# Hierarchical Attention Network for Image Captioning

**Weixuan Wang, Zhihong Chen, Haifeng Hu**[*]
School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China
{wangwx25, chenzhh45}@mail2.sysu.edu.cn, huhaif@mail.sysu.edu.cn

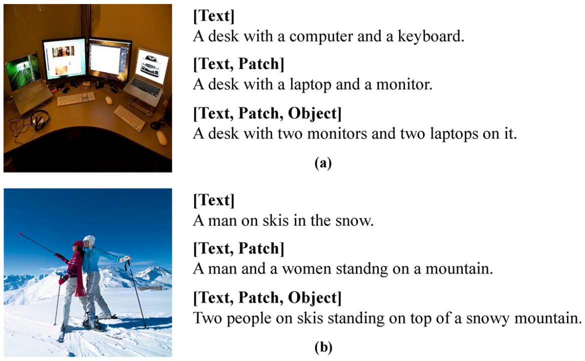
## Abstract

Recently, attention mechanism has been successfully applied in image captioning, but the existing attention methods are only established on low-level spatial features or high-level text features, which limits richness of captions. In this paper, we propose a Hierarchical Attention Network (HAN) that enables attention to be calculated on pyramidal hierarchy of features synchronously. The pyramidal hierarchy consists of features on diverse semantic levels, which allows predicting different words according to different features. On the other hand, due to the different modalities of features, a Multivariate Residual Module (MRM) is proposed to learn the joint representations from features. The MRM is able to model projections and extract relevant relations among different features. Furthermore, we introduce a context gate to balance the contribution of different features. Compared with the existing methods, our approach applies hierarchical features and exploits several multimodal integration strategies, which can significantly improve the performance. The HAN is verified on benchmark MSCOCO dataset, and the experimental results indicate that our model outperforms the state-of-the-art methods, achieving a BLEU1 score of 80.9 and a CIDEr score of 121.7 in the Karpathy's test split.

## Introduction

Image captioning, which aims to describe the content of an image, has emerged as a prominent attractive research problem in both academia and industry. It is a multidisciplinary task involving computer vision and natural language processing. Image captioning is difficult for machine since it not only requires a comprehensive understanding of objects, scene and their mutual relations, but also needs to describe the content of an image with semantically and syntactically correct sentence. In real life, image captioning has a wide range of applications, ranging from helping visually impaired people to personal assistants.

Recently, attention mechanism, which plays an important role in image captioning, has been developed into diverse forms. For example, (Xu et al. 2015; Lu et al. 2017; Wang et al. 2017) produce a spatial mask via calculating the similarity between image patches and generated words at each time step. (Anderson et al. 2018) generates bounding boxes with an object detector and highlights regions associated with the generated words. (You et al. 2016; Wu et al. 2016; Gan et al. 2017) apply the soft attention to selectively focus on visual concepts. These attention mechanisms are used to distill more discriminative visual information to improve generated sentences. In the aforementioned methods, the attention mechanisms are based on single level features, such as semantically weak patch features, semantically moderate object features or semantically strong text features. However, different words in description sentence are relevant to features of different levels. Figure 1 illustrates captions generated by leveraging visual features of different levels. In Figure 1(a), the model applying text features attends to describe all objects in the image without focusing on salient objects. With both text features and patch features, the model prefers to describe objects with higher salience, but the quantity of described objects is inaccurate. When leveraging text features, object features and patch features simultaneously, the model counts objects in images accurately, which indicates different words are relevant to fea-



**[Text]**
A desk with a computer and a keyboard.
**[Text, Patch]**
A desk with a laptop and a monitor.
**[Text, Patch, Object]**
A desk with two monitors and two laptops on it.
**(a)**

**[Text]**
A man on skis in the snow.
**[Text, Patch]**
A man and a women standng on a mountain.
**[Text, Patch, Object]**
Two people on skis standing on top of a snowy mountain.
**(b)**

Figure 1: The illustration of captions generated by features of different levels. Text, Patch and Object indicate the semantically strong text features, semantically weak patch features and semantically moderate features respectively.

[*] Corresponding author

tures of different levels, and it is of great significance to incorporate features of different levels to predict sentences. In this paper, we propose a novel features hierarchy, which consists of features on diverse semantic levels. On top of this hierarchical structure, we construct several independent parallel attention modules to refine the features of different levels, which allow us to leverage different features as the dominant role to predict words at different time steps.

Since features from different attention networks are in different modalities, it is necessary to explore a multimodal embedding strategy to combine them. However, most existing image captioning models (You et al. 2016; Rennie et al. 2017), directly integrate features of different modalities via addition or concatenation, which is unable to fully capture the complex correlations between features of different modalities. Motivated by the researches (Kim et al. 2016; 2017; Ben-younes et al. 2017) in Visual Question Answer (VQA), in this paper, we propose a novel Multivariate Residual Module (MRM) to model the richer multimodal representation. MRM is able to project features of each modality into the target space and exploit the relevant relations among source spaces. Compared with the traditional methods, the MRM preserves valuable information of each modality and extracts more discriminative multimodal features. On the other hand, inspired by the neuroscience researches (Ungerleider and Haxby 1994; Lu et al. 2018b), we further design a parallel MRM (pMRM) for integrating features of different levels gradually.

Since the importance of the features generated by MRM varies from word to word at the different time step, we introduce a context gate structure to adaptively balance the contribution of features of different levels.

Overall, the contributions of this paper are four fold:

- In order to generate accurate and informative sentences, we construct a feature pyramid leveraging patch features, object features and text features, and build an attention network to refine features.
- We propose a parallel Multivariate Residual Network to integrate features of different levels, which aims at projecting features into a unified target space and exploring the intrinsic relationship between different source spaces.
- A context gate mechanism is introduced into our model to adaptively balance the contribution of features on different levels.
- We conduct a number of experiments on the MSCOCO dataset and the results show that our model outperforms the state-of-the-art approaches, achieving a BLEU1 score of 80.9 and a CIDEr score of 121.7 in the Karpathy's test split.

## Related Work

**Attention mechanism** Recently, plenty of researchers are devoted to studying diverse attention mechanisms in order to refine visual information for image captioning. (Xu et al. 2015) presents a spatial attention mechanism to focus on the fixed-size patches which are most relevant to the generated words. (You et al. 2016) applies an attribute predictor to propose semantic concepts which are selectively attended and combined with the semantic information of recurrent neural network to predict words. (Anderson et al. 2018) proposes a bottom-up and top-down attention mechanism that computes the relevance between the generated words and salient objects. Although these attention mechanisms are effective, they have the deficiency of predicting words by using only one kind of features. In a sentence, different words could be determined by the features of different levels. For example, the color words could be predicted by low-level features and the quantifier could be generated by mid-level features. Predicting on one-level features may generate words that are inconsistent with the content of images.

**Multimodal embedding** In the previous works (You et al. 2016; Rennie et al. 2017) in image captioning, features of different modalities are usually integrated by the concatenation or addition operators. However, in the Visual Question Answering (VQA), Kim (Kim et al. 2016) proposes Multimodal Residual Network (MRN) for multimodal residual learning, which adopts Hadamard multiplication for the joint residual mappings. Kim (Kim et al. 2017) further presents a Multimodal Low-rank Bilinear Attention Networks (MLB) so as to approximate bilinear pooling to learn the multimodal features. It is beneficial to introduce these achievements into the image captioning task.

## Methodology

### Overall

Given an image $I$, the image captioning model needs to generate a caption sequence $w = \{w_1, w_2, \ldots, w_T\}$, $w_t \in D$, where $D$ is the vocabulary dictionary and $T$ is the sequence length. We adopt the variant CNN-RNN architecture for image captioning. In particular, CNN, which plays the role of an visual encoder, extracts four different features including global features $V_g$, patch features $V_p$, object features $V_o$ and text features $V_t$ to establish hierarchical features. As a semantic decoder, RNN is leveraged to guide the generation of attention and caption sequences. In order to alleviate the vanishing gradient problem, we adopt LSTM (Hochreiter and Schmidhuber 1997) as a decoder in this paper.

To decouple attention guidance and sequence generation, we construct a cascaded LSTM structure that includes a visual LSTM and a language LSTM. The former is applied to perceive global information of images and guide different attention mechanisms to generate attention features $A_p, A_o, A_t$, while the latter guides caption generation. In order to better integrate attention features of different levels, we construct MRM to extract the internal relationship between features.

The overall structure of our model is shown in Figure 2. During the generation process, the visual encoder extracts different features. The visual LSTM reviews the global information of the image at each moment and guides attention models to refine features. The attention features of different levels are input into pMRM to project into a unified target space for integration. The language LSTM generates a word at each moment given last word and multimodal features.

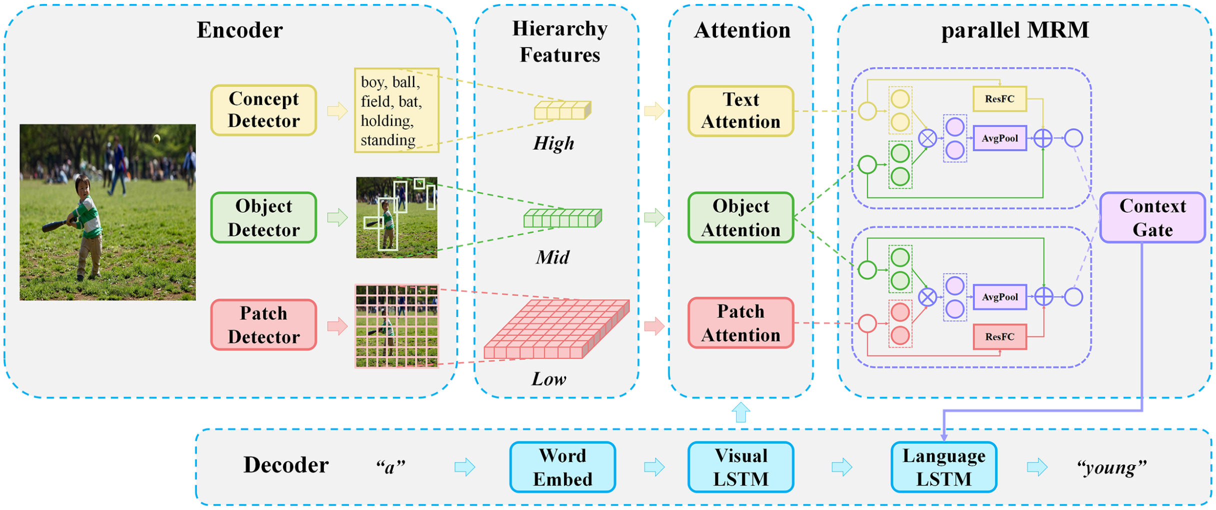


Figure 2: The illustration of the overall structure of HAN, which is composed of Encoder module, Attention module, parallel MRM and Decoder module. The model takes the images and the words generated at last time step as input and outputs the next words.

The process can be defined by the following formulas:

$$V_g, V_p, V_o, V_t = Detectors(I) \tag{1}$$

$$h_t^V = LSTM_V([h_{t-1}^L, V_g, E(w_t)]) \tag{2}$$

$$A_p, A_o, A_t = Attentions(h_t^V, V_p, V_o, V_t) \tag{3}$$

$$M = pMRM(A_p, A_o, A_t) \tag{4}$$

$$h_t^L = LSTM_L(M) \tag{5}$$

$$w_t \sim Softmax(h_t^L) \tag{6}$$

where $Detectors()$ represents feature extractors and $E()$ is the embedding function which maps the one-hot representation into the embedding space. $LSTM_V$, $Attentions()$, $pMRM()$ and $LSTM_L$ represent visual LSTM, attention module, pMRM and language LSTM, respectively.

Given an image $I$ and the corresponding caption $w = \{w_1, w_2, \ldots, w_T\}$, the model aims to maximize the following loss function:

$$\theta^* = argmax \sum_{(I,y)} \log p(y|I;\theta) \tag{7}$$

where $\theta$ is the parameters of our model. Applying chain rule, we can model the joint probability on $w = \{w_1, w_2, \ldots, w_T\}$ and the cross entropy loss function (XE) is adopted to minimize a negative log-likelihood:

$$L = -\sum_{t=1}^{T} \log p(w_t|w_1, ..., w_{t-1}, I) \tag{8}$$

## Hierarchy feature pyramid

In the previous work, most attention mechanisms are based on only one-level features. For example, (Xu et al. 2015) is based on semantically weak patch features. (Anderson et al. 2018) is established on semantically moderate object features and (You et al. 2016) leverages semantically strong text features. However, as shown in Figure 1, different words are associated with features of different levels. To leverage different features to generate words synchronously, we propose a hierarchy feature pyramid structure. The bottom, middle and top layers of the hierarchy are the patch features, object features and text features respectively.

**Patch features** Patch features refer to the abstract feature expression of each patch in an image. We utilize a Resnet-101 pre-trained on ImageNet to extract features of the last convolution layer as patch features $V_p \in R^{r \times r \times d}$. In particular, $V_p = [V_{p(1)}, V_{p(2)}, \ldots, V_{p(r \times r)}]$, $V_{p(i)} \in R^d$ is a $d$-dimensional visual patch vector and the number of patch features is $Np = r \times r = 196$.

**Object features** Object features refer to the feature representation of salient objects. In order to accurately capture objects, we introduce the Faster RCNN into our model. In our work, we take $N_o$ object features $V_o$ with the highest confidence, where $N_o = 15$. In particular, $V_o = [V_{o(1)}, V_{o(2)}, \ldots, V_{o(N_o)}]$, $V_{o(i)} \in R^d$ is a $d$-dimensional visual object vector.

**Text features** Text features refer to semantic concepts related to images, including adjectives, verbs, and nouns. In order to obtain the text features, we construct a $K = 2000$ classification text predictor, where the $K$ represents the number of the most frequent words in the dataset. We take $N_t = 10$ text concepts $T = [T_1, T_2, \ldots, T_{N_t}]$ with the highest score. In particular, when constructing feature hierarchy, the semantic concepts are converted to text features $V_t \in R^{N_t \times d}$ via the Embedding function, where $V_{t(i)} \in R^d$ is a $d$-dimensional text vector.

The text predictor, composed of the backbone of ResNet101 and three novel fully connected layers, is shown in Figure 3. In training stage, we fix the parameters of the ResNet and only optimize the fully connected layers to predict the texts that are related to the input image. The objective function we adopted is defined as follows:

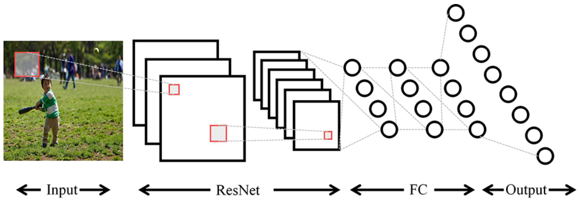


Figure 3: The illustration of text predictor based on Resnet101 with three fully connected layers.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[p_i^* \log(p_i) + (1-p_i^*)\log(1-p_i)] \tag{9}$$

where $N$ represents the number of texts and the $p_i^*$ is 1 if the corresponding word exists in the groundtruth.

## Attention module

In order to focus on the features that are most relevant to words at the current time step, the soft attention mechanism (Xu et al. 2015) is introduced into our framework. We construct three independent attention networks on three levels for refining features. Given the features $V$ of one level and the output $h_t^V$ of the visual LSTM, we apply a neural network to normalize attention weights:

$$z(t) = W_\alpha^T \tanh(W_V V + W_h h_t^V) \tag{10}$$

$$\alpha(t) = softmax(z(t)) \tag{11}$$

where $W_V, W_h \in R^{d\times d}$ and $W_\alpha \in R^{d\times 1}$ are the trainable matrices and $\alpha(t) \in R^N$ are the attention weights. Based on the weight matrix, attention features $A(t)$ can be calculated by weighted sum at the current time step $t$:

$$A(t) = \sum_{i=1}^{N} \alpha_i(t) V_i \tag{12}$$

Specially, non-visual words in captions are not relevant to object and text features. Thus, we concatenate object features with global features, text features with semantic features of the Language LSTM at the last time step in order to provide the extra global information to attend.

## Multivariate Residual Module

The existing image captioning methods do not employ the strategy of combining various features, but directly fuse features with concatenation or addition operators. Motivated by the researches (Kim et al. 2016)(Kim et al. 2017), we propose a novel Multivariate Residual Module (MRM) to integrate information of different modalities. The MRM consists of a projection part and a relation part.

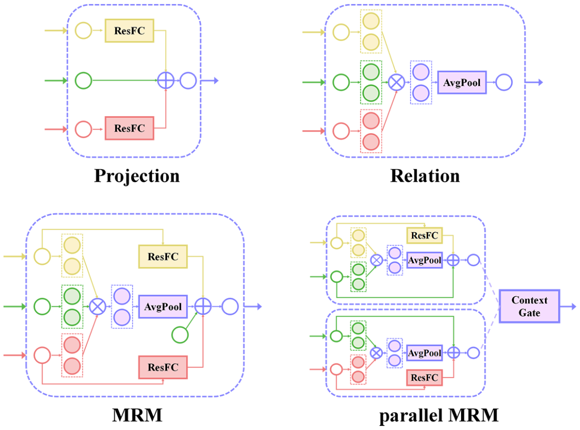


Figure 4: The illustration of Multivariate Residual Module and parallel Multivariate Residual Module.

**Projection** Inspired by ResNet (He et al. 2016), the projection part (show in Figure 4(a)) is proposed to learn the relationship between the input data and nonlinear residual function, rather than directly learn the desired mapping. Thus, we construct two independent residual networks to project the patch attention feature and the text attention feature into the object space:

$$H_p = A_p + ReLU(W_{mp}A_p) \tag{13}$$

$$H_t = A_t + ReLU(W_{mt}A_t) \tag{14}$$

where $A_p$ and $A_t$ are the patch attention feature and text attention feature respectively. $ReLU$ is a nonlinear activation function and the overall projected feature $H$ defined as

$$H = A_o + H_p + H_t \tag{15}$$

where $A_o$ is the object attention feature.

**Relation** Motivated by (Kim et al. 2016)(Kim et al. 2017), the relation part is proposed to apply the multimodal bilinear strategy to exploit the inherent relationship among diverse spaces. The relation network is shown in Figure 4(b). Given the patch attention feature $A_p \in R^m$ and object feature $A_o \in R^n$, multimodal bilinear strategy is defined as follows:

$$Z_i = A_p^T W_i A_o \tag{16}$$

where $W_i \in R^{m\times n}$ is the weight matrix. In order to obtain the output $Z \in R^o$, we need to learn $o$ matrices $W = [W_1, ..., W_o] \in R^{m\times n\times o}$. According to (Kim et al. 2017), we can rewrite the above formula to decrease the dimensions of parameter matrices:

$$Z_i = A_p^T W_i A_o = A_p^T U_i V_i^T A_o = U_i^T A_p \circ V_i^T A_o \tag{17}$$

$$Z = U^T A_p \circ V^T A_o \tag{18}$$

where $U = [U_1, ..., U_o] \in R^{m\times o}$, $V = [V_1, ..., V_o] \in R^{n\times o}$, and $\circ$ represents the Hadamard product. Further, we can extend this strategy to merge the features of three modalities and rewrite it as follows:

$$Z = U^T A_p \circ V^T A_o \circ W^T A_t \tag{19}$$

where $A_t$ represents text attention feature. $U, V, W$ are weight matrices. Finally, we apply an average pooling layer to condense relation features:

$$R = AvgPool(Z) \tag{20}$$

**MRM** The MRM is shown in Figure 4(c). The output of the multivariate residual module is determined by projection features $H$ (refers to the equation (15)) and relation features $R$ (refers to the equation (20)), which is defined as:

$$M = H + R \tag{21}$$

**pMRM (parallel MRM)** The neuroscience (Ungerleider and Haxby 1994) proves that there is a content pathway in the brain for responding selectively to relevant object identification, and a position pathway for responding selectively to spatial aspects of stimuli. Furthermore, the latest research (Lu et al. 2018b) indicates that there is a cluster of neurons in the brain that provides fine-grained object information to the pathways when identifying objects. In our work, similar to these research, we propose a parallel MRM to integrate features of different levels. The pMRM first provide the object information to the text features and the patch features respectively. Then a context gate is introduced to selectively focus on high-level content features and low-level position features. The pMRM is shown in Figure 4(d).

## Context Gating

Inspired by the gating mechanism in LSTM (Hochreiter and Schmidhuber 1997) and the work (Wang et al. 2018) in dense video captioning, we introduce a context gating mechanism into our model to balance the contribution of low-level context and high-level context. When obtaining the low-level attended features $M_L$ and the high-level attended features $M_H$, we learn a context gate to balance them. We project the two different features into the same space:

$$\tilde{M_L} = tanh(W_L M_L) \tag{22}$$

$$\tilde{M_H} = tanh(W_H M_H) \tag{23}$$

where $W_L, W_H$ are the projection matrices. The context gate is then calculated by a nonlinear sigmoid function:

$$g_{ctx} = \sigma(W_g[\tilde{M_L}, \tilde{M_H}, h_t^V]) \tag{24}$$

where $h_t^V$ is the previous visual LSTM state and $g_{ctx}$ is a 512-d weight vector. We could fuse the low-level features and the high-level features as follows:

$$M = [(1 - g_{ctx}) \circ M_L, g_{ctx} \circ M_H] \tag{25}$$

## Objective

Firstly, we adopt the usual cross entropy loss (XE) to optimize our model. Considering the XE may result in the discrepancy of evaluation between training and inference, we further adopt the CIDEr (Rennie et al. 2017) as the objective function to finetune our model. Specially, we minimize the negative expectation score of CIDEr as follows:

$$L(\theta) = -E_{w^s \sim p_\theta}[CIDEr(w^s)] \tag{26}$$

According to (Rennie et al. 2017), the expected gradient for single sample $w^s \sim p_\theta$ is:

$$\triangledown_\theta L(\theta) \approx -(CIDEr(w^s) - CIDEr(w)) \triangledown_\theta \log p_\theta(w^s) \tag{27}$$

where $w^s = (w_1^s, ..., w_T^s)$, $w_t^s$ is the word sampled at time $t$ and $CIDEr(w)$ is the baseline score obtained by greedily decoding the current model.

# Experiment

## Datasets and Evaluation metrics

The MSCOCO dataset (Lin et al. 2014) is the benchmark dataset for image captioning, which contains 82,783, 40,504, and 40775 images for training, validation and test respectively. For offline evalution, we employ the Karpathy's splits (Karpathy and Li 2015) which contain 113,287 images for training, 5,000 images for validation and 5,000 images for test. For fair comparison, we also report the results of the online MSCOCO testing server. To evaluate the quality of captioning, we adopt the evaluation criteria widely applied in previous works: BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011), ROUGEL (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016).

## Implement details

We use the ResNet101 pretrained on ImageNet to extract global feature and patch features, and use the Faster RCNN pretrained on MSCOCO object detection dataset to obtain object features, and train another ResNet101 to predict visual concepts. The dimension of these features are reduced to 512. Due to the limitation of the hardware, we only adopt 196 patch features, 15 object features and 10 text features for each image. The dimensions of embedding layers and both LSTMs are set to 512. Firstly, we train our model under cross entropy (XE) loss using ADAM optimizer with a learning rate 5e-4 and do not finetune the CNN. Afterwards, we perform the CIDEr standard optimization on the XE-trained model, and also use the Adam optimizer. In the decoding process, we use beam search and set beam size to 3.

## Comparison with the state-of-the-art methods

For offline evaluation on MSCOCO dataset, we compare our model with the current state-of-the-art methods: Adaptive (Lu et al. 2017), Att2in (Rennie et al. 2017), Updown (Anderson et al. 2018) and NeuralBabyTalk (Lu et al. 2018a) on MSCOCO. The Adaptive method proposes an adaptive attention to decide where and when to attend. Att2in modifies the architecture of the traditional spatial attention mechanism and optimizes the model with reinforcement learning. Updown proposes a bottom-up mechanism to detect image regions, and a top-down mechanism to determine feature weightings. NBT combines the patch attention with object attention, and generates sentences template with slots, which are then filled in by visual concepts. Table 1 shows the results on the Karpathy's test split. From the table we can find that our model achieves the best performances in all metrics. With XE objective, our model has an superiority over NBT model with an improvement of 7.1% in terms of the CIDEr metric. With CIDEr objective, our model improves 9.2% relative to att2in model. Considering the fact that Updown applies the object detector trained on Visual Genome dataset (Krishna et al. 2017) and thus obtains better object features, we do not compare its performance with other models.

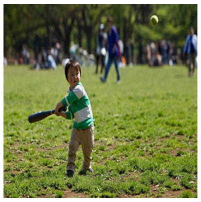
We further report the results on the official MSCOCO evaluation server in Table 2. The models in first row are optimized with XE, and the one in second row are finetuned

Table 1: The performance of HAN on the MSCOCO Karpathy's test split. The XE is the cross entropy objective and the RL is the reinforcement learning objective. * uses better object features, and are thus not directly comparable.

| Dataset | Objectvie | Models | B1 | B2 | B3 | B4 | M | R | C | S |
|---|---|---|---|---|---|---|---|---|---|---|
| MSCOCO | XE | Att2in(Rennie et al. 2017) | - | - | - | 31.3 | 26.0 | 54.3 | 101.3 | - |
| | | Adaptive(Lu et al. 2017) | 74.2 | 58.0 | 43.9 | 33.2 | 26.6 | - | 108.5 | 19.5 |
| | | NBT(Lu et al. 2018a) | 75.5 | - | - | 34.7 | 27.1 | - | 107.2 | 20.1 |
| | | Updown*(Anderson et al. 2018) | 77.2 | - | - | 36.2 | 27.0 | 56.4 | 113.5 | 20.3 |
| | | **ours** | **77.2** | **61.2** | **47.7** | **36.2** | **27.5** | **56.6** | **114.8** | **20.6** |
| | RL | Att2in(Rennie et al. 2017) | - | - | - | 33.3 | 26.3 | 55.3 | 111.4 | - |
| | | Updown*(Anderson et al. 2018) | 79.8 | - | - | 36.3 | 27.7 | 56.9 | 120.1 | 21.4 |
| | | **ours** | **80.9** | **64.6** | **49.8** | **37.6** | **27.8** | **58.1** | **121.7** | **21.5** |

Table 2: Comparison with the state-of-the-art methods on the online MSCOCO test server. † indicates the results of ensemble models.

| Models | B1 | | B2 | | B4 | | M | | R | | C | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | c5 | 40 | c5 | 40 | c5 | 40 | c5 | 40 | c5 | 40 | c5 | 40 |
| SCA(Chen et al. 2017) | 72.5 | 89.2 | 55.6 | 80.3 | 30.6 | 58.2 | 24.6 | 32.9 | 52.8 | 67.2 | 91.1 | 92.4 |
| NIC(Vinyals et al. 2015) | 71.3 | 89.5 | 54.2 | 80.2 | 30.9 | 58.7 | 25.4 | 34.6 | 53.0 | 68.2 | 94.3 | 94.6 |
| ATT(You et al. 2016) | 73.1 | 90.0 | 56.5 | 81.5 | 31.6 | 59.9 | 25.0 | 33.5 | 53.5 | 68.2 | 94.3 | 95.8 |
| Adaptive(Lu et al. 2017) | 74.8 | 92.0 | 58.4 | 84.5 | 33.6 | 63.7 | 26.4 | 35.9 | 55.0 | 70.5 | 104.2 | 105.9 |
| MIXER(Ranzato et al. 2016) | 74.7 | 90.9 | 57.9 | 82.7 | 31.7 | 60.0 | 25.8 | 34.0 | 54.5 | 68.6 | 99.1 | 101.2 |
| SPIDEr(Liu et al. 2017) | 75.4 | 91.8 | 59.1 | 84.1 | 33.2 | 62.4 | 25.7 | 34.0 | 55.0 | 69.5 | 101.3 | 103.2 |
| AC(Zhang et al. 2017) | 77.8 | 92.9 | 61.2 | 85.5 | 33.7 | 62.5 | 26.4 | 34.4 | 55.4 | 69.1 | 110.2 | 112.1 |
| SCST†(Rennie et al. 2017) | 78.1 | 93.1 | 61.9 | 86.0 | 35.2 | 64.5 | 27.0 | 35.5 | 56.3 | 70.7 | 114.7 | 116.7 |
| **HAN** | **80.4** | **94.5** | **63.8** | **87.7** | **36.5** | **66.8** | **27.4** | **36.1** | **57.3** | **71.9** | **115.2** | **118.2** |



**Faster RCNN**
person, ball, bat, backpack, frisbee, umbrella, handbag, car, bench, racket

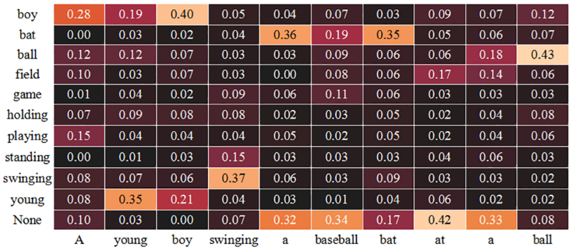
**Our Text Predictor**
boy, bat, ball, field, game, holding, playing, standing, swinging, young

Figure 5: The classes predicted by Faster RCNN and the texts predicted by our proposed text predictor.

with CIDEr. We can find that HAN outperforms other models, and achieves an impressive result.

## Ablation study

To better understand the effect of our hierarchical features strategy, we conduct experiments on a unified framework to compare the model leveraging hierarchical features by concatenation (Hierarchy) with three models leveraging patch features (Patch), object features (Object) and text features (Text) respectively. In particular, the framework mainly consists of Visual LSTM, Attention module and Language LSTM. The MRM is removed in order to provide the same experimental conditions. The results are shown in Table 3. The Hierarchy model achieves an improvement of 5.2%, 3.8% and 18.6% in terms of the CIDEr metric compared with the Patch model, Object model and Text model. Better performance can be attributed to the fact that Hierarchy model can provide more extensive features for Attention module.



| | A | young | boy | swinging | a | baseball | bat | at | a | ball |
|---|---|---|---|---|---|---|---|---|---|---|
| boy | 0.28 | 0.19 | 0.40 | 0.05 | 0.04 | 0.07 | 0.03 | 0.09 | 0.07 | 0.12 |
| bat | 0.00 | 0.03 | 0.02 | 0.04 | 0.36 | 0.19 | 0.35 | 0.05 | 0.06 | 0.07 |
| ball | 0.12 | 0.12 | 0.07 | 0.03 | 0.03 | 0.09 | 0.06 | 0.06 | 0.18 | 0.43 |
| field | 0.10 | 0.03 | 0.07 | 0.03 | 0.00 | 0.08 | 0.06 | 0.17 | 0.14 | 0.06 |
| game | 0.01 | 0.04 | 0.02 | 0.09 | 0.06 | 0.11 | 0.06 | 0.03 | 0.03 | 0.03 |
| holding | 0.07 | 0.09 | 0.08 | 0.08 | 0.02 | 0.03 | 0.05 | 0.02 | 0.04 | 0.08 |
| playing | 0.15 | 0.04 | 0.04 | 0.04 | 0.05 | 0.02 | 0.05 | 0.02 | 0.04 | 0.06 |
| standing | 0.00 | 0.01 | 0.03 | 0.15 | 0.03 | 0.03 | 0.03 | 0.04 | 0.06 | 0.03 |
| swinging | 0.08 | 0.07 | 0.06 | 0.37 | 0.06 | 0.03 | 0.09 | 0.03 | 0.03 | 0.02 |
| young | 0.08 | 0.35 | 0.21 | 0.04 | 0.03 | 0.01 | 0.04 | 0.06 | 0.02 | 0.02 |
| None | 0.10 | 0.03 | 0.00 | 0.07 | 0.32 | 0.34 | 0.17 | 0.42 | 0.33 | 0.08 |

Figure 6: The visualization of text attention. The numbers represent the attention weights between our generated caption(bottom) and our generated concepts(left).

To illustrate the effect of our multivariate residual embedding strategy, we further carry out another ablation study and the results are exhibited in Table 4. Compared with the previous methods that apply concatenation or addition to integrate features, our MRM achieves an improvement of 3.2% and 3.1% in terms of the CIDEr metric. This indicates that our MRM is able to learn the valuable representations from features of different levels. Moveover, the pMRM network promotes the CIDEr score from 114.2 to 114.8, because it integrates different features and introduces the gate mechanism to adaptively adjust the contribution of features on different levels.

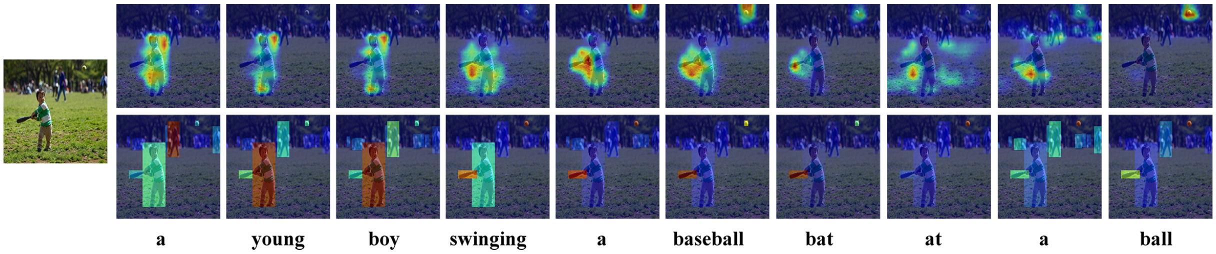


Figure 7: The visualization of patch attention(top) and object attention(bottom). The red and blue color represent the highest and lowest score respectively.

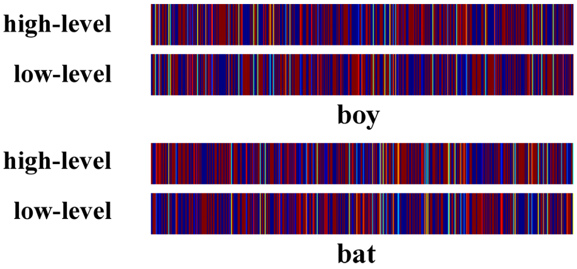


Figure 8: The visualization of the context gate. The red and blue colors represent the highest and lowest score respectively.

## Qualitative Analysis

In order to qualitatively analyze our model, we first exhibit the categories proposed by Faster RCNN trained on MSCOCO and the texts generated by our text predictor in Figure 5. The results reveal that the texts proposed by our predictor are more related to image content. Secondly, we visualize the attention masks of features of different levels for each word in the generated caption. The visualization of text attention is shown in Figure 6, and the visualization of patch and object attention are shown in Figure 7. In Figure 6, when inputting the word "baseball", the text attention assigns the highest confidence to the concrete concepts for guiding the next word "bat". In Figure 7, when predicting the word "bat", our patch attention focuses on the area around the baseball bat, and the object attention can accurately attend to the region of "bat". Figure 8 exhibits the visualizations of the context gate when predicting the words "boy" and "bat". Each row represents a weight vector, and the red and blue colors denote the highest and lowest score respectively. When predicting different words, the context gate can adaptively focus on the different channels of the high-level feature and the low-level feature.

## Conclusions

In this paper, we propose a Hierarchical Attention Network (HAN) for image captioning. The key of our work is adopting semantically weak patch features, semantically moderate object features and semantically strong text features to construct a pyramidal hierarchy of features, which allows predicting different words according to different features. We also propose an MRM to model projections and extract relevant relations among different features. Inspired by the latest research in neuroscience, we further construct the parallel MRM to combine features gradually. Moreover, a context gate is introduced to balance the contribution of different features. We verify our model on the benchmark datasets MSCOCO, and achieve the state-of-the-art results.

Table 3: The performance of the ablation experiment on single feature and hierarchical features.

| Model | B1 | B2 | B4 | M | R | C |
|---|---|---|---|---|---|---|
| Patch | 75.2 | 59.1 | 34.4 | 26.2 | 55.3 | 106.8 |
| Object | 75.5 | 59.3 | 34.9 | 26.4 | 55.6 | 108.3 |
| Text | 71.2 | 54.3 | 30.8 | 24.7 | 52.8 | 94.8 |
| Hierarchy | **76.3** | **60.4** | **35.5** | **27.1** | **56.4** | **112.4** |

Table 4: The performance of the ablation experiment on hierarchical features with different combination.

| Model | B1 | B2 | B4 | M | R | C |
|---|---|---|---|---|---|---|
| Add | 76.4 | 60.2 | 35.3 | 27.2 | 56.3 | 111.2 |
| Concat | 76.3 | 60.4 | 35.5 | 27.1 | 56.4 | 112.4 |
| Projection | 76.5 | 60.6 | 35.7 | 27.2 | 56.3 | 112.3 |
| Relation | 76.7 | 60.8 | 35.8 | 27.3 | 56.5 | 113.2 |
| MRM | 76.8 | 60.7 | 36.1 | 27.5 | 56.6 | 114.2 |
| pMRM | **77.2** | **61.2** | **36.2** | **27.5** | **56.6** | **114.8** |

## Acknowledgement

This work was supported in part by the NSFC (61673402), the NSF of Guangdong Province (2017A030311029), the Science and Technology Program of Guangzhou (201704020180), and the Fundamental Research Funds for the Central Universities of China.

## References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.

Ben-younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2631–2639.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 6298–6306.

Denkowski, M., and Lavie, A. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 85–91. Association for Computational Linguistics.

Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017. Semantic compositional networks for visual captioning. In *CVPR*, 1141–1150.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Karpathy, A., and Li, F. F. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

Kim, J.-H.; Lee, S.-W.; Kwak, D.; Heo, M.-O.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Multimodal residual learning for visual qa. In *NIPS*, 361–369.

Kim, J.; On, K. W.; Lim, W.; Kim, J.; Ha, J.; and Zhang, B. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 873–881.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 3242–3250.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018a. Neural baby talk. In *CVPR*, 7219–7228.

Lu, Y.; Yin, J.; Chen, Z.; Gong, H.; Liu, Y.; Qian, L.; Li, X.; Liu, R.; Andolina, I. M.; and Wang, W. 2018b. Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4. *Neuron* 98(2):417–428.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 1179–1195.

Ungerleider, L. G., and Haxby, J. V. 1994. 'what' and 'where' in the human brain. *Current Opinion in Neurobiology* 4(2):157–165.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.

Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; and Cottrell, G. W. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *CVPR*, 7378–7387.

Wang, J.; Jiang, W.; Ma, L.; Liu, W.; and Xu, Y. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 7190–7198.

Wu, Q.; Shen, C.; Liu, L.; Dick, A.; and Hengel, A. V. D. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 203–212.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*, 4651–4659.

Zhang, L.; Sung, F.; Liu, F.; Xiang, T.; Gong, S.; Yang, Y.; and Hospedales, T. M. 2017. Actor-critic sequence training for image captioning. In *CVPR*.