Supportive and Self Attentions for Image Caption

Jen-Tzung Chien and Ting-An Lin
Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Attention over an observed image or natural sentence is run by spotting or locating the region or position of interest for pattern classification. The attention parameter is seen as a latent variable, which was indirectly calculated by minimizing the classification loss. Using such an attention mechanism, the target information may not be correctly identified. Therefore, in addition to minimizing the classification error, we can directly attend the region of interest by minimizing the reconstruction error due to supporting data. Our idea is to learn how to attend through the so-called supportive attention when the supporting information is available. A new attention mechanism is developed to conduct the attentive learning for translation invariance which is applied for image caption. The derived information is helpful for generating caption from input image. Moreover, this paper presents an association network which does not only implement the word-to-image attention, but also carry out the image-to-image attention via self attention. The relations between image and text are sufficiently represented. Experiments on MS-COCO task show the benefit of the proposed supportive and self attentions for image caption with the key-value memory network.**

***Index Terms*—image caption, attention mechanism, encoder-decoder network, association network**

I. Introduction

Deep learning has been successfully developing for natural language processing [1], [2] and computer vision [3] such as speech recognition [4], machine translation [5], dialogue system, language understanding [6], reading comprehension, image caption [7], image comprehension [8], image classification [9], [10] and question answering [11] where the temporal information in natural language can be learned by recurrent neural network [12] or long short-term memory (LSTM) [13]. There are twofold limitations in LSTM. First, temporal information in LSTM is stored in an internal memory which is basically limited to store abundant information in long and rich history data. Second, the temporal information without attention is basically a loose and insufficient representation for natural language. This study presents a series of new attention mechanisms for memory-augmented neural networks where supportive attention is merged in memory network which provides external memory for information storage in the application of image caption [14]–[16].

In general, attention is used to identify the position of interest for pattern classification [17]–[20]. The attention parameter is seen as a latent variable which is usually estimated according to the classification loss. This paper directly optimizes the neural network which controls the attention by minimizing the reconstruction error due to the supporting data. We deal with image caption task which aims to find the most likely natural sentence to describe the meaning in an input image. We propose two attention methods to image caption based on memory network. The first method is to adjust the supportive attention in a question answering system [21] to work for image caption. Using this method, an input image is first encoded by convolutional layer. Convolutional weights are trained by using the ImageNet dataset [22]. A high-dimensional feature vector is encoded. This vector is then attended multiple times to sequentially calculate the hidden codes to produce different words in the transcription. In the implementation, the supporting data are automatically acquired by using the other attention method. The supportive attention sequentially zooms in different objects of an image for text generation. The second method is developed by combining a self attention and a word-to-image attention which act as complementary evidence for image caption. Again, an image is encoded into a number of feature maps as an input memory by a convolutional layer. Objects in an image are then represented by memory slots. Self-attention is performed to attend the pairs of objects of an image which are helpful for generating natural language. The word-to-image attention is applied to attend the object from individual words. A desirable text transcription does not only reflect the individual objects in lexical level but also characterize the relations of objects in syntactic level. This method further incorporates a residual scheme to allow single attention mechanism in sequential learning at each word. A set of experiments on image caption are conducted to illustrate the merit of the proposed methods.

II. Attention Mechanism

Automatically generating caption for an image plays a crucial role for scene understanding which is one of the most challenging tasks combining the technologies in computer vision and natural language processing. A baseline system, called the show, attend and tell (SAT) [7], was proposed to mimic human capability to compress a large amount of salient visual features into a meaningful text transcription. SAT is basically implemented by an architecture consisting of a convolutional neural network (CNN) as the encoder and a long short-term memory (LSTM) network as the decoder. The CNN encoder is used to extract high-dimensional visual features from input image as shown in Figure 1. The pretrained VGG 19 [23] from ImageNet dataset was adopted. This model takes the feature maps in convolutional layer as the input memory similar to that in end-to-end memory networks for question answering [11], [21]. The feature memory of an image is then denoted by a three-dimensional cube $M \in \mathcal{R}^{H \times W \times C}$ where the H stands for the height, the W stands for the width, and the C stands for the number of channels or feature maps. SAT

then reshapes the memory cube into a two-dimensional matrix $M = \{\mathbf{m}_i\} \in \mathcal{R}^{HW \times C}$.

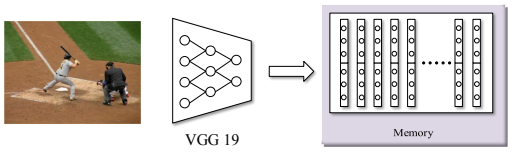

Fig. 1: Feature extraction by convolutional neural network.

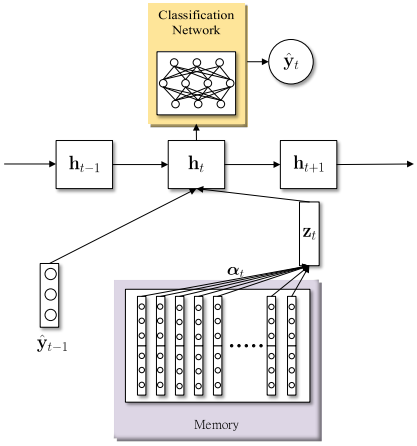

Fig. 2: Architecture for show, attend, and tell (SAT) model

LSTM decoder is then applied to update the hidden state $\mathbf{h}_t$ at each time by using the previously-generated word $\hat{\mathbf{y}}_{t-1}$ and a context vector $\mathbf{z}_t$ by LSTM machine with parameter θ_h

$$\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1}, \hat{\mathbf{y}}_{t-1}, \mathbf{z}_t, \theta_h) \tag{1}$$

and decode the output word $\hat{\mathbf{y}}_t$ by using a classification network $\hat{\mathbf{y}}_t = f_c(\mathbf{h}_t, \theta_c)$ with parameter θ_c. The context vector $\mathbf{z}_t$ is calculated by $\mathbf{z}_t = \sum_i \alpha_{ti}\mathbf{m}_i$ using the attention weights $\boldsymbol{\alpha}_t = \{\alpha_{ti}\}$. The attention weights are calculated by using attention network $f_a(\cdot)$ with parameter θ_a

$$\alpha_{ti} = \frac{\exp e_{ti}}{\sum_k \exp e_{tk}}, \quad \text{where } e_{ti} = f_a(\mathbf{m}_i, \mathbf{h}_{t-1}, \theta_a) \tag{2}$$

or equivalently $\alpha_{ti} = \text{softmax}(f_a(\mathbf{m}_i, \mathbf{h}_{t-1}, \theta_a))$. The LSTM, attention and classification parameters $\{\theta_h, \theta_a, \theta_c\}$ are jointly estimated according to stochastic gradient descent (SGD) algorithm by minimizing the classification errors in terms of cross-entropy error function between SAT outputs $\{\hat{\mathbf{y}}_t\}$ and true words $\{\mathbf{y}_t\}$ in caption. The overall architecture of show, attend and tell is depicted in Figure 2. However, such a baseline system is too simple to achieve significant result. This paper presents two attention schemes to improve the performance of encoder-decoder architecture for image caption. One is the key-value supportive attention, and the other is the self attention. Translation invariance or regularization is performed. Multi-hop attention is implemented.

III. Key-Value Supportive Attention

A new encoder-decoder architecture for image caption is illustrated in left-hand-side of Figure 3 where the key-value (KV) supportive attention (SA) is implemented. The attention weights $\boldsymbol{\alpha}_t$ are sequentially calculated at each time t by an attention controller which is shown in right-hand-side of this figure. For ease of expression, the extension of SAT is denoted by SAT-KVSA which consists of feature extractor, attention controller, support reconstructer and word classifier. SAT-KV is also implemented if only key-value attention is performed.

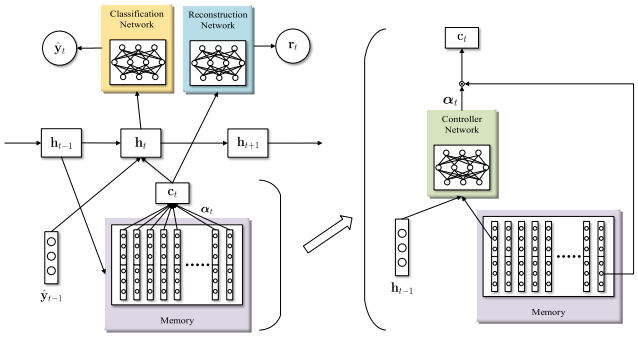

Fig. 3: Architecture for key-value supportive attention (left) where the attention controller (right) is shown.

A. Feature extractor and attention controller

Similar to SAT, the proposed SAT-KVSA uses the same VGG 19 as the pretrained CNN feature extractor to calculate the high-dimensional memory matrix $M = \{\mathbf{m}_i\}$ from input image. $\mathbf{m}_i$ denotes the supervector of features at location i. To fulfill key-value attention, the same memory slot $\mathbf{m}_i$ is used to produce the key memory $\mathbf{m}_i^k$ and value memory $\mathbf{m}_i^v$ by

$$\mathbf{m}_i^k = \mathbf{W}_k^T \mathbf{m}_i + \mathbf{b}_k, \quad \mathbf{m}_i^v = \mathbf{m}_i \tag{3}$$

where $\{\mathbf{W}_k, \mathbf{b}_k\}$ denote the parameters to calculate key vector. The attention controller is performed to calculate the attention weights $\boldsymbol{\alpha}_t$ by using key vector $\mathbf{m}_i^k$. To do so, we first transform the hidden state $\mathbf{h}_t$ of LSTM network at time t to the corresponding vector $\mathbf{h}_t^k$ in key space via a controller network with parameters $\{\mathbf{W}_a, \mathbf{b}_a\}$

$$\mathbf{h}_t^k = \text{ReLU}(\mathbf{W}_a^T \mathbf{h}_{t-1} + \mathbf{b}_a). \tag{4}$$

The activation function based on the rectified linear unit (ReLU) is applied. This key space specifically characterizes the relations among different features i. We therefore calculate inner product between memory and hidden state in key space, and then run a couple of feedforward layers $f_a(\cdot)$ with parameter θ_α to find the attention weights using a softmax function similar to Eq. (2)

$$\boldsymbol{\alpha}_t = \{\alpha_{ti}\} = \{\text{softmax}(f_a(\mathbf{m}_i^k, \mathbf{h}_{t-1}^k, \theta_a))\}. \tag{5}$$

On the other hand, the value space is constructed to reflect the semantic meaning by calculating the context vector $\mathbf{c}_t$. This

context vector is calculated by a weighted sum over different value memories $\mathbf{c}_t = \sum_i \alpha_{ti}\mathbf{m}_i^v$, which is the summary at each time t. This summary contains the compressed information for text transcription at each time t. The parameters of key network $\{\mathbf{W}_k, \mathbf{b}_k\}$, controller network $\{\mathbf{W}_a, \mathbf{b}_a\}$ and attention network θ_a for finding attention weights in attention controller are formed. Different from conventional attention in SAT, the calculation of attention vector $\boldsymbol{\alpha}_t$ and context vector $\mathbf{c}_t$ in SAT-KV or SAT-KVSA uses individual key $\mathbf{m}_i^k$ and value memories $\mathbf{m}_i^v$, respectively. Separate memories in attention scheme are helpful to find precise word-to-image summary.

B. Support reconstructer and word classifier

Traditional attention in SAT is fulfilled by minimizing the classification loss due to caption words. This paper presents the supportive attention for image caption where not only the classification loss due to text transcription but also the regression loss due to supporting data are minimized. In the implementation, there is no additional data collection. The context vector $\mathbf{z}_t$ using SAT is treated as the useful guidance or supporting data in implementation of SAT-KVSA. Our motivation is to regularize the proposed key-value attention by *aligning* the context vectors $\{\mathbf{z}_t, \mathbf{c}_t\}$ between SAT and SAT-KVSA at each time t during training procedure. Supportive attention is performed for model regularization. The attention controller is regularized by minimizing the regression loss

$$\mathcal{L}_r = \sum_{t=1}^{T}(\mathbf{z}_t - \mathbf{r}_t)^2, \quad \text{where } \mathbf{r}_t = f_r(\mathbf{c}_t, \theta_r). \tag{6}$$

Translation invariance is preserved. An additional reconstruction network or support reconstructer $f_r(\cdot)$ with parameter θ_r is merged. At last, a classification network or word classifier is introduced in SAT-KVSA to calculate the classification loss

$$\mathcal{L}_c = -\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{V}|} y_{tk} \log \hat{y}_{tk} \tag{7}$$

which is a cross-entropy error function between classification outputs $\hat{\mathbf{y}}_t = \{\hat{y}_{tk}\}$ and true caption words $\mathbf{y}_t = \{y_{tk}\}$. k is the word index in a dictionary with vocabulary size $|\mathcal{V}|$. The classification output $\hat{\mathbf{y}}_t$ is initialized at the first time $t = 1$ with the flag $<$ START $>$ as a trigger to LSTM decoder. The prediction of classifications outputs is calculated by a classification network $\hat{\mathbf{y}}_t = f_c(\mathbf{h}_t, \theta_c)$ at each time where the hidden state of decoder is sequentially updated by a recurrent machine $\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1}, \hat{\mathbf{y}}_{t-1}, \mathbf{c}_t, \theta_h)$. In this machine, the previous state $\mathbf{h}_{t-1}$, the previously-generated word $\hat{\mathbf{y}}_t$ and the context vector $\mathbf{c}_t$ are concatenated as the inputs to LSTM generator for next state $\mathbf{h}_t$. Accordingly, the regression and classification losses are combined with a hyperparameter γ as

$$\mathcal{L} = \mathcal{L}_r + \gamma\mathcal{L}_c \tag{8}$$

and jointly minimized to estimate the parameters of key network $\{\mathbf{W}_k, \mathbf{b}_k\}$, controller network $\{\mathbf{W}_a, \mathbf{b}_a\}$, attention network θ_a, reconstruction network θ_r, LSTM network θ_h and classification network θ_c in the proposed SAT-KVSA. The word-to-image attention is implemented.

IV. Self Attention in Association Network

Using SAT-KVSA, the word-to-image attention is captured by using mutual relation between image memory $\mathbf{m}_i^k$ and word state $\mathbf{h}_{t-1}^k$ where the hidden state of word $\mathbf{h}_{t-1}^k$ is used as a query to attend the image memory $\mathbf{m}_i^k$. This scheme is merged in supportive attention for image caption . Alternatively, the self-attention is introduced to measure the image-to-image attention where the image memories are mutually attended. Memory $\mathbf{m}_i$ is acted as the query to attend different memories $\mathbf{m}_j$. All feature vectors in memory matrix are used to attend all the other feature vectors in the same matrix. The word-to-image attention and image-to-image attention are both implemented to build an association network (ANet) which is composed of feature extractor, encoder, decoder and transducer as shown in Figure 4 where $\hat{\mathbf{y}}_t$ is the classifier output as the prediction of caption word and $\mathbf{x}_i$ is the image input or memory slot $\mathbf{m}_i$. A state transducer is functioned to continuously update the hidden state in decoder $\mathbf{h}_t$ which is used in classifier. An encoder-decoder network is built as a memory module in ANet which is detailed in Figure 5.

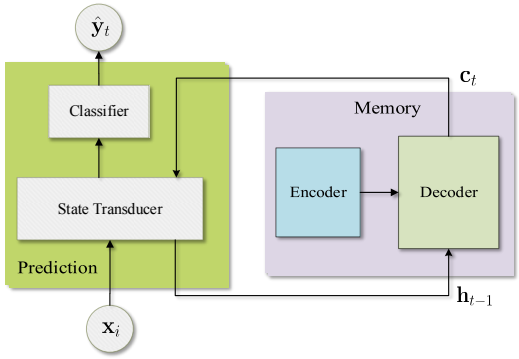

Fig. 4: Architecture for the proposed association network.

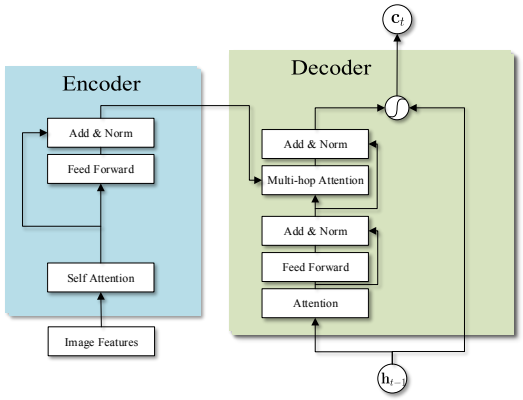

Fig. 5: Encoder-decoder network as a memory module.

A. Feature extractor and self-attention encoder

The feature extractor in ANet is implemented by CNN similar to that in SAT and SAT-KVSA. The feature vectors

$\{\mathbf{m}_i\}_{i=1}^N$ are used to generate the key memories $\{\mathbf{m}_i^k\}_{i=1}^N$ in word-to-image attention and the key memories $\{\mathbf{m}_i^s\}_{i=1}^N$ in image-to-image attention as a kind of self-attention. The key networks $f_k(\cdot)$ and $f_s(\cdot)$ are used as the transformation functions to produce key memories $\mathbf{m}_i^k = f_k(\mathbf{m}_i, \theta_k)$ and $\mathbf{m}_i^s = f_s(\mathbf{m}_i, \theta_s)$ for word-to-image attention and image-to-image attention with parameters θ_k and θ_s, respectively. The value memories in both attentions are the same as $\mathbf{m}_i^v = \mathbf{m}_i$. Using this method, the encoder is implemented for self attention. The features in memory matrix are used as the queries to attend the other features in the same matrix. An image-to-image attention is performed by using mutual relation between any two features in memory matrix. This self attention is able to generate some words which describe the relation between two objects in an image. Image caption can be improved accordingly. In particular, the key-value self attention is performed by finding the attention weights α_{ij} using the attention network $f_a(\cdot)$ with two key memories $\mathbf{m}_i^s$ and $\mathbf{m}_j^s$ at locations i and j, and the parameter θ_a

$$\alpha_{ij} = \text{softmax}(f_a(\mathbf{m}_i^s, \mathbf{m}_j^s, \theta_a)) \tag{9}$$

where $1 \le i, j \le N$. Self-attended context vector is then computed by using the value memories $\{\mathbf{m}_j^v\}_{i=1}^N$ in a form of $\hat{\mathbf{m}}_i = \sum_j \alpha_{ij} \mathbf{m}_j^v$. In this self-attention encoder, a residual block is further added and normalized with the feedforward block. Learning convergence and model flexibility are both enhanced. The self-attended context vector $\hat{\mathbf{m}}_i$ is refined as new input to decoder. Decoder uses this refined context vector to attend and extract semantic information in an image.

B. Decoder and transducer

Decoder is designed to generate the final context where the word-to-image and image-to-image attentions are sequentially performed by using two memories. The first memory is the original feature or value memory $\mathbf{m}_i^v$, and the second memory is the self-attended context memory $\hat{\mathbf{m}}_i$. The decoder module uses the hidden state of transducer (or recurrent machine based on LSTM) $\mathbf{h}_{t-1}$ as the query to attend the first memory $\mathbf{m}_i^v$ to fulfill the word-to-image attention by using weights $\{\alpha_{ti}^{(1)}\}$

$$\mathbf{z}_t^{(1)} = \sum_{i=1}^N \alpha_{ti}^{(1)} \mathbf{m}_i^v, \quad \alpha_{ti}^{(1)} = \text{softmax}(f_a(\mathbf{m}_i^k, \mathbf{h}_{t-1}, \theta_a^{(1)})). \tag{10}$$

The object in the image is found by the word in the caption. After this first-stage attention, the context memory $\mathbf{z}_t^{(1)}$ is refined by a residual network $\mathbf{z}_t = f_r(\mathbf{z}_t^{(1)}, \theta_r)$ to obtain the second query. Then, the second-stage attention is performed by using this query to attend the second memory $\hat{\mathbf{m}}_i$, generated by the encoder, to implement the image-to-image attention by using the second attention weights $\{\alpha_{ti}^{(2)}\}$

$$\mathbf{z}_t^{(2)} = \sum_{i=1}^N \alpha_{ti}^{(2)} \hat{\mathbf{m}}_i, \quad \alpha_{ti}^{(2)} = \text{softmax}(f_a(\mathbf{m}_i^s, \mathbf{z}_t, \theta_a^{(2)})). \tag{11}$$

The resulting context memory $\mathbf{z}_t^{(2)}$ conveys the information including the relation between word and object and the relation between objects themselves which are helpful for image caption. A kind of multi-hop key-value attention is implemented in decoder. In this study, the attention networks $f_a(\cdot)$ in Eqs. (9), (10), (11) are functioned differently with individual parameters θ_a, $\theta_a^{(1)}$ and $\theta_a^{(2)}$, respectively. The inputs to these networks are based on the key memories $\{\mathbf{m}_i^s, \mathbf{m}_j^s, \mathbf{m}_i^k\}$. In decoder module, we additionally arrange a residual shortcut to calculate the final combined context vector

$$\mathbf{c}_t = \beta \hat{\mathbf{z}}_t^{(2)} + (1-\beta)\mathbf{h}_{t-1}, \quad \beta = \sigma(\mathbf{W}_\beta^T \mathbf{h}_{t-1} + \mathbf{b}_\beta) \tag{12}$$

where β denotes the coefficient from an interpolation network between context memory $\hat{\mathbf{z}}_t^{(2)}$ in encoder-decoder network and static memory $\mathbf{h}_{t-1}$ in transducer. The parameters of interpolation network consist of $\mathbf{W}_\beta$ and $\mathbf{b}_\beta$. In practice, the residual shortcut possibly captures the relation between objects, the information of objects themselves, and the other semantic meaning in this association network. The hidden state $\mathbf{h}_t$ acts as an important role to predict caption word $\hat{\mathbf{y}}_t$ and to extract the information from memory module via $\mathbf{c}_t$ at each time t. Similar to SAT and SAT-KVSA, the caption words $\{\hat{\mathbf{y}}_t\}_{t=1}^T$ are predicted by the classification network $\hat{\mathbf{y}}_t = f_c(\mathbf{h}_t, \theta_c)$ driven by the hidden state of LSTM transducer which is calculated by $\mathbf{h}_t = \text{LSTM}(\mathbf{h}_{t-1}, \hat{\mathbf{y}}_{t-1}, \mathbf{c}_t, \theta_h)$ at each time t. The association network (ANet) is accordingly constructed. The parameters of key networks $\{\theta_k, \theta_s\}$, attention networks $\{\theta_a, \theta_a^{(1)}, \theta_a^{(2)}\}$, residual network θ_r, interpolation network $\{\mathbf{W}_\beta, \mathbf{b}_\beta\}$, transducer network θ_h and classification network θ_c are estimated by minimizing a single objective, i.e. classification loss $\mathcal{L}_c$ in Eq. (7). There is no supportive attention performed in ANet.

V. Experiments

A series of experiments were conducted to evaluate the key-value supportive and self attentions for image caption.

A. Experimental setup

The MS COCO caption task [24] contained the human generated captions for the images in Microsoft Common Objects in COntext (COCO) dataset [25] where the captions were collected on Flickr images using the Amazon Mechanical Turk. The images were gathered by searching for pairs of 80 object categories and various scene types on Flickr. The goal of MS COCO image collection process was to gather the images containing multiple objects. A user interface was designed to gather the captions. There were 414K captions for 83K images in training, 203K captions for 41K images in validation, and 379K captions for 41K images in testing. The baseline system was implemented for image caption using the show, attend and tell (SAT) model [7]. The proposed SAT-KV (key-value attention), SAT-KVSA (key-value supportive attention) and ANet (self attention in association network) were carried out for conducting the comparison. The experimental settings of memory networks were referred to [21]. The same architecture of LSTM was used in different methods. Selection of γ in Eq. (8). In the experiments, the attention weights were analyzed. Different models were evaluated by the

metric of BLEU [26] which measured the closeness between the estimated caption and the human caption by matching the unigram, bigram, trigram and fourth-gram (BLEU-1, BLEU-2, BLEU-3 and BLEU-4).

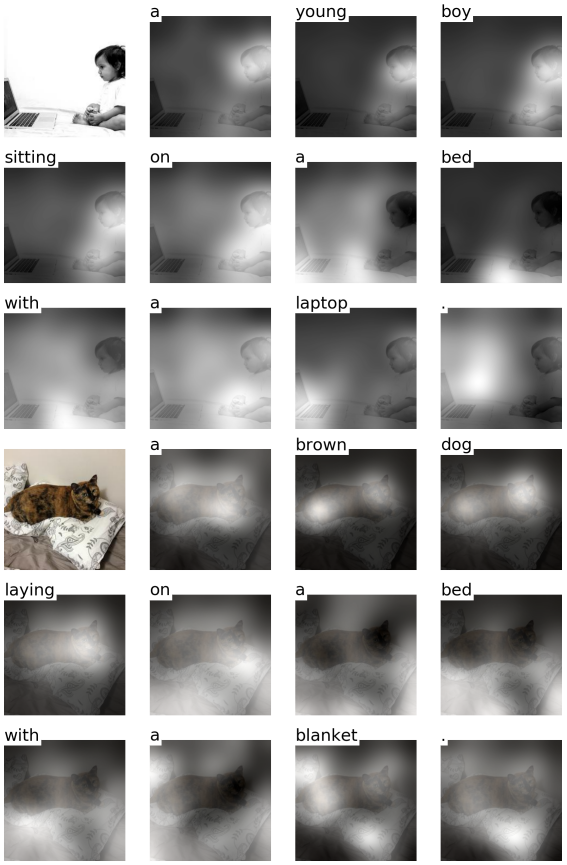

Fig. 6: Illustration for the estimated attention weights of the estimated words corresponding to two example images.

B. Experimental result

First of all, the attention weights $\boldsymbol{\alpha}_t$ of the generated caption words $\hat{\mathbf{y}}_t$ at each time t corresponding to two example images are displayed in Figure 6. SAT-KVSA is evaluated in this caption. The value of attention is reflected by the degree of whiteness. It is obvious that the attention weights clearly spotlight on the location of objects in the first image while generating the words including 'boy', 'bed' and 'laptop'. The attention weights are gradually changed to move the focus not only on the objects but also the adjectives ('young', 'sitting') and the determiner ('a') which construct the meaning of a sentence. Similar performance is obtained when spotlighting on the objects ('bed' and 'blanket') and adjectives ('brown' and 'laying') in the second example. But, the object 'cat' is misclassified into a 'dog'. Basically, supportive attention can improve identifying the location of objects and adjectives. Table I reports the BLEU values of using different methods. The proposed SAT-KV, SAT-KVSA and ANet consistently perform better than baseline method using SAT. The precise attention with separate key memory and value memory does work. The supportive attention with external supporting information from standard SAT is helpful. In this comparison, the self attention in ANet works as good as the supportive attention in SAT-KVSA. ANet does not require supporting data but needs additional computation and storage to perform word-to-image and image-to-image attentions. Table II show the experimental result on ablation study using ANet. The complete implementation of ANet is further simplified to the variants without using residual block and without using both residual block and gating shortcut in Eq. (12). It is found that the residual blocks in encoder and decoder and the information shortcut in decoder are equally important in implementation of ANet in terms of BLEU values. However, the number of parameters in complete ANet is increased as well.

model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SAT	65.83	45.07	30.82	21.36
SAT-KV	67.13	46.16	31.78	22.36
SAT-KVSA	68.73	48.23	33.73	23.21
ANet	69.11	48.53	33.97	23.38

TABLE I: BLEU scores by using different SATs and ANet.

model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ANet	69.11	48.53	33.97	23.38
ANet w/o res	68.23	46.23	32.31	22.66
ANet w/o res & short	67.82	46.03	32.03	21.92

TABLE II: BLEU scores for ablation study on ANet.

VI. Conclusions

This paper has investigated the importance of attention mechanisms to spotlight useful information for text transcription of an input image. The supportive attention and self attention were proposed in an encoder-decoder architecture under the show, attend and tell model and the association network model, respectively. The key memory and value memory were separate in a memory network for image representation so as to carry out the attention weights and context vector, respectively, for precise word prediction. The key-value supportive attention and key-value self attention were performed. There was no extra cost in collection of supporting data, but additional computation from baseline system was required. Supportive attention acted as an alignment or regularization for attention scheme. Association network captured the word-to-image and image-to-image attentions via multi-hop attention procedure where traditional and self attentions were both performed. Experimental results showed that both advanced attention schemes improved the performance of image caption. Supportive attention did identify the locations or evidences in word caption. Residual blocks and information shortcut were helpful in association network for image caption.

REFERENCES

[1] Dong Yu, Geoffrey Hinton, Nelson Morgan, Jen-Tzung Chien, and Shigeki Sagayama, "Introduction to the special section on deep learning for speech and language processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 4–6, 2011.

[2] Jen-Tzung Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[3] Jen-Tzung Chien, "Deep bayesian multimedia learning," in *Proc. of ACM International Conference on Multimedia*, 2020, pp. 4791–4793.

[4] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] Jen-Tzung Chien and Ying-Lan Chang, "Bayesian sparse topic model," *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 375–389, 2014.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[9] Fangyi Zhu, Zhanyu Ma, Xiaoxu Li, Guang Chen, Jen-Tzung Chien, Jing-Hao Xue, and Jun Guo, "Image-text dual neural network with decision strategy for small-sample image classification," *Neurocomputing*, vol. 328, pp. 182–188, 2019.

[10] Jen-Tzung Chien and Yi-Ting Bao, "Tensor-factorized neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1998–2011, 2017.

[11] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus, "End-to-end memory networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2440–2448.

[12] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Proc. of Annual Conference of International Speech Communication Association*, 2010, pp. 1045–1048.

[13] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, and Weisheng Dong, "Image caption generation with part of speech guidance," *Pattern Recognition Letters*, vol. 119, pp. 229–237, 2019.

[15] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[16] Jason Weston, Sumit Chopra, and Antoine Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[18] Yu-Min Huang, Huan-Hsin Tseng, and Jen-Tzung Chien, "Stochastic fusion for multi-stream neural network in video classification," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2019, pp. 69–74.

[19] Jen-Tzung Chien and Che-Yu Kuo, "Markov recurrent neural network language model," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 807–813.

[20] Jen-Tzung Chien and Chun-Wei Wang, "Self Attention in Variational Sequential Learning for Summarization," in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 1318–1322.

[21] Jen-Tzung Chien and Ting-An Lin, "Supportive attention in end-to-end memory networks," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2018, pp. 1–6.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[23] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *Proc. of European Conference on Computer Vision*, 2014, pp. 740–755.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.