This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Entangled Transformer for Image Captioning

Guang Li Linchao Zhu Ping Liu Yi Yang

ReLER, University of Technology Sydney

lguang@live.cn, linchao.zhu@uts.edu.au, pino.pingliu@gmail.com, yi.yang@uts.edu.au

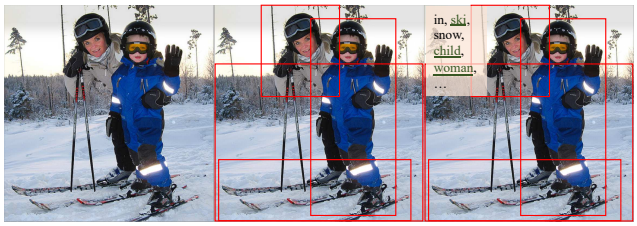
Abstract

In image captioning, the typical attention mechanisms are arduous to identify the equivalent visual signals especially when predicting highly abstract words. This phenomenon is known as the semantic gap between vision and language. This problem can be overcome by providing semantic attributes that are homologous to language. Thanks to the inherent recurrent nature and gated operating mechanism, Recurrent Neural Network (RNN) and its variants are the dominating architectures in image captioning. However, when designing elaborate attention mechanisms to integrate visual inputs and semantic attributes, RNN-like variants become unflexible due to their complexities. In this paper, we investigate a Transformer-based sequence modeling framework, built only with attention layers and feed-forward layers. To bridge the semantic gap, we introduce EnTangled Attention (ETA) that enables the Transformer to exploit semantic and visual information simultaneously. Furthermore, Gated Bilateral Controller (GBC) is proposed to guide the interactions between the multimodal information. We name our model as ETA-Transformer. Remarkably, ETA-Transformer achieves state-of-the-art performance on the MSCOCO image captioning dataset. The ablation studies validate the improvements of our proposed modules.

1. Introduction

Image captioning [39, 18] is one of the essential tasks [4, 39, 47] that attempts to break the semantic gap between vision and language. To generate good captions for images, it involves not only the understanding of many concepts, such as objects, actions, scenes, human-objects interactions but also expressing these factors and their relations in a natural language. Recently, the attention mechanism [41, 44, 12] was introduced to dynamically recap the salient information of the input image for every word.

In previous image captioning works [41, 44, 12], the attention mechanism mainly lies in two fields based on the modality of the information they employed: *Visual Attention* and *Semantic Attention*. On the one hand, visual attention exploits the low-level feature maps [41] or high-level object ROI-pooled features [29, 2] to identify the most relevant regions for the words. However, due to the semantic gap, not every word in the caption has corresponding visual signals [25], especially for the tokens associated with abstract concepts and complex relationships. Figure 1 shows an example of this obstacle. On the other hand, researchers develop the semantic attentions [44, 12] which can leverage the high-level semantic information directly. Nevertheless, because of the recurrent nature, RNNs [11, 27, 34] have difficulties in memorizing the inputs many steps ago, especially the initial visual input. Consequently, such approaches tend to collapse into high-frequency phrase fragments without regard to the visual cues.



(a). A person is standing in the snow.
(b). Two children are standing in the snow.
(c). A woman and a child are skiing in the snow.

Figure 1: The image captioning results when given different modality information. (a) provides an unsatisfactory caption result only using low-level visual features. When provided with high-level visual information guided from region proposals, (b) can make some improvement, e.g., predict "two children" in the picture. However, it still fails to grab abstract concepts in the image, e.g., "skiing". (c) is the result when utilizing information from complementary modalities: visual and semantic. It is the most accurate result among the three descriptions.

As shown in Figure 1(c), the combination of the two

complementary attention paradigms can alleviate the harmful impacts of the semantic gap. Therefore, Li *et al.* [22] propose a two-layered LSTM [17] that the visual and semantic attentions are separately conducted at each layer. Yao *et al.* [42] employ graph convolutional neural networks to explore the spatial and semantic relationships. They use late fusion to combine two LSTM language models that are independently trained on different modalities. However, due to the inherent recurrent nature and the complex operating mechanism, RNNs fail to explore the two complementary modalities concurrently.

To solve these problems above, we extend the efficient and straightforward Transformer [37] framework with our proposed Entangled Attention (ETA) and Gated Bilateral Controller (GBC) to explore visual and semantic information simultaneously. The design of ETA is inspired by the studies [7, 35] about the human visual system, showing the selection of attentive regions in human visual attention can be influenced by a prior linguistic input. To mimic this phenomenon, we use an information injection operation to infuse the input query with the information from the preliminary modality. Then the attention over the target modality can be conducted under the guidance of the preliminary modality. Subsequently, the representations of the target visual and semantic modalities propagate to the next layers under the channel-wise control of GBC.

The advantages of our method are as follows. First, the simplicity of the Transformer [37] framework relieves us from the limitations of recurrent neural networks. Second, the application of self-attention in the encoder encourages our model to explore the relationships between the detected entities. Our method can efficiently leverage the information in the target modality under the guidance of preliminary modality. Third, the proposed bilateral gating, GBC, can jointly facilitate our module to provide sophisticated control for the propagation of multimodal information. Because of the cohesiveness, our attention module can be readily applied to the Transformer without violating its parallel nature and modularity.

Our **contributions** can be summarized as follows:

(1) We devise the EnTangled Attention – a unique attention mechanism which enables the Transformer framework to exploit the visual and semantic information simultaneously.

(2) We propose the Gated Bilateral Controller – a novel bilateral gating mechanism which can provide sophisticated control for the forward propagation of multimodal information as well as their backpropagating gradients.

(3) We comprehensively evaluate our approach on the MSCOCO dataset [24], and our method achieves the state-of-the-art performance.

2. Related Work

Attention in Visual Captioning. Desipite the efforts [41, 29, 44, 12, 25, 2, 40] investigate the attention over monomodal information, many works also try to combine visual and semantic information semoutanouly. Yao *et al.* [43] prove multimodal information can contribute to the image captioning problem and investigate how to employ semantic attributes under LSTM framework. Li *et al.* [22] propose a two-layer visual-semantic LSTM which conducts visual attention and semantic attention at different layers. To explore the relationship between objects and semantic attributes, Yao *et al.* [42] apply graph convolution neural networks in the encoding stage. Tang *et al.* [36] leverage scene graph to align the relations between vision and language. Conducted only in each modality separately, these methods fail to explore the complementary nature of the visual and semantic information.

Co-attention in VQA. The widely used co-attention mechanism [26, 45, 13, 21] in visual question answering (VQA) can explore the visual and semantic information jointly. But the major concern of VQA is to identify the most relevant visual regions based on the question. Hence, the attention mechanism in VQA mainly queries the visual regions with the semantic feature. However, in image captioning, the most salient semantic attributes should also be identified.

Model Structures. The recurrent nature of RNN dilutes the long-term information at every time step [33]. To get rid of the catastrophic forgetting in long-term memory, Gu *et al.* [15] introduce temporal CNN to impose the experienced semantic information at every step of the generation procedure. Additionally, to overcome the inherently recurrent nature of the RNNs, Gehring *et al.* [14] propose to use Convolutional Neural Networks (CNN) to model the sequence-to-sequence problem. Afterward, Aneja *et al.* [3] adapt this model to image captioning. Different from the local convolution operation, whose receptive field is determined by the kernel size and layer depth, the self-attention can access the information globally. Besides, there are only a few attempts [5, 46, 31] to employ the Transformer in visual captioning.

3. Preliminary

To overcome the inherent recurrence in RNN model, the Transformer reformulate the calculation of the hidden state in Eq. 1. Thus, the hidden state of current time step $\mathbf{h}_t$ only depends on the feature embeddings of the input image and history words, rather than the previous hidden state $\mathbf{h}_{t-1}$. This formulation enables the Transformer model to execute in parallel.

$$\mathbf{h}_t = TransformerDecoder(I; \mathbf{w}_1, \ldots, \mathbf{w}_{t-1}) \quad (1)$$

To handle the variable-length inputs, such as image regions and word sequence, Transformer employs attention

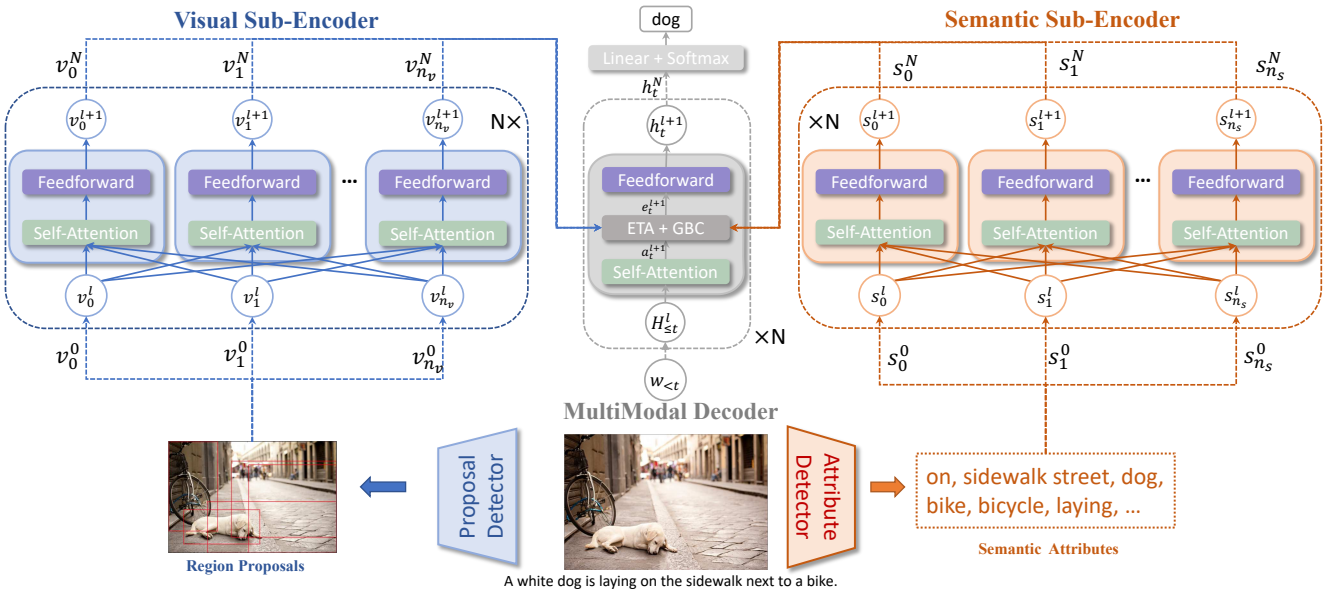


Figure 2: The overall architecture of ETA-Transformer. Our model consists of three components: the visual sub-encoder, the semantic sub-encoder, and the multimodal decoder. The generation procedure has three steps: (1) detecting region proposals and semantic attributes; (2) encoding the visual and semantic features separately; (3) decoding word by word to obtain the final caption. Notice that the Residual Connections, Layer Normalizations, and Embedding Layers are omitted.

to convert the unfixed number of inputs to a unified representation. Moreover, positional encoding [37] is employed both in the encoder and decoder to inject sequential information.

There are two particular attention mechanisms in the Transformer model. Here we start with the *scaled dot-product attention* [37], in which the inner product is applied to calculate the attention weights. Given a query $\mathbf{q}_i$ from all m queries, a set of keys $\mathbf{k}_t \in \mathbb{R}^d$ and values $\mathbf{v}_t \in \mathbb{R}^d$ where $t = 1, \ldots, n$, the scaled dot-product attention outputs a weighted sum of values $\mathbf{v}_t$, where the weights are determined by the dot-products of query $\mathbf{q}_i$ and keys $\mathbf{k}_t$. In order to implement the dot product operation by highly optimized matrix multiplication code, the queries, keys, and values are packed together into matrices $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_m)$, $\mathbf{K} = (\mathbf{k}_1, \ldots, \mathbf{k}_n)$, and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. In practice,

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \quad (2)$$

where d is the width of the input feature vectors.

To extend the capacity of exploring subspaces, Transformer employs the *multi-head attention* [37] which consists of h parallel scaled dot-product attentions named *head*. The inputs including queries, keys, and values are projected into h subspaces, and the attention performs in the subspaces seperately:

$$\begin{aligned} MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= Concat(\mathbf{H_1}, \ldots, \mathbf{H_h})\mathbf{W}^O, \\ \mathbf{H}_i &= Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{aligned} \quad (3)$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{\frac{d}{h} \times d}$ are the independent head projection matrices, $i = 1, 2, \ldots, h$ and $\mathbf{W}_i^O \in \mathbb{R}^{d \times d}$ denotes the linear transformation. Note that the bias terms in linear layers are omitted for the sake of concise expression, and the subsequent descriptions follow the same principle.

4. Methodology

In this section, we devise our ETA-Transformer model. As shown in Figure 2, the overall architecture follows the encoder-decoder paradigm. First, a dual-way encoder maps the original inputs into highly abstract representations and then the decoder incorporates the multimodal information simultaneously to generate the caption word by word.

4.1. Dual-Way Encoder

In most cases, CNNs like VGG [32] or ResNet [16] are first considered for encoding the visual information, while the transformer encoder is originally designed for sequence modeling. However, we argue that a transformer encoder with sophisticated design can better explore the inter- and intra- relationships between the visual entities and semantic attributes. Specifically, we devise a dual-way encoder

that consists of two sub-encoders. Each sub-encoder is self-attentive and of the same structure, i.e., a stack of N identical blocks.

Take the output of the l-th ($0 \leq l < N$) block $\mathbf{O}^l \in R^{d \times n}$ as an example. They are first fed into the multi-head self-attention module in the $(l+1)$-th block:

$$\mathbf{M}^{l+1} = MultiHead(\mathbf{O}^l, \mathbf{O}^l, \mathbf{O}^l), \tag{4}$$

where $\mathbf{M}^{l+1}$ is the hidden state calculated by multi-head attention. The query, key and value matrices have the same shape. Notice that the $\mathbf{O}^0$ is the output of the embedding layer.

The subsequent sub-layer is a position-wise feed-forward network (FFN) which consists of two linear transformations with a ReLU activation in between:

$$\begin{aligned} FFN(\mathbf{x}) &= \mathbf{W}_2 \cdot ReLU(\mathbf{W}_1 \cdot \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2, \\ O^{l+1} &= [FFN(\mathbf{M}^{l+1}_{:,1}); \ldots; FFN(\mathbf{M}^{l+1}_{:,n})], \end{aligned} \tag{5}$$

where $\mathbf{W}_2 \in \mathbb{R}^{d \times d_m}$, $\mathbf{W}_1 \in \mathbb{R}^{d_m \times d}$, $\mathbf{O}^{l+1} \in R^{d \times n}$ are the outputs of the $(l+1)$-th block, and $\mathbf{M}^{l+1}_{:,i}$ represents column i of matrix $\mathbf{M}$, thus the i-th feature vector. The two equivalent expressions are used interchangeably in the subsequent description. Same to [37], the residual connection and layer normalization are used after the forementioned sub-layers, and we omit them for a concise explanation.

The structure described above can be used for encoding both the visual and semantic features. Before feeding into the sub-encoder, the n_v visual features are mapped into $\mathbf{V}^0 \in \mathbb{R}^{d \times n_v}$ by a linear transformation, and the n_s one-hot semantic attributes are projected into $\mathbf{S}^0 \in \mathbb{R}^{d \times n_s}$ by an embedding layer. Furthermore, we share the word embeddings between the semantic encoder and the decoder so that our model can utilize the target information directly.

4.2. Multimodal Decoder

In addition to the basic block of the encoder, the decoder block inserts an ETA module and a GBC module between the self-attention sub-layer and the feed-forward sub-layer, which empowers the decoder block to perform attention over the visual outputs $\mathbf{V}^N$ and semantic outputs $\mathbf{S}^N$ of the dual-way encoder simultaneously. Similar to the encoder, the decoder consists of N identical blocks, and we employ residual connections around each of the sub-layers, followed by layer normalization.

Suppose the decoder is generating the t-th word in the target sentence. We denote $\mathbf{w}_t \in \mathbb{R}^{d \times 1}$ as the vector representation of the t-th word, which is the sum of word embedding and positional encoding. Therefore, the input matrix representation for time step t is:

$$\mathbf{W}_{<t} = [\mathbf{w}_0; \ldots; \mathbf{w}_{t-1}], \tag{6}$$

where $\mathbf{W}_{<t} \in \mathbb{R}^{d \times t}$ and $\mathbf{w}_0$ is the feature vector of the token representing the start of sentence.

For the $(l+1)$-th block, the inputs $\mathbf{H}^l_{\leq t} \in \mathbb{R}^{d \times t} = (\mathbf{h}^l_1, \ldots, \mathbf{h}^l_t)$ are fed into a multi-head self-attention sub-layer, notice that $\mathbf{h}^0_t$ corresponds to $\mathbf{w}_{t-1}$:

$$\mathbf{A}^{l+1}_{:,t} = MultiHead(\mathbf{H}^l_{:,t}, \mathbf{H}^l_{<t}, \mathbf{H}^l_{<t}), \tag{7}$$

where $\mathbf{H}^l_{:,t} \in \mathbb{R}^{d \times 1}$, $\mathbf{A}^l_{:,t} \in \mathbb{R}^{d \times 1}$, and $\mathbf{h}^0_t = \mathbf{w}_{t-1}$. Notice that $\mathbf{W}_{<t}$ is the inputs for the first layer. Subsequently, the self-attention output $\mathbf{a}^{l+1}_t$ is passed into the ETA to incorporate with visual and semantic features:

$$\mathbf{E}^{l+1}_{:,t} = ETA(\mathbf{A}^{l+1}_{:,t}, \mathbf{V}^N, \mathbf{S}^N), \tag{8}$$

where $\mathbf{E}^{l+1}_{:,t} \in \mathbb{R}^{d \times 1}$ contains the visual and semantic information which is elaborately integrated according to the importance of modalities in channel level. After the process of FFN, we obtain the output $\mathbf{h}^{l+1}_t = FFN(\mathbf{e}^{l+1}_t)$ of current layer.

Finally, the output of layer N is fed into the classifier over vocabulary to predict next word. Notice that the procedure described above illustrates the incremental generation in inference. Because all the input tokens are known in the training stage, the attention is implemented with highly optimized matrix multiplication.

4.3. EnTangled Attention

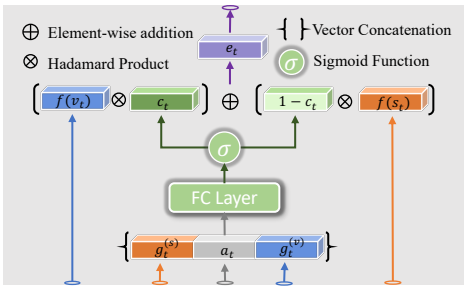
Most of the previous attempts trying to integrate multimodal information for image captioning only perform attention over the multiple modalities separately and then fuse the independent attention representations. Therefore, they fail to leverage the complementary nature of visual and semantic information in attention operations. Differently, as shown in Figure 3 (b), we implement the attention in an entangled manner so that it can be affected by the preliminary modality while performing attention over the target one.

Here we take the visual pathway in ETA as an illustration. To mimic the attention mechanism of the human vision system, we need a function which can inject the information of preliminary modality $\mathbf{S}^N$ into the self-attention output $\mathbf{a}_t$ (see Eq. 7) so that the generated representation $\mathbf{g}^{(s)}_t \in \mathbb{R}^{d \times 1}$ (the superscript (s) is donated for the sign of modality) can provide proper guidance for the attention in target modality. In order to handle the variable number of semantic attributes, we choose multi-head attention as the preliminary information injection function:

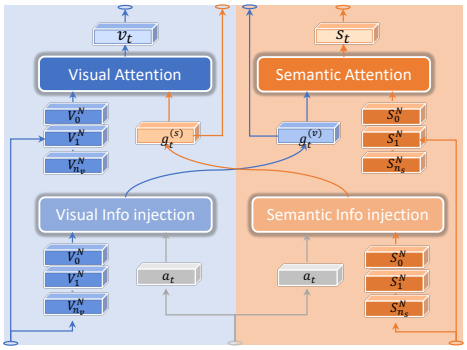
$$\mathbf{g}^{(s)}_t = MultiHead(\mathbf{a}_t, \mathbf{S}^N, \mathbf{S}^N). \tag{9}$$

Next, we use the semantic guidance $\mathbf{g}^s_t$ to perform multi-head attention over the target modality $\mathbf{V}^N$:

$$\mathbf{v}_t = MultiHead(\mathbf{g}^{(s)}_t, \mathbf{V}^N, \mathbf{V}^N), \tag{10}$$



(a) Gated Bilateral Controller



(b) EnTangled Attention

Figure 3: The multimodal representations are first fed into ETA to conduct EnTangled Attention, then to GBC to obtain the final representation.

where $\mathbf{v}_t \in \mathbb{R}^{d\times 1}$ is the final representation generated with the guidance of semantic modality. And in a similar manner but reversed order, we could obtain the semantic representation $\mathbf{s}_t \in \mathbb{R}^{d\times 1}$. Notice that all the attention layers in ETA are followed with residual connection and layer normalization which are omitted for concise expression.

4.4. Gated Bilateral Controller

In this section, we present the *Gated Bilateral Controller* (GBC) specially designed for the integration of the generated representations $\mathbf{s}_t$ and $\mathbf{v}_t$. The gating mechanisms controlling the path through which information flows to the subsequent layers are widely used in the famous sequence models like LSTM [17], GRU [6], and ConvS2S [14]. Such multiplicative gates are adept at dealing with gradient explosion and vanishing, which enable the information to propogate unimpededly through long timesteps or deep layers. As illustrated in Figure 3 (a), the context gate $\mathbf{c}_t$ in GBC is determined by the current self-attention output $\mathbf{a}_t$, the visual guidance $\mathbf{g}^{(v)}$ and the semantic guidance $\mathbf{g}^{(s)}$:

$$\mathbf{c}_t = \sigma\left(\mathbf{W}_c \cdot [\mathbf{g}_t^{(s)}, \mathbf{g}_t^{(v)}, \mathbf{a}_t]\right), \tag{11}$$

where $\mathbf{c}_t \in \mathbb{R}^{d\times 1}$, $\mathbf{W}_c \in \mathbb{R}^{d\times 3d}$ and $\sigma(\cdot)$ denotes the sigmoid function.

Different from the previous gating mechanism managing only one pathway, we extend it with a bilateral scheme. The gate value $\mathbf{c}_t$ controls the flow of visual guidance $\mathbf{v}_t$ while the complement part $(1-\mathbf{c}_t)$ governs the propagation of semantic information $\mathbf{s}_t$:

$$\mathbf{e}_t = f(\mathbf{v}_t) \odot \mathbf{c}_t + f(\mathbf{s}_t) \odot (1-\mathbf{c}_t), \tag{12}$$

where $\odot$ represents the hadamard product, $f(\cdot)$ can be an activation function or identity function, and $\mathbf{e}_t \in \mathbb{R}^{d\times 1}$ denotes the output of ETA.

The Effect of f Function. In LSTM or GRU, the left part of the Hadamard product is always activated with function f which can be Sigmoid, Tanh or ReLU [20], *etc.* Whereas, we do not apply any activation over v_t and s_t which are merely the outputs of the linear transformation in multi-head attention. Compared with the saturate activations mentioned above, the identity function $id(x) = x$ allows gradients to propagate through the linear part without downscaling. Here, following the analysis in [8], we take the left part of the Eq. 12 as an example, whose gradient is:

$$\nabla\left[f(\mathbf{x}) \odot \mathbf{c}_t\right] = f'(\mathbf{x})\nabla\,\mathbf{x} \odot \mathbf{c}_t. \tag{13}$$

As shown in the Eq. 13, the $f'(x)$ can act as a scale factor of the gradients. Additionally, $tanh'(\cdot) \in (0, 1]$, $\sigma'(\cdot) \in (0, 0.25]$, while $id'(\cdot) = 1$. Thus, the saturate activations will downscale the gradient and make gradient vanishing even worse with the stacking of layers. Although the non-saturate activation ReLU has similar property with identity function, here we argue the activated gate $\mathbf{c}_t$ has equipped the module with non-linearity [9]. For the principle of simplicity, we do not apply any activations over $\mathbf{v}_t$ and $\mathbf{s}_t$. By comparing the effect of f function experimentally in Section 5.4.3, we find the activations deteriorate the performance greatly while the identity function achieves the best.

5. Experiments

5.1. Datasets and Evaluation

We use the MSCOCO 2014 captions dataset [24] to evaluate our proposed captioning model. In offline testing, we use the Karpathy splits [18] that have been used extensively for reporting results in previous works. This split contains 113,287 training images with five captions each, and 5K images respectively for validation and testing. Our MSCOCO test server submission is trained on the Karpathy's training split, and chosen on the Karpathy's test split.

Data processing We follow standard practice and perform only minimal text pre-processing, converting all sentences to lower case, tokenizing on white space, and keeping words that occur at least five times, resulting in a model vocabulary of 9,487 words. To evaluate caption quality,

	Proposal	Semantic	Cross-Entropy Loss						Sequence-Level Optimization					
			B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
SCST [30]	✗	✗	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
LSTM-A [43]	✗	✓	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
VS-LSTM [22]	✓	✓	76.3	34.3	26.9	-	110.2	-	78.9	36.3	27.3	-	120.8	-
Up-Down [2]	✓	✗	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM$_{fuse}$ [42]	✓	✓	77.4	37.1	28.1	57.2	117.1	21.1	80.9	38.3	28.6	58.5	**128.7**	22.1
ETA	✓	✓	77.3	37.1	28.2	57.1	117.9	21.4	81.5	39.3	28.8	58.9	126.6	**22.7**
ETA$_{fuse}$	✓	✓	**77.6**	**37.8**	**28.4**	**57.4**	**119.3**	**21.6**	**81.5**	**39.9**	**28.9**	**59.0**	127.6	22.6

Table 1: MSCOCO Offline Evaluation. The ETA denotes the ETA-Transformer. ✓ indicates the corresponding features (region proposals or semantic attributes) are applied, and ✗ means otherwise. All values are reported as percentage (%).

	B@1	B@4	M	R	C	S
VS-LSTM$_s$	74.3	33.3	26.5	-	105.1	-
VS-LSTM$_v$	75.1	33.5	26.5	-	105.8	-
VS-LSTM	76.3	34.3	26.9	-	110.2	-
GCN-LSTM$_s$	77.3	36.8	27.9	57.0	116.3	20.9
GCN-LSTM$_v$	77.2	36.5	27.8	56.8	115.6	20.8
GCN-LSTM$_{fuse}$	**77.4**	**37.1**	28.1	**57.2**	117.1	21.1
Transformer$_s$	71.1	29.0	25.3	52.8	96.2	18.2
Transformer$_v$	75.9	34.0	27.5	56.1	112.2	21.0
ETA	77.3	**37.1**	**28.2**	57.1	**117.9**	**21.4**
ETA^{oracle}	97.0	76.7	47.9	84.2	204.2	34.7

Table 2: The results on single modality. The ETA denotes the ETA-Transformer. Subscript indicates that the visual modality or semantic modality is applied.

we use the standard automatic evaluation metrics, namely SPICE [1], CIDEr-D [38], METEOR [10], ROUGE-L [23] and BLEU [28].

5.2. Implementation Details

Visual & Semantic Features. For visual features, we use the region proposals as the visual representations. To select the salient regions, we follow the settings in Up-Down [2]. When comparing with some previous methods [18, 3], we also encode the full-sized input image with the final convolutional layer of VGG-16 [32] and use adaptive pooling to resize the outputs into a fixed size of 7x7. For semantic features, we follows the settings of Fang *et. al* [12] to detect semantic attributes. The backbone of attribute detector is fine-tuned from VGG16 equipped with a noisy-OR version of multiple instance loss. We only keep the top-1000 frequent words as labels. And in the training stage, we use the detected semantic attributes rather than the ground truth.

Model Settings & Training. We follow the same hyper-parameter settings in [37]. We use $N = 6$ identical layers in both encoder and decoder. The output dimension of the word embedding layers is 512, and the input visual features are also mapped into 512 with a linear projection. The inner-layer of the feed-forward network has dimentionality $d_m = 2048$. And $h = 8$ parallel attention layers are employed in multi-head attention. Besides, we also share the word embedding between semantic sub-encoder and the decoder in order to leverage the target word representation directly. In training stage, we use the same learning rate schedule as [37]. The input batch size is 75 image-sentence pairs and the warm-up step is 20000. We use the Adam optimizer [18] with $\beta_1 = 0.9$, $\beta_2 = 0.98$.

5.3. Comparision with State-of-the-Art Methods

Offline Evaluation. Table 1 shows the performance of our model and state-of-the-art approaches in recent two years. Note that the comparative methods are all based on LSTM and its variants, which is the dominant framework in image captioning. All the baselines adapt ResNet-101 as the backbone network of visual representation. The self-critical sequence-level training strategy devised in SCST [30] is applied by Up-Down [2], GCN-LSTM [42] and ETA-Transformer for optimizing the CIDEr-D score, while VS-LSTM [22] employs an improved version of SCST. LSTM-A [43] investigates how to utilize the predicted semantic attributes efficiently. We use them as the LSTM baselines. Up-Down [2] presents a two-layer LSTM to conduct attention over bottom-up and top-down visual features separately. VS-LSTM [22] use a similar design but replace the low-level visual features with semantic attributes. Restricted by the complexity of LSTM, the models have difficulties in stacking deep layers. Benefits from the scalability of the Transformer and the cohesiveness of our proposed modules, the multimodal attention can be conducted at different levels of abstraction. In our experiments, we employ N=6 multimodal attentions in the decoding stage. Thus, our method outperforms them with a large margin. Aiming at modeling the relations of objects, GCN-LSTM [42] introduced graph convolutional neural network to encode the detected entities. To make fair comparison, we also provide the late-fused performance of two models with different initialization. The result shows that our model achieves superior performance on the cross-entropy training. And in sequence level training, our model produces higher performance in five out of six metrics, especially the BLEU@4(39.9%) and SPICE(22.7%).

To provide a more detailed comparison, we also report

Model	B@1		B@2		B@3		B@4		M		R-L		C-D	
-	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.3	56.3	70.7	114.7	116.0
LSTM-A	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
VS-LSTM	78.8	94.6	62.8	87.5	47.9	77.3	35.9	66.3	27.0	35.3	56.5	70.3	116.6	119.5
Up-Down	80.2	**95.2**	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GCN-LSTM	-	-	**65.5**	**89.3**	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	**125.3**	**126.5**
ETA	**81.2**	95.0	**65.5**	89.0	**50.9**	**80.4**	**38.9**	**70.2**	**28.6**	**38.0**	**58.6**	**73.9**	122.1	124.4

Table 3: MSCOCO Online Evaluation. The ETA denotes the ETA-Transformer. cX means evaluation on X captions. All values are reported as percentage (%).

	B@1	B@4	M	R	C	S
LSTM [18]	71.3	30.3	24.7	52.5	91.2	17.2
Convolution [3]	71.1	28.7	24.4	52.2	91.2	17.5
Transformer$_s$	71.1	29.0	25.3	52.8	96.2	18.2
- Encoder	70.3	28.5	24.8	52.0	93.1	17.7
Transformer$_v$	71.0	30.2	24.9	52.6	93.8	18.0
- Encoder	70.2	28.2	24.2	51.6	91.8	17.2
ETA	**72.2**	**31.9**	**25.7**	**53.4**	**99.2**	**18.6**

Table 4: Comparison with different model structures. And "-Encoder" implies the Encoder is removed from the model. All results are reported in token-level training.

	B@1	B@4	M	R	C	S
Transformer$_v$	80.6	38.3	28.5	58.3	124.1	22.3
Transformer$_s$	76.6	32.6	25.5	54.4	102.8	19.1
T$_v$ & T$_{s\ fuse}$	79.6	37.5	27.6	57.6	118.7	19.8
Parallel	80.9	38.7	28.8	58.7	124.9	22.4
Stacked$_v$	80.7	39.1	28.6	58.6	125.0	22.4
Stacked$_s$	80.8	38.8	28.6	58.5	124.5	22.5
ETA	**81.5**	**39.3**	**28.8**	**58.9**	**126.6**	**22.7**

Table 5: Ablation experiments. ETA is denotes the ETA-Transformer. And all results are trained on sequence-level criterion.

the results on single modality. ETA-Transformer and VS-LSTM use weak semantic labels generated from the ground truth captions (see [12] for more details), but the GCN-LSTM employs a fully-supervised model trained on the region-level annotations of Visual Genome [19]. Therefore, the GCN-LSTM$_s$ has superior performance to the Transformer$_s$ and VS-LSTM$_s$. However, as shown in Table 2, our model provides the most significant improvements when combining the two modalities. This comparison further proves the effectiveness of our proposed modules in leveraging the complementary information. We also report the performance of our model under an Oracle setting (see the ETA^{oracle}), where the semantic attributes tokenized from the ground truth captions are provided during test time. This can be viewed as the upper bound of our method when we have a perfect attribute detector.

Online Evaluation. We ensembled three models trained on sequence-level criterion with different initalization, and submitted our results to the online testing server. Table 3 includes the top-5 methods which have been officially published, and it shows that the ETA-Transformer is among the top-2 performance over all the metrics. In particular, the B@3, B@4, METEOR, and ROUGE-L are superior on both c5 and c40 testing sets. The submission results named *ETA-Transformer* have been public on the leaderboard [1].

[1] https://competitions.codalab.org/competitions/3221

5.4. Ablation Study

5.4.1 Comparison with Different Frameworks

In the ablation study, we first compare the Transformer with the other two classical sequence model LSTM [41] and ConvS2S [3, 14]. The two models are all equipped with visual attention mechanism. Following the feature extraction settings in [3], we use 7x7 feature maps of the fifth convolution layer in VGG-16 as our visual representations. The performance on Table 4 shows that the standard Transformer is comparable with LSTM and ConvS2S model in the image captioning problem.

Further, to validate the previous declaration that the self-attention can benefit the feature representation from modeling the relationships of input entities, we report the results of the encoder-removed transformer on both modalities. As shown in Table 4, the performance has dropped significantly over all the metrics.

5.4.2 Comparison with Strong Baselines

In this section, we provide other two simplified versions of our proposed modules, and the late-fusion of Transformer$_s$ (T$_s$) & Transformer$_v$ (T$_v$), as strong baselines. In the first one, we remove the GBC module and extract one pathway of ETA as the first version. We refer this version as Stacked Attention (SA) because it has two stacked multi-head attentions. In the second one, we remove the preliminary information injection blocks in ETA and simply use GBC to integrate the outputs of encoder (S^N and V^N) directly.

Figure 4: Qualitative examples of different methods. Compared with Transformer$_v$ (T$_v$) and Transformer$_s$ (T$_s$), the ETA-Transformer (ETA) generates more descriptive and more accurate captions.

	B@1	B@4	M	R	C	S
Sigmoid	74.5	32.1	26.3	54.8	104.9	19.4
Tanh	74.8	32.0	26.2	54.8	104.1	19.6
ReLU	**76.3**	36.1	27.9	56.2	114.0	20.8
Linear	**76.3**	**36.3**	**28.1**	**56.5**	**115.2**	**21.0**

Table 6: The effect of activation functions in GBC. All results are reported in token-level training.

This version is named as Parallel Attention (PA). In T$_s$ & T$_{v\ fuse}$, we train the two standard transformer model separately and late-fused the results of them.

The late fusion of monomodal models can only have limited gains, sometimes, even severe degeneration. As shown in Table 5, the performance of T$_s$ & T$_{v\ fuse}$ is worse even compared with T$_v$. This is mainly caused by the inferior single model T$_s$. Differently, the ETA-Transformer, which integrates the multimodal information at the feature level, obtains significant and stable improvement in performance.

In Table 5, compared results of the multimodal versions with Transformer$_s$ or Transformer$_v$, we can find that visual and semantic modalities are complementary. The integration of visual and semantic information can contribute to better performance despite that the semantic representations are considerably worse than the visual representations. Notwithstanding the huge performance gap between Transformer$_v$ and Transformer$_s$, SA$_s$ and SA$_v$ (see the Stacked$_s$ and Stacked$_v$ in Table 5 have near performance on all the metrics. These experimental results show the EnTangled Attention can benefit from fusing the visual and semantic information with an ordered manner. Besides, the widely used skip connection, which equally combines the preliminary and target representations without any adaptive trade-off, sustains the impact of the preliminary modality. Thus the performance of semantic information is enhanced.

Without using the EnTangled Attention mechanism, the parallel attention only employs the gated bilateral controller to combine the encoded visual and semantic representations adaptively. And PA gains comparable and slightly better performance than SA$_s$ and SA$_v$. Furthermore, The ETA can be viewed as the combination of PA and SA, which incorporates the advantages of both. Shown in Table 5, the ETA achieves superior performances against the two strong baselines in all the metrics noteworthily.

5.4.3 The Effect of Activation in GBC

As shown in Table 6, the saturated activation functions like Sigmoid and Tanh deteriorate the performance of GBC significantly, while the identity function and the non-saturated activation ReLU do not suffer from this degeneration. The identity function only outperforms ReLU slightly. Following the analysis in 4.4, bacause $tanh'(\cdot) \in (0, 1]$ has a larger range compared with $\sigma'(\cdot) \in (0, 0.25]$, Tanh should outperform Sigmoid. We think that the saturated area, where the gradients are close to zero, occupies most of the feasible domain in saturated activation functions – consequently, Tanh still suffers serious deterioration as Sigmoid.

Further, we compare the design principle of the gating mechanism between RNN and Transformer. For RNN, the supervision information is provided for every time step. Thus the gating mechanism should be able to restrict gradient explosion in the backpropagation through time. Differently, the supervision only provided in the last layer of the Transformer Decoder, where the gradient vanishing becomes the dominant problem. Therefore, the identity function should be considered first when stacking deep layers.

6. Conclusion

In this work, we devise an effective multimodal sequence modeling framework for image captioning. By introducing the EnTangled Attention and Gated Bilateral Controller, the Transformer model is extended to exploit complementary information of visual regions and semantic attributes simultaneously. Moreover, comprehensive comparisons with state-of-the-art methods and adequate ablation studies demonstrate the effectiveness of our framework.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 2016. 6

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *ICCV*, 2018. 1, 2, 6

[3] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *CVPR*, 2018. 2, 6, 7

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1

[5] Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. Tvt: Two-view transformer network for video captioning. In *ACML*, 2018. 2

[6] Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. 2014. 5

[7] Roger M Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 1974. 2

[8] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, 2017. 5

[9] Yann N. Dauphin and David Grangier. Predicting distributions with linearizing belief networks. In *ICLR*, 2016. 5

[10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *SMT-W*, pages 376–380, 2014. 6

[11] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. 1

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015. 1, 2, 6, 7

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2

[14] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017. 2, 5, 7

[15] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. An empirical study of language cnn for image captioning. In *ICCV*, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 2, 5

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 5, 6, 7

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 7

[20] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *CoRR*, abs/1504.00941, 2015. 5

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 2

[22] Nannan Li and Zhenzhong Chen. Image cationing with visual-semantic lstm. In *IJCAI-18*, 2018. 2, 6

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 6

[24] 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 *ECCV*. Springer, 2014. 2, 5

[25] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *ICCV*, 2018. 1, 2

[26] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Neurips*, 2016. 2

[27] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010. 1

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*. Association for Computational Linguistics, 2002. 6

[29] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *ICCV*, 2017. 1, 2

[30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024. 6

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3, 6

[33] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Neurips*, 2015. 2

[34] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML-11*, pages 1017–1024, 2011. 1

[35] Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 1995. 2

[36] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neurips*, 2017. 2, 3, 4, 6

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 6

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1

[40] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM MM*. ACM, 2018. 2

[41] 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 *ICML*, 2015. 1, 2, 7

[42] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 2, 6

[43] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017. 2, 6

[44] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 1, 2

[45] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017. 2

[46] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 2

[47] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *IJCV*. 1