-speech. So it can also contribute to image captioning. The transformer outperforms both recurrent and convolutional models on academic English to German and English to French translation benchmarks. The transformer proposed by Google also complies with sequence-to-sequence structure, consisting of encoder and decoder. The encoder is made up of multi-head attention layer and feed forward layer for extracting features from source and the decoder part consist of masked multi-head attention layer, multi-attention layer and feed forward layer. The decoder part of the full transformer model is employed for decoding visual information. In transformer based image captioning (TIC) model, bi-direction long short-term memory (LSTM) decoder is replaced by transformer decoder for less training time and better captions generation.

This paper has 2 main contributions:

(1) Open-vocabulary image captioning is feasible by encoding (rare) words via subword units is proved. Moreover, byte pair encoding (BPE)[6] is utilized for the task of fine-grained word segmentation and caption generation. BPE allows for the representation of an open vocabulary, which makes it suitable for word segmentation in neural network architecture.

(2) Transformer based image captioning model is proposed, it adopts a self-attention based neural network to the task of image captioning. Other than taking advantage of the full transformer model, the decoder part of transformer is extracted for the generation of sentence and the experimental results show that the proposed method outperforms baseline model.

1 Related work

Most modern approaches[1,2] encode an image using a convolutional neural network (CNN), and feed this as input to a recurrent neural network or its variants, typically with some form of gating or memory mechanism. The RNN can generate an arbitrary length sequence of words. Within this common framework, many research work[10,11] explored different encoder-decoder structures including attention-based models. Multi-kinds of attention mechanism are applied to the output of one or more layers of a CNN, by predicting weights distribution on CNN output of the input image. Whereas, choosing the optimal number of image regions invariably leads to an unwinnable trade-off between coarse and fine levels of detail. Moreover, the arbitrary positioning of the regions with respect to image content may make it more difficult to detect objects that are poorly aligned to regions and to bind visual concepts associated with the same object.

Comparatively few previous works have considered addressing caption generation problem from a language perspective. Sennrich et al.[6] proposed byte pair encoding to segment words, which enable the encoder-decoder machine translation model to generate open-vocabulary translation. Applied originally for neural machine translation (NMT), BPE is based on the intuition that various word classes are made of smaller units than words such as compounds and loanwords. In addition to making the vocabulary smaller and the length of sentences shorter, the subword model is able to productively generate new words that are not seen at training time.

Neural networks, in particular recurrent neural network, has been the center of leading approaches to sequence modeling tasks such as image captioning, question answering and machine translation for years. However, it takes long time to train an RNN model in that it can only process the input data step by step. The transformer proposed by Google has received much attention in the last two years. In contrast to RNN-based approaches, the transformer used no recurrence, instead processing all words or symbols in the sequence in parallel while making use of a self-attention mechanism to incorporate context from words or features farther away. By processing all words in parallel and letting each word attend to other words in the sentence over multiple processing steps, the transformer was much faster to be trained than recurrent models. Remarkably, experiments on machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to be trained. Transformer achieves state-of-the-art performance on the machine translation task. Besides, given large or limited training data, the transformer model generalizes well to other sequence modeling problem. However, on smaller and more structured language understanding tasks, or even simple algorithmic tasks such as copying a string (e. g. to transform an input of ‘abc’ to ‘abcabc’), the transformer does not perform very well. In contrast, models that perform well on these tasks fail on large-scale language understanding tasks like translation and caption generation.

2 Approach

Given an image I, the image captioning model takes as input a possibly variably-sized set of k image features, $VI = \{v_1, \cdots, v_k\}$, $V_i \in R^D$, such that each image crop feature encodes a sematic region of the image. The spatial image features V can be variously defined as the output of bottom-up attention model, which extracts multi crop features under the architecture of Faster R-CNN[12]. The same approach in Ref. [1] is followed to implement a bottom-up attention model and the details are described in Ref. [1]. In Section 2.1, the practical use of BPE algorithm for captions segmentation is demonstrated. In Section 2.2, the architecture of TIC model is outlined.

2.1 Byte pair encoding

Byte pair encoding is a technique designed for simple data compression. BPE iteratively replaces the most frequent pair of bytes in a captioning sentence with a single, unused byte. This algorithm is adopted for subword segmentation. Instead of merging frequent pairs of bytes, it uses merge characters or character sequences. Following the work of Ref. [6], the BPE preprocess consists of 2 stages: learning BPE and applying BPE.

First of all, in learning BPE stage, the symbol vocabulary is initialized with the character vocabulary, and each word in image caption sentences is represented as a sequence of characters, plus a special end-of-word symbol ‘ · ’, which allows it to restore the original tokenization after caption generation. All symbol pairs are iteratively counted and replaced each occurrence of the most frequent pair (‘A’, ‘B’) with a new symbol ‘AB’. Each merge operation produces a new symbol which represents a character n-gram. Frequent character n-grams (or whole words) are eventually merged into a single symbol, thus BPE requires no shortlist. The final symbol vocabulary size is equal to the size of the initial vocabulary, plus the number of merge operations, the latter is the only hyperparameter of the algorithm. For efficiency, pairs that cross word boundaries are not taken into consideration. The algorithm can thus be run on the dictionary extracted from a text, with each word being weighted by its frequency.

After learning BPE stage, a fixed dictionary is completed. In applying BPE stage, all words in all sentences from training data are substituted for subword units according to the BPE dictionary. Then this dictionary is used to represent each subword units. At last, the one-hot vector x for each word is acquired. The embedding model embeds the one-hot vector into a d_{model} dimensional vector. All these embedding vectors in one sentence are combined into a matrix $L \times d_{\text{model}}$ as the input to the transformer decoder, where L is the length of the sentence.

Two methods of applying BPE are evaluated: learning encodings only for image captioning training dataset, or learning the encoding on the union of the MSCOCO2014 image captioning training dataset and VQA v2.0 dataset[13] (which is called expand BPE). The former has the advantage of being more compact in terms of text and vocabulary size, whereas the latter leads to accurate semantic units by taking the larger vocabulary into account.

2.2 Transformer based image captioning

Transformer based image captioning model contains 2 parts, the encoder and the decoder, as is shown in Fig. 1.

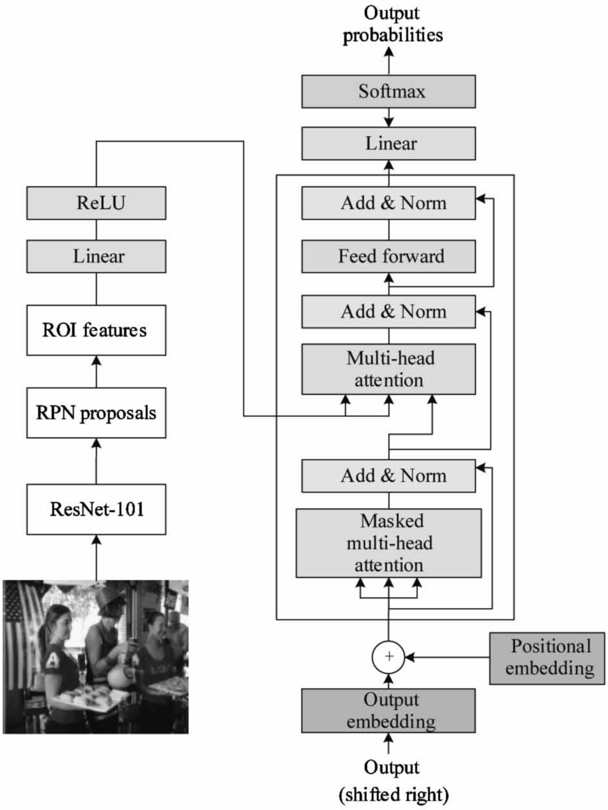

Fig. 1 The framework of the proposed TIC model

Most image captioning models are made up of the encoder-decoder structure. The encoder used in this work is a bottom-up attention model borrowed from Ref. [1]. Bottom-up attention model utilizes Faster R-CNN for mapping an image to a context feature VI. This process is shown as

$$VI = Faster\ R\text{-}CNN(\boldsymbol{I}) \tag{1}$$

where, $\boldsymbol{I}$ is vector of input image, $VI = \{v_1, \cdots, v_k\}$ is the image features processed by Faster R-CNN based bottom-up attention model.

Faster R-CNN is an object detection model designed to localize and recognize objects in a project given image with bounding boxes. Objects are detected by Faster R-CNN in 2 stages. The first stage, described as a region proposal network (RPN), predicts object proposals. Then the predicted top box proposals are selected as input to the second stage for labels classification and class-specific bounding box refinements. In this work, ResNet-101 CNN is used as feature extractor in Faster R-CNN model. The final output of the model is selected as the input caption model. For each selected region i, *VI* is defined as the mean-pooled convolutional feature from this region, such that the dimension D of the image feature vectors is 2048. Faster R-CNN is served as a 'hard' attention mechanism using this fashion, as only a relatively small number of image bounding box features are selected from a large number of possible configurations.

The decoder part of the transformer with stacked attention mechanisms is taken to decode the encoded image feature into the sentence. The transformer model composes of a stack of N identical layers and contains no RNN structure. Note that each layer has 3 sub-layers. The first sub-layer uses the multi-head self-attention mechanism. The multi-head attention is shown as

$$h_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$H = Concat(h_1, \cdots, h_n) \tag{3}$$

$$O = HW_h \tag{4}$$

where the projections are parameter matrices $W_i^Q \in R^{d_{\text{model}} \times d_k}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, $W_i^V \in R^{d_{\text{model}} \times d_v}$. $Q \in R^{L \times d_{\text{model}}}$, $K \in R^{L \times d_{\text{model}}}$, $V \in R^{L \times d_{\text{model}}}$ are the inputs of the multi-head attention. Attention is the scaled dot-product attention. Concat is the concatenating function. $h_i \in R^{L \times d_v}$ is the output of the scaled dot-product attention. n scaled dot-product attention is concatenated to generate $H \in R^{L \times (n \times d_v)}$. $W_h \in R^{(n \times d_v) \times d_{\text{model}}}$ is used to project H into the output $O \in R^{L \times d_{\text{model}}}$.

Fig. 1 shows that the inputs of this layer are fixed to output embedding plus positional embedding. This sub-layer makes use of a masked mechanism for preventing this model from seeing the future information, which ensures the generation of the current word with only the previous generated words. In contrast to the first sub-layer, the second sub-layer is a multi-head attention layer without the masked mechanism in that it takes all the image features into consideration in every time step. The multi-head attention is employed over preprocessed image features and the output of the first sublayer. This sublayer is vital importance to blend the text information with the image information using attention mechanism. The third sublayer is a position-wise fully connected feed-forward network aiming at selecting the most relevant information for generating image captions. In addition, a residual connection is utilized around each of the 3 sub-layers in the transformer decoder, followed by layer normalization. Finally, a full connected layer and a softmax layer is used to project the output of the transformer decoder to the probabilities distribution of the vocabulary. Using the notation $y_{1:T}$ to refer to a sequence of words $(y_1, \cdots, y_T)$, at each time step t the conditional distribution over possible output words is given by

$$p(y_t \mid y_{1:t-1}) = softmax(W_p h_t^2 + b_p) \tag{5}$$

where, $W_p \in R^{|\Sigma| \times M}$ and $b_p \in R^{|\Sigma|}$ are learned weights and biases.

Given a target ground truth sequence $y_{1:T}^*$ and a captioning model with parameters θ, the training of the model minimizes the following cross entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*)) \tag{6}$$

For fair comparison with recent work[14], results optimized for CIDEr is also reported. Initializing from the cross-entropy trained model, the training seeks to minimize the negative expected score:

$$L_R(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})] \tag{7}$$

where, r is the score function (e.g., CIDEr). Following the approach described as self-critical sequence training (SCST), the gradient of this loss can be approximated:

$$\nabla_\theta L_R(\theta) \approx -[r(y_{1:T}^s) - r(\hat{y}_{1:T})] \nabla_\theta \log p_\theta(y_{1:T}^s) \tag{8}$$

where, $y_{1:T}^s$ is a sampled caption and $r(\hat{y}_{1:T})$ defines the baseline score obtained by greedily decoding the current model. SCST (like other reinforce[15] algorithms) 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.

3 Experiments and results

3.1 Datasets

The MSCOCO2014 captions dataset[5] is employed to evaluate the proposed transformer based image captioning model. For validation of model hyperparameters and offline testing, this paper uses the 'Karpathy' splits[16] that have been used extensively for reporting results in prior work. This split contains 113287 training images with 5 captions each, and 5 K images respectively for validation and testing. To explore the performance of BPE, all sentences are converted to lower case, tokenized on white space, and substituted words with subword units according to BPE vocabulary. To evaluate caption quality, this work uses the standard automatic evaluation metrics, namely SPICE[17], CIDEr, METEOR, ROUGE-L[18] and BLEU[19].

To evaluate the proposed expand BPE model, the recently introduced VQA v2.0 dataset[13] is used. VQA v2.0 is proposed to minimize the effectiveness of learning dataset priors by balancing the answers to each question, but in the experiment this dataset only takes advantage of expanding BPE corpus with 1.1 M questions and 11.1 M answers relating to MSCOCO images.

3.2 Experiment settings

For fair comparison with bottom-up and top-down baseline model, TIC model takes the same pretrained image features of bottom-up and top-down baseline model as inputs. To pretrain the bottom-up attention model, Anderson et al. [1] initialized Faster R-CNN with ResNet-101 pretrained for classification on ImageNet, then trained it on visual genome[20] data. For the six-layer-stacked transformer model, this work sets the model size which is d_{model} to be 512 and the minibatch size to be 32. The Adam method is adopted to update the parameters of transformer. The initial learning rate of the transformer is 4×10^{-4}. The momentum and the weight-decay are set as 0.9 and 0.999 respectively. All implements of neural networks are based on PyTorch deep learning framework. In evaluation stage, the beam search size is set to 5 for high-quality caption generation at the sacrifice of decoding time.

3.3 Image captioning results

Table 1 shows single-model image captioning performance on the MSCOCO Karpathy test split. TIC + BPE-10K-exp stands for expanding BPE trained on MSCOCO2014 captions dataset and VQA v2.0 dataset with a dictionary of 10 000. The TIC model obtains similar results to baseline, the existing state-of-the-art on this test set. TIC plus BPE training model achieves significant (2% –7%) relative gains across all metrics regardless of whether cross-entropy loss or CIDEr optimization is used, which illustrates the contribution of transformer and BPE algorithm to image captioning task.

In Table 1 the performance of the improved TIC model and the existing state-of-the-art bottom-up and top-down baseline is demonstrated in comparison to SCST approach on the test portion of the Karpathy splits. For fair comparison, results are reported for models trained with both standard cross-entropy loss, and models optimized for CIDEr score. Note that the SCST[14] takes advantage of reinforcement learning to optimize evaluation metrics. And it also uses ResNet-101[21-23] encoding of full images, similar to the bottom-up and top-down baseline model and TIC model. All results are reported for a single model with no finetuning-of the input ResNet/Faster R-CNN model.

Table 1 Performance of different models on MSCOCO2014

	XE loss						CIDEr optimization					
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SCST: Att2in[14]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST: Att2all[14]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Baseline	75.7	35.7	27.6	56.2	112.0	20.4	79.8	36.3	27.7	56.9	120.1	21.4
TIC	75.9	35.8	27.6	56.4	112.9	20.4	80.1	36.3	27.6	57.0	119.6	21.4
TIC + BPE-10K	76.0	36.9	27.8	**56.9**	113.8	20.7	80.2	**37.4**	28.5	58.2	126.0	21.9
TIC + BPE-20K	75.8	36.0	27.5	56.6	113.4	20.5	80.2	36.8	28.3	58.0	124.9	21.7
TIC + BPE-30K	75.7	35.8	27.3	56.4	114.0	20.4	79.8	36.8	28.4	57.7	123.7	21.6
TIC + BPE-10K-exp	**76.2**	**36.9**	**27.8**	56.8	**115.2**	**20.9**	**80.3**	37.3	**28.8**	**58.5**	**126.7**	**22.3**

Compared to the bottom-up and top-down baseline model, TIC model obtains slightly better performance under both cross-entropy loss and CIDEr optimization loss, which shows the feasibility of replacement of RNN with transformer. Moreover, instead of using word-level model with a back-off dictionary, BPE subword units model brings improvements in the generation of rare and unseen words and outperforms the bottom-up and top-down baseline by 0.1 – 1.2 BLEU-4 and 0.9 – 3.2 CIDEr under XE loss training. Regardless of whether cross-entropy loss or CIDEr optimization is used, Tabel 1 shows that TIC models acquire improvements across all metrics using just a single transformer decoder model and BPE method. The TIC model achieved the best reported performance on the Karpathy test split as illustrated in Table 1.

In addition, the results about the effect of the different sizes of BPE dictionary is explored. Three different sizes are implemented to find the appropriate settings. The TIC + BPE-10K model means that BPE dictionary size is set to 10 000. From these scores in Table 1, it can be implied that all TIC with BPE model is improved over the baseline model. And when the vocabulary size is set to 10 000 and trained on multi-dataset, the TIC + BPE-10K-exp model gets the best performance. According to these scores, it can be inferred that fixed dictionary size is necessary for the generation common description. Whereas, it is believed that larger dictionary size is needed given larger image captioning dataset.

4 Conclusions

This work proposes a novel transformer image captioning model which is improved by training on subword units. It is shown that image captioning systems are capable of open-vocabulary generation by representing rare and unseen words as a sequence of subword units. The transformer decoder with multi-head self-attention modules enables the caption model to memorize dependencies between vision and language context. With these innovations, performance gains have been obtained over the baseline with both BPE segmentation and transformer decoder. The state-of-the-art performance is achieved on the test portion of the Karpathy MSCOCO2014 splits. In addition, the proposed models can be taken into consideration in vision to language problems like visual question answering and text-to-speech.

Reference

[1] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and VQA[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA 2018: 5676-5685

[2] Lu J, Xiong C, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6077-6086

[3] Yang Z, Yuan Y, Wu Y, et al. Review networks for caption generation[C]//Advances in Neural Information Processing Systems, Barcelona, Spain, 2016: 2361-2369

[4] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[J]. *arXiv*:1502.03044, 2015

[5] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//European Conference on Computer Vision, Zürich, Switzerland, 2016: 740-755

[6] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 1715-1725

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, California, America, 2017: 5998-6008

[8] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. *arXiv*:1810.04805, 2018

[9] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[J]. *IEEE Computational Intelligence Magazine*, 2018, 13(3):55-75

[10] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[C]//Proceedings of 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 55-75

[11] Jia X, Gavves E, Fernando B, et al. Guiding the long-short term memory model for image caption generation[C]//Proceedings of the IEEE International Conference on Computer Vision, Boston, USA, 2015: 2407-2415

[12] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems, Montreal, Canada, 2015: 91-99

[13] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 6904-6913

[14] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2017: 7008-7024

[15] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. *Machine Learning*, 1992, 8(3-4):229-256

[16] Johnson J, Karpathy A, Fei-Fei L. DenseCap: fully convolutional localization networks for dense captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 4565-4574

[17] Anderson P, Fernando B, Johnson M, et al. SPICE: semantic propositional image caption evaluation[J]. *Adaptive Behavior*, 2016, 11(4):382-398

[18] Lin C. ROUGE: a package for automatic evaluation of summaries[C]//ACL Workshop, Barcelona, Spain, 2004: 1-10

[19] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Pennsylvania, USA, 2002: 311-318

[20] Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. *International Journal of Computer Vision*, 2017, 123(1):32-73

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778

[22] Zhu A, Zhang Z, Zhang X, et al. A novel framework for semantic segmentation with generative adversarial network[J]. *Journal of Visual Communication and Image Representation*, 2019, 58: 532-543

[23] Zhu X, Li Z, Zhang X, et al. Residual invertible spatio-temporal network for video super-resolution[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Hawaii, USA, 2019: 191-198

Cai Qiang, born in 1969. He is a professor at School of Computer and Information Engineering, Beijing Technology and Business University, China. He received his M. S. degree in computer science from Beijing Technology and Business University in 1994. He received Ph. D degree at Beihang University. His research interests include computer vision, information system and computer graphics.