Bridge the Gap: High-level Semantic Planning for Image Captioning

Chenxi Yuan and **Yang Bai**
Department of Computer Science
and Technology
Tsinghua University
ycx18@mails.tsinghua.edu.cn
bai-y18@mails.tsinghua.edu.cn

Chun Yuan *
Tsinghua ShenZhen International
Graduate School
Tsinghua University
Peng Cheng Laboratory
yuanc@sz.tsinghua.edu.cn

Abstract

Recent image captioning models have made much progress for exploring the multi-modal interaction, such as attention mechanisms. Though these mechanisms can boost the interaction, there are still two gaps between the visual and language domains: (1) the gap between the visual features and textual semantics, (2) the gap between the disordering of visual features and the ordering of texts. To bridge the gaps we propose a high-level semantic planning (HSP) mechanism that incorporates both a semantic reconstruction and an explicit order planning. We integrate the planning mechanism to the attention based caption model and propose the High-level Semantic PLanning based Attention Network (HS-PLAN). First, an attention based reconstruction module is designed to reconstruct the visual features with high-level semantic information. Then we apply a pointer network to serialize the features and obtain the explicit order plan to guide the generation. Experiments conducted on MS COCO show that our model outperforms previous methods and achieves the state-of-the-art performance of 133.4% CIDEr-D score.

1 Introduction

Image captioning which aims to generate textual descriptions of images, is a significant task in both computer vision and natural language process. It not only requires recognizing and understanding the objects and attributes from the given image but also needs to verbalize them with natural language in proper order.

Previous works with neural models follow the encoder-decoder paradigm that uses Convolutional Neural Network (CNN) to encode the input image and apply Recurrent Neural Network (RNN) as decoder to generate the textual descriptions (as shown in Figure 1(a)) (Vinyals et al., 2015; Gan et al., 2017a; Chen et al., 2018; Gan et al., 2017b; Lu et al., 2017; Yang et al., 2016). To explore the multi-modal interaction between the visual content and textual description, some recent methods (Xu et al., 2015; Anderson et al., 2018) apply visual attention mechanism to model the interaction. The visual attention works by learning to selectively attend to image features extracted by the encoder when generating each word. For better interaction, a large number of works focus on boost the performance of neural models with improved attention mechanisms (Huang et al., 2019a; Huang et al., 2019b; Pan et al., 2020). However, there are still **two gaps** between the visual and language domains that visual attention does not address: **(1)** the gap between the visual features and textual semantics, **(2)** the gap between the disordering of visual features and the ordering of texts. For one thing, it is hard for the decoder to associate each word in the caption with the features without a high-level semantic understanding. For another thing, with the visual attention these neural models implicitly select which features to focus on at each decoding step without any explicit guidance or exterior supervision, which makes the generation process uncontrollable and inexplicable.

There have been some researches focusing on alleviating both of the two problems. Some of them apply semantic attention to leverage the high-level semantic information to narrow the first gap (Fang

*Corresponding author: yuanc@sz.tsinghua.edu.cn

Proceedings of the 28th International Conference on Computational Linguistics, pages 3157–3167
Barcelona, Spain (Online), December 8-13, 2020

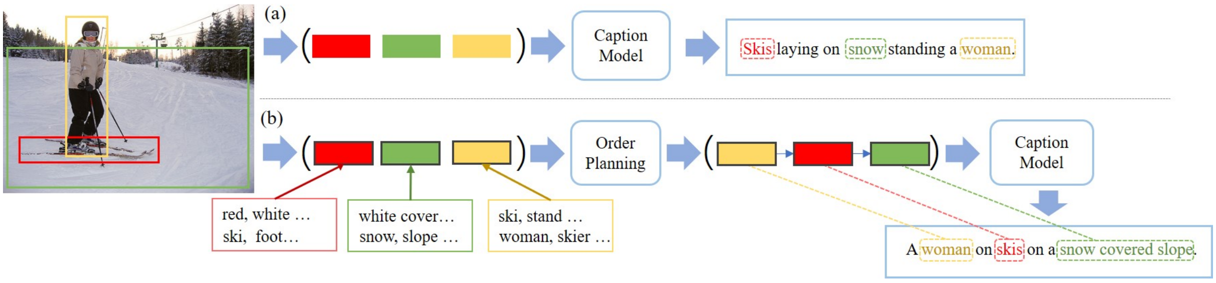


Figure 1: Comparison between the caption model without planning (a) and planning based model (b) where the visual features are first reconstructed with high-level concepts and then serialized with order planning. Thus the visual features can be grounded to the words in the caption (the dotted lines).

et al., 2015; You et al., 2016). Wu et al.(2016) explicitly represent high-level semantic concepts and incorporate them into the CNN-RNN approach. Li et al. (2019) propose the Entangled Attention based on the Transformer architecture (Vaswani et al., 2017) to explore visual and semantic information simultaneously. These methods focus more on leveraging the high-level semantic information to enhance neural models but pay little attention to the relation between the detected semantic information and the extracted visual features. For the second gap, Cornia et al. (2019) propose a controllable framework that can generate captions grounded on a sequence of image regions which are sorted by a sorting network. Though achieving the controllability to some extent, the method struggles to avoid the problems of inflexibility and error propagation between the sorting and generation.

To address these issues mentioned above, we propose a High-level Semantic PLanning based Attention Network (HS-PLAN) that incorporates both a **high-level semantic reconstruction** and an **explicit order planning** as shown in Figure 1 (b). (1) To narrow the gap between the visual features and textual semantics, an attention based reconstruction module is designed to re-represent the visual feature of each image region with the corresponding high-level concepts predicted by the object detector and attribute classifier. (2) To bridge the gap between the disordering of visual features and the ordering of textual sentences, we implement an attention based pointer network to make explicit order-plan to guide the caption generation. After the planning stage, the planned features are fed to an attention based caption model. The caption model first applies an order-sensitive encoder to encode the planned features further and learn the absolute and relative order information of features with position encoding. Then a visual attention based decoder is employed to generate the textual description of the input image guided by the determined plan.

We conduct experiments on a large benchmark dataset named MS COCO (Lin et al., 2014) to evaluate our proposed model. The results show that our model outperforms all the baselines and achieves the state-of-the-art performance: achieving 133.4% CIDEr-D score with a single model and 134.8% with an ensemble of four models on “Karpathy” test split. The qualitative human evaluation also demonstrates that our model can generate more fluent, faithful and coherent captions.

2 Related Work

Begin with show and tell(Vinyals et al., 2015), numbers of neural-based encoder-decoder models are proposed for image captioning. They utilize CNN-RNN based frameworks by encoding images into features and then translating image features into sentences, and achieve significant improvements on captioning. Recently attention mechanisms are widely used in image captioning, which provide guidance for choosing the most relevant image region when generating words of sentences (Xu et al., 2015; Anderson et al., 2018; Huang et al., 2019a; Huang et al., 2019b; Pan et al., 2020). Specifically, Huang et al. (Huang et al., 2019a) propose an enhanced attention mechanism to determine the relevance of attention results for better multi-modal interaction. Moreover, (Rennie et al., 2017) applies reinforcement learning with a self-critical reward to models for a more efficient training process. However, these methods are limited to the generation of the word in sentences from image features. It is still hard for these methods to bridge

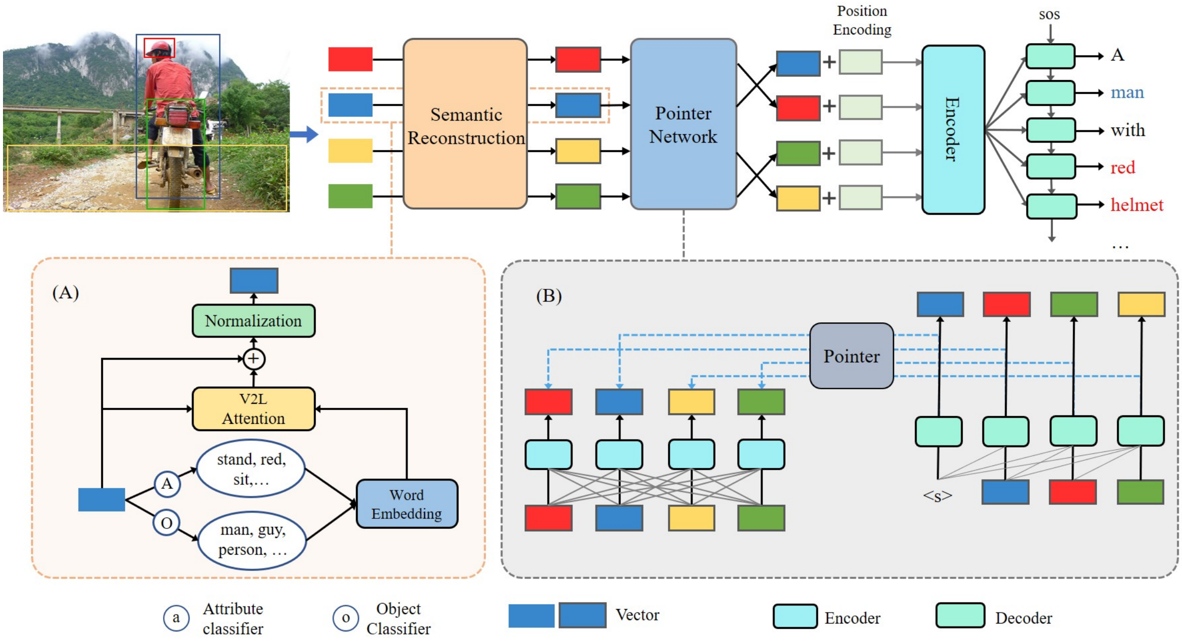


Figure 2: The architecture of HS-PLAN, where (A) is the semantic reconstruction module and (B) is a multi-head attention based pointer network used for order planning.

the gaps between the visual and language domains.

Previous captioning approaches focus on two different dimensions to alleviate the problems. Some focus on a better understanding of images or presentation of image features with high-level semantic information (Fang et al., 2015; You et al., 2016). Specifically, Wu et al.(2016) explicitly represent high-level semantic concepts and incorporate them into the CNN-RNN approach. Yang et al. (2019) leverage scene graph for more meaningful semantic representation to transfer the inductive bias from the pure language domain to the vision-language domain. Li et al. (2019) propose the Entangled Attention based on the Transformer architecture (Vaswani et al., 2017) to explore visual and semantic information simultaneously. And others concentrate more on the controllability of the generation stage. Cornia et al. (2019) propose a controllable framework that can generate captions grounded on a sequence of image regions which are sorted by a sorting network. Different from it, our method is flexible that just makes an explicit order plan to guide the generation instead of generates words step-by-step depending on the control signal in a fixed order.

3 Methodology

In this section, we devise HS-PLAN to model explicit high-level planning to guide the image captioning. The target of captioning model is to generate a textual sentence $\mathcal{Y} = \{y_1, y_2, ..., y_T\}$ of the given image $\mathcal{I}$. Traditional encoder-decoder models formulate the problem as a two-stage process: feature extraction and caption generation. But our model further decompose the problem into a three-stage process:

$$\mathcal{I} \rightarrow \mathcal{V} \rightarrow \mathcal{Y} \;\Rightarrow\; \mathcal{I} \rightarrow \mathcal{V} \rightarrow \mathcal{Z} \rightarrow \mathcal{Y} \tag{1}$$

where $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_n\}$ represents the visual features captured by the CNN-based encoder, $\mathcal{Z}$ is the explicit plan and $\mathcal{V} \rightarrow \mathcal{Z}$ represents the planning stage.

The architecture of HS-PLAN is shown in Figure 2. After extracting visual features from a given image, our model first applies a semantic reconstruction module to integrate textual semantic information into visual features to re-represent them. Then an attention-based pointer network is applied to make explicit order-plan to guide the caption generation. After the planning stage, an order-sensitive encoder is employed to further encode the features which are then used for generating textual descriptions with the decoder guided by the determined plan.

3.1 High-level Planning

3.1.1 Semantic Reconstruction

Given the image, first the visual features $\boldsymbol{v} \in \mathbb{R}^{n \times d_v}$ are extracted by a pre-trained Faster-RCNN (Ren et al., 2015), which is also used as the object detector to determine the object of each image region. Further, we use an attribute classifier to detect the possible attributes of each object. Then an attention-based reconstruction module (Figure 2.(A)) is designed to integrate the information of textual semantics into the visual features to narrow the gap between the textual and visual semantics. The textual description of each feature $\boldsymbol{v}_i$ is presented as a bag-of-words $w = \{o_1, o_2, o_3, ...a_1, a_2, a_3, ...\}$ including the possible objects and attributes, which are first embedded as word vectors $\boldsymbol{w} \in \mathbb{R}^{m \times d_w}$ where m is the scale of the bag-of-words and d_w is the dimension of word embedding. Then we design a vision-to-language attention (V2L) to estimate the similarity between the visual feature and the word embeddings to reconstruct the feature, which is computed as follows:

$$\boldsymbol{v}'_i = \mathrm{ReLU}(\boldsymbol{W}_v \boldsymbol{v}_i + \boldsymbol{b}_v), \tag{2}$$

$$\alpha_{i,j} = \frac{\exp(\boldsymbol{v}'^{\top}_i \boldsymbol{w}_{i,j})}{\sum_j \exp(\boldsymbol{v}'^{\top}_i \boldsymbol{w}_{i,j})}, \tag{3}$$

$$\boldsymbol{v}^{att}_i = \sum_j \alpha_{i,j} \boldsymbol{w}_{i,j}, \tag{4}$$

where $\boldsymbol{w}_{i,j}$ is the j-th word in the bag-of-words, $\boldsymbol{W}_v \in \mathbb{R}^{d_w \times d_v}, \boldsymbol{b}_v \in \mathbb{R}^{d_w}$ are parameters. After a linear layer and layer normalization, the feature is re-represented by integrating the information of the bag-of-words into the visual feature.

$$\boldsymbol{v}^r_i = \mathrm{LayerNorm}(\boldsymbol{v}'_i + \mathrm{ReLU}(\boldsymbol{W}_w \boldsymbol{v}^{att}_i + \boldsymbol{b}_w)), \tag{5}$$

where $\boldsymbol{W}_w \in \mathbb{R}^{d_w \times d_w}, \boldsymbol{b}_w \in \mathbb{R}^{d_w}$ are parameters. Finally each visual feature is reconstructed to a more informative one which we refer to as semantic feature.

3.1.2 Order Planning

After reconstructing the features, an attention based pointer network is designed to make explicit order plan to guide the captioning. As shown in Figure 2, the pointer network is a multi-head attention based encoder-decoder architecture with a designed pointer attention module to serialize the semantic features.

First a multi-head attention based encoder which is order-insensitive is applied to encode the features where a multi-head self-attention layer is used to capture the dependency between different image regions:

$$\boldsymbol{e}^p = \mathrm{Multihead}(\boldsymbol{v}^r, \boldsymbol{v}^r, \boldsymbol{v}^r), \tag{6}$$

where Multihead represents the multi-head attention which takes queries, keys and values as inputs and consists of h parallel scaled dot-product attentions performing in different sub-spaces separately:

$$\mathrm{Multihead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}(\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_h)\boldsymbol{W}^o, \tag{7}$$

$$\boldsymbol{H}_i = \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}^Q_i, \boldsymbol{K}\boldsymbol{W}^K_i, \boldsymbol{V}\boldsymbol{W}^V_i), \tag{8}$$

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}})\boldsymbol{V}, \tag{9}$$

where $\boldsymbol{W}^o \in \mathbb{R}^{d_w} \times d_w, \boldsymbol{W}^Q_i, \boldsymbol{W}^K_i, \boldsymbol{W}^V_i \in \mathbb{R}^{d_k \times d_k}$ are parameters and $d_k = d_w/h$.

Then a multi-head attention based decoder is used to decode and predict the order of features. The decoder is order-sensitive which implements a masked multi-head self-attention layer to capture the dependency from the predicted features to predict the next feature. The output of decoder is then fed into the pointer attention, which is designed to point to the input features one-by-one to serialize them.

A plan is a sequence of features $\mathcal{Z} = \{z_1, z_2, ..., z_n\}$ with a certain order where n is the number of features extracted from the input image. The probability of $P(z_t = \mathcal{V}_i | z_{<t}, \mathcal{V})$ is modeled as an attention over the input features as follows:

$$P(z_t = \mathcal{V}_i | z_{<t}, \mathcal{V}) = \frac{\exp(\boldsymbol{h}_t \boldsymbol{W}_p \boldsymbol{e}_i^p)}{\sum_i \exp(\boldsymbol{h}_t \boldsymbol{W}_p \boldsymbol{e}_i^p)}, \tag{10}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{d_w \times d_w}$ are parameters of the pointer attention, $\boldsymbol{h}_t$ is the hidden state of the t-th decoding step of the pointer decoder. Following the obtained probabilities the disordered features are finally serialized to a sequence $\boldsymbol{v}^p \in \mathbb{R}^{n \times d_w}$ for caption generation.

The detected image regions can be grounded to the words in the caption according to the detected objects as illustrated in Figure 1. According to the corresponding relationship, we can assume the order that the word appears in the caption is the order of the relative feature. Following this rule we can obtain the oracle order of each feature by aligning the detected objects of different image regions to the words in the caption. With the oracle order-plan, we can train the order-planning stage supervised.

3.2 Caption Model

After the high-level planning, a caption model is applied to generate the textual description guided by the determined plan, where an order-sensitive encoder is applied to further encode the planned features and to capture the order information and a decoder is employed to decode and generate the caption of the given image.

3.2.1 Order-sensitive Encoder

Since the semantic features have been serialized after the order-planning, a position encoding module is first used to model the relative or absolute order information of the sequence of features. Inspired by the Transformer we add the position embedding to each semantic feature which is calculated as follows:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), \tag{11}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}), \tag{12}$$

where pos is the position of the feature in the sequence and i is the dimension.

Then a multi-head attention based encoder is used to further encode and represent the semantic features. The encoder is a stack of N identical blocks with the same structures, each of which consists of a multi-head self-attention layer and a position-wise feedforward layer:

$$\boldsymbol{v}^a = \mathrm{LayerNorm}(\boldsymbol{v}^p + \mathrm{Multihead}(\boldsymbol{v}^p, \boldsymbol{v}^p, \boldsymbol{v}^p)), \tag{13}$$

$$\boldsymbol{v}^e = \mathrm{LayerNorm}(\boldsymbol{v}^a + FFN(\boldsymbol{v}^e)), \tag{14}$$

where $\mathrm{Multihead}$ is calculated the same as Eq.(6), FFN is the position-wise feedforward layer including two linear transformations with a GeLU activation (Hendrycks and Gimpel, 2016) in between and $\mathrm{LayerNorm}$ represents layer normalization.

3.2.2 Decoder

The caption decoder of HS-PLAN basically follows the same spirit of the Transformer which is used to generate the target caption $\mathcal{Y}$ with the encoded semantic features $\boldsymbol{v}^e$. Inspired by the AoANet (Huang et al., 2019a) we further implement the attention-on-attention module on the Transformer decoder which can determine the relevance between the attention result and the query to improve the performance of attention module.

At each decoding step t, first a masked multi-head self-attention is used to capture the dependency from the the input of the decoder, the embeddings of the predicted output $\boldsymbol{y}_{<t}$, and obtain the hidden state $\boldsymbol{h}_t$. Then a multi-head attention layer modified by the AOA module is used to obtain the context vector, which is fed with the hidden state $\boldsymbol{h}_t$ and output of encoder $\boldsymbol{v}^e$ and calculated as follows:

$$\begin{aligned} \boldsymbol{c}_t &= \mathrm{AoA}(\mathrm{Multihead}, \boldsymbol{h}_t, \boldsymbol{v}^e, \boldsymbol{v}^e) \\ &= \sigma(\boldsymbol{W}_g(\boldsymbol{h}_t + \mathrm{Multihead}(\boldsymbol{h}_t, \boldsymbol{v}^e, \boldsymbol{v}^e) + \boldsymbol{b}_g) \odot (\boldsymbol{W}_i(\boldsymbol{h}_t + \mathrm{Multihead}(\boldsymbol{h}_t^s, \boldsymbol{v}^e, \boldsymbol{v}^e)) + \boldsymbol{b}_i), \end{aligned} \tag{15}$$

where $\boldsymbol{W}_g, \boldsymbol{W}_i \in \mathbb{R}^{d_w \times d_w}, \boldsymbol{b}_g, \boldsymbol{b}_i \in \mathbb{R}^{d_w}$ are parameters and $\mathrm{Multihead}$ is calculated the same as Eq.(6). With the context vector the conditional probabilities of the output word y_t is calculated:

$$P(y_t|y_{<t}, \mathcal{V}, \mathcal{Z}) = \mathrm{Softmax}(\boldsymbol{W}_d \boldsymbol{c}_t), \tag{16}$$

where $\boldsymbol{W}_d \in \mathbb{R}^{d_w \times D}$ are parameters and D is the vocabulary size.

3.3 Objective

3.3.1 Pretraining

We pretrain the pointer network on MS-COCO and with the oracle plan illustrated in Section 3.1.2 by minimizing the negative log-likelihood of the oracle order-plan:

$$\mathcal{L}_{op} = -\sum_{(\mathcal{V},\mathcal{Z})\in\mathcal{D}} \sum_{t}^{|\mathcal{Z}|} \log P(z_t = \mathcal{V}_i | z_{<t}, \mathcal{V}), \tag{17}$$

where $\mathcal{D}$ represents all the training samples including the input features $\mathcal{V}$, the oracle plans $\mathcal{Z}$ and target captions $\mathcal{Y}$. Then we pretrain the caption model with the oracle-plan by optimizing the cross entropy (XE) loss:

$$\mathcal{L}_{cm} = -\sum_{(\mathcal{V},\mathcal{Z},\mathcal{Y})\in\mathcal{D}} \sum_{t}^{T} \log P(y_t | y_{1:t}, \mathcal{V}, \mathcal{Z}), \tag{18}$$

where T is the length of the ground truth caption.

3.3.2 Training

After the pretraining, we train our model end-to-end with a joint learning of both planning and captioning by aggregating the losses over the two stages:

$$\mathcal{L}_{XE} = \lambda \mathcal{L}_{op} + (1 - \lambda) \mathcal{L}_{cm}, \tag{19}$$

where λ is the hyperparameter. Then we follow the previous works that directly optimize the non-differentiable metrics with Self-Critical Sequence Training(Rennie et al., 2017):

$$\mathcal{L}_{RL} = -\mathbf{E}_{y_{1:T} \sim p_\theta}[r(y_{1:T})], \tag{20}$$

where r is the CIDEr (Vedantam et al., 2015) score function.

4 Experiment

4.1 Experimental Settings

4.1.1 Dataset and Metrics

We evaluate our proposed model on the popular benchmark dataset MS-COCO (Lin et al., 2014) containing 123,287 images labeled with 5 captions for each. We use the offline “Karpathy” data split (Karpathy and Li, 2015) for the performance comparisons, where $5,000$ images are used for validation, 5,000 images for testing and the rest for training. Following the previous works we also used five standard automatic evaluation metrics: CIDEr-D(Vedantam et al., 2015), BLEU(Papineni et al., 2002), METEOR(Banerjee and Lavie, 2005), ROUGE-L(Lin, 2004) and SPICE (Anderson et al., 2016). We also implement qualitative human evaluations to further evaluate the quality of the generated captions.

4.1.2 Implementation Details

We use Faster-RCNN in conjunction with ResNet-101 similarly as (Anderson et al., 2018) to extract visual features from images, which have been pretrained on ImageNet (Deng et al., 2009). Further we use the Faster-RCNN as the object detector to detect the objects in different image regions and obtain the textual description of each object, and an attribute classifier to obtain the attributes of the objects

Model	Cross-Entropy Loss						CIDEr Score Optimization					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
	Single Model											
LSTM (2015)	-	29.6	25.2	52.6	94.0	-	-	31.9	25.5	54.3	106.3	-
SCST (2017)	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Up-Down (2018)	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 (2018)	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
ETA (Li et al., 2019)	77.3	37.1	28.1	57.2	117.9	21.4	81.5	39.3	28.8	58.9	126.6	22.7
SGAE (2019)	-	-	-	-	-	-	80.8	38.4	28.4	58.6	127.8	22.1
AAT (2019b)	-	37.0	28.1	57.3	117.2	21.2	-	38.7	28.6	58.5	128.6	22.2
AoANet (2019a)	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
HS-PLAN	**78.5**	**38.5**	**29.1**	**58.9**	**121.8**	**22.3**	**81.6**	**40.3**	**29.7**	**59.9**	**133.4**	**23.6**
	Ensemble/Fusion											
GCN-LSTM$^{\Sigma}$ (2018)	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$^{\Sigma}$ (Li et al., 2019)	77.6	37.8	28.4	57.4	119.3	21.6	81.5	39.9	28.9	59.0	127.6	22.6
SGAE$^{\Sigma}$ (2019)	-	-	-	-	-	-	81.0	39.0	28.4	58.9	129.1	22.2
AoANet$^{\Sigma}$ (2019a)	78.7	38.1	28.5	58.2	122.7	21.7	81.6	40.2	29.3	59.4	132.0	22.8
HS-PLAN$^{\Sigma}$	**79.0**	**39.0**	**29.4**	**59.2**	**124.9**	**22.7**	**82.3**	**40.8**	**29.9**	**60.2**	**134.8**	**24.0**

Table 1: The performances of different models with automatic evaluation on COCO Karpathy test split, where B@N, M, R, C and S are short for BLEU@N, METEOR,ROUGE-L, CIDEr and SPICE scores. All values are reported as percentage (%).

which are utilized to reconstruct the features. The dimension of the original vectors is $d_v = 2048$ and we project them to a new space with the dimension of $d_w = 1024$, which is also the embedding size and the hidden size of the pointer network and the caption model. The pointer contains a 2-layer Transformer encoder and a 2-layer Transformer decoder, the number of heads is $h = 8$.

During the pretraining stage, we train the pointer network for 20 epochs and the caption model with oracle plan for 20 epochs. During the training stage we train our whole model jointly with $\mathcal{L}$ for 20 epochs with the mini-batch size of 10 and $\lambda = 0.3$. Learning rate is $2e-4$ with an Adam optimizer (Kingma and Ba, 2015). Then we optimize the CIDEr-D score with SCST for another 15 epochs.

4.2 Baselines

We compare our model with some following strong baselines: **LSTM** (Vinyals et al., 2015) which use CNN to encode the image and use LSTM-based decoder to generate the caption; **SCST** (Rennie et al., 2017) which first use SCST to directly optimize the evaluation metrics; **Up-Down** (Anderson et al., 2018) which propose the Bottom-Up and Top-Down attention mechanism to identify selective spatial regions; **GCN-LSTM** (Yao et al., 2018) which encodes the relationships between the objects in the image into feature vectors; **ETA** (Li et al., 2019) which propose the Entangled Attention to explore visual and semantic information simultaneously. **SGAE** (Yang et al., 2019), which introduces auto-encoding scene graphs into caption model; **AAT** (Huang et al., 2019b) which proposes an Adaptive Attention Time to align the source and the target adaptively; **AoANet** (Huang et al., 2019a) which proposes an Attention on Attention module to improve the multi-head attention based caption model.

4.3 Overall Results

The performances of the baselines and our proposed model on the COCO Karpathy test split are shown in Table 1. For fair comparison, we report the results of each model optimized with both cross entropy loss and CIDEr Score and separately show the performances for single models and ensemble/fused models. We can see that our proposed model outperforms all the baselines on all the automatic evaluation metrics with both XE loss training and CIDEr-D Score Optimization, achieving the state-of-the-art performance. Specifically, on CIDEr-D score our model achieves 121.8% with XE loss training and 133.4% with

Models	Fluency			Faithfullness			Coherence		
	win (%)	lose (%)	tie (%)	win (%)	lose (%)	tie (%)	win (%)	lose (%)	tie (%)
v.s.Transformer	50.5	15.0	34.5	65.0	16.5	18.5	68.5	14.0	17.5
v.s. SGAE	45.5	19.5	35.0	56.5	18.0	25.5	59.0	19.5	21.5
v.s. AoANet	43.0	23.5	33.5	50.5	20.5	29.0	56.5	21.0	22.5

Table 2: Results of human evaluation.

ID	Setting	B@1	B@4	M	R	C	S
1	base	76.4	36.7	27.5	56.0	117.2	20.8
2	+Position Encoding	76.2	36.2	27.4	56.3	116.5	20.7
3	+Reconstruction	77.7	37.0	28.2	57.3	118.0	21.5
4	+Order Plan, Position Encoding	77.2	37.7	28.6	57.8	118.7	21.9
5	+Reconstruction, Position Encoding	77.1	37.1	27.8	56.7	117.6	21.2
6	+Reconstruction, Order Plan	77.6	37.4	28.3	57.5	117.9	21.5
7	**Full (HS-PLAN)**	**78.5**	**38.5**	**29.1**	**58.9**	**120.9**	**22.3**

Table 3: Results of different ablation settings of our model on COCO Karpathy test split. The results are reported after XE training stage.

CIDEr-D optimization, which makes a significant improvement over the previous best model AoANet by 3.6%. With an ensemble of four models, HS-PLAN further achieves 134.8% CIDEr-D score. The results demonstrate that the proposed high-level semantic planning is able to facilitate the performance of image captioning model.

4.4 Human Evaluation

To further evaluate the quality of the captions generated by our model, we implement qualitative human evaluation on three different aspects: **Fluency** which measures whether the caption is fluent and has no grammatically error; **Faithfulness** which measures whether the caption is faithful to the given image and contains enough objects (too much or too little would be deducted); **Coherence** which measures whether the generated caption is logically coherent and is described in a proper order. For pair-wise comparison we randomly select 100 images with captions generated by our model and three strong baselines. We invite ten annotators with enough knowledge to give preference (win, lose or tie) to each pair of texts (ours vs. a baseline, 600 pairs in total).

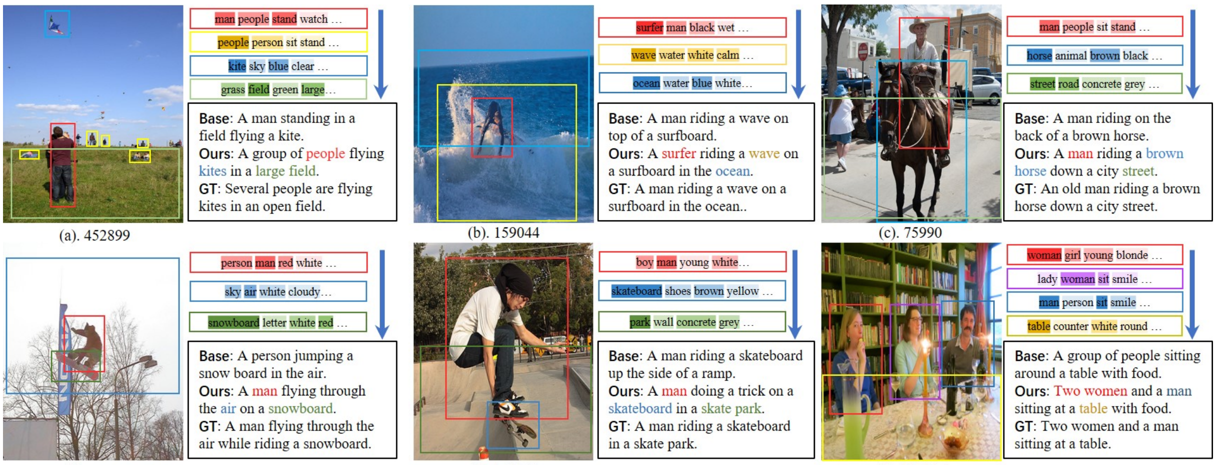
The results reported in Table 2 show that our model HS-PLAN outperforms the baselines on the three metrics, which further demonstrate the effectiveness of the proposed high-level semantic planning method. We also find that our model has a significant improvement on Faithfulness and Coherence compared with the baselines, illustrating that the high-level semantic plan is able to improve the quality of generated captions.The results demonstrate that our proposed model can generate more fluent, faithful and coherent captions.

4.5 Ablation Study

To further evaluate the effectiveness of the proposed high-level semantic planning method, we conduct ablation study by comparing the performance of different settings of HS-PLAN. The results are reported in Table 3. We can find that:

(1) The comparisons between the models with or without semantic reconstruction demonstrate that the semantic reconstruction module can heavily improve the performance of caption model and prove that semantic reconstruction can narrow the gap between the visual features and textual words.

(2) Without position encoding, order planning does not lead to obvious improvements, since the encoder is still order-insensitive and can not learn the order information. Without order-planning, the position encoding causes the performance to decrease. It might because position encoding introduces and propagates error from the disordered features.


man people stand watch ...
people person sit stand ...
kite sky blue clear ...
grass field green large...
surfer man black wet ...
wave water white calm ...
ocean water blue white...
man people sit stand ...
horse animal brown black ...
street road concrete grey ...
person man red white ...
sky air white cloudy...
snowboard letter white red ...
boy man young white...
skateboard shoes brown yellow ...
park wall concrete grey ...
woman girl young blonde ...
lady woman sit smile ...
man person sit smile ...
table counter white round ...

Base: A man standing in a field flying a kite.
Ours: A group of people flying kites in a large field.
GT: Several people are flying kites in an open field.

(a). 452899

Base: A man riding a wave on top of a surfboard.
Ours: A surfer riding a wave on a surfboard in the ocean.
GT: A man riding a wave on a surfboard in the ocean..

(b). 159044

Base: A man riding on the back of a brown horse.
Ours: A man riding a brown horse down a city street.
GT: An old man riding a brown horse down a city street.

(c). 75990

Base: A person jumping a snow board in the air.
Ours: A man flying through the air on a snowboard.
GT: A man flying through the air while riding a snowboard.

(d). 41007

Base: A man riding a skateboard up the side of a ramp.
Ours: A man doing a trick on a skateboard in a skate park.
GT: A man riding a skateboard in a skate park.

(e). 346011

Base: A group of people sitting around a table with food.
Ours: Two women and a man sitting at a table with food.
GT: Two women and a man sitting at a table.

(f). 560744

Figure 3: Examples of the captions generated by our model (Ours) and AoANet (Base) as well as the ground truth (GT). The plans are also visualized for each image (the background color represents the V2L attention weight and arrows represent the order plan).

(3) However, order planning plus position encoding can lead to better performances of models, proving that order planning can bridge the gap between the disordering of visual features and the ordering of textual sentence. Since the order planning is able to serialize the features, position encoding can learn correct order information to guide the generation.

(4) During the experiments we surprisingly find that the semantic reconstruction module can also improve the performance of the pointer network, thus making a better order plan to guide the generation.

4.6 Case Study

Figure 3 shows six examples of the captions generated by our model and a baseline randomly selected from the Karpathy test split. We also visualize the plans to further show the effectiveness of the proposed planning method. We show the high-level concepts of each image region and use background colors to represent the V2L attention weights, darker is higher. The arrows illustrate the order of the features after the order planning. We find that the baseline still suffers from the problems of information missing like (b) and misunderstanding the objects in images such as (a) and (e). But our model can better understand the objects with the benefit of the semantic reconstruction, such as ”surfer” in (b) and ”two women” in (f). Further, the captions generated by our model basically follow the order plans, demonstrating that explicit order planning can guide the neural model to generate more informative and well-ordered captions. Generally, the captions generated by our model are more informative, faithful and coherent.

5 Conclusion

In this paper we integrate the planning strategy to attention based neural models and propose a novel high-level semantic planning method to bridge the gap between the visual features and textual semantics. We design a high-level semantic planning based attention network (HS-PLAN) that incorporates both a semantic reconstruction and an explicit order planning to guide the caption generation. Experiments are conducted on a large benchmark dataset MSCOCO and show that our model outperforms the baselines on both automatic and human evaluation. The experimental results also demonstrate that our model can generate more fluent, faithful and coherent captions.

Acknowledgements

This work was supported by NSFC project Grant No.U1833101, SZSTI under Grant No.JCYJ20190809172201639 and the Joint Research Center of Tencent and Tsinghua.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086. IEEE Computer Society.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *IEEvaluation@ACL*, pages 65–72. Association for Computational Linguistics.

Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2018. Less is more: Picking informative frames for video captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 367–384. Springer.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, pages 8307–8316. Computer Vision Foundation / IEEE.

Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pages 248–255, 06.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*, pages 1473–1482. IEEE Computer Society.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017a. Stylenet: Generating attractive visual captions with styles. In *CVPR*, pages 955–964. IEEE Computer Society.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017b. Semantic compositional networks for visual captioning. In *CVPR*, pages 1141–1150. IEEE Computer Society.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). cite arxiv:1606.08415Comment: Trimmed version of 2016 draft.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019a. Attention on attention for image captioning. In *ICCV*, pages 4633–4642. IEEE.

Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. 2019b. Adaptively aligned image captioning via adaptive attention time. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, pages 8940–8949.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137. IEEE Computer Society.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR (Poster)*.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, October.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

C. Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*, Barcelona, Spain, July 25-26.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 3242–3250. IEEE Computer Society.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. *CoRR*, abs/2003.14080.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, pages 91–99.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164. IEEE Computer Society.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212. IEEE Computer Society.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.

Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Review networks for caption generation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *NIPS*, pages 2361–2369.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694. Computer Vision Foundation / IEEE.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 711–727. Springer.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*, pages 4651–4659. IEEE Computer Society.