Channel and spatial attention mechanism for fashion image captioning

Bao T. Nguyen[1], Son T. Nguyen[1], Anh H. Vo[2]
[1]Faculty of Information Technology, HCMC University of Technology and Education, Ho Chi Minh City, Vietnam
[2]Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Sep 25, 2022
Revised Jan 13, 2023
Accepted Feb 4, 2023

Keywords:

Attention mechanism
Channel wise attention
Deep learning
Fashion image captioning
Spatial attention

ABSTRACT

Image captioning aims to automatically generate one or more description sentences for a given input image. Most of the existing captioning methods use encoder-decoder model which mainly focus on recognizing and capturing the relationship between objects appearing in the input image. However, when generating captions for fashion images, it is important to not only describe the items and their relationships, but also mention attribute features of clothes (shape, texture, style, fabric, and more). In this study, one novel model is proposed for fashion image captioning task which can capture not only the items and their relationship, but also their attribute features. Two different attention mechanisms (spatial-attention and channel-wise attention) is incorporated to the traditional encoder-decoder model, which dynamically interprets the caption sentence in multi-layer feature map in addition to the depth dimension of the feature map. We evaluate our proposed architecture on Fashion-Gen using three different metrics (CIDEr, ROUGE-L, and BLEU-1), and achieve the scores of 89.7, 50.6 and 45.6, respectively. Based on experiments, our proposed method shows significant performance improvement for the task of fashion-image captioning, and outperforms other state-of-the-art image captioning methods.

Corresponding Author:

Bao T. Nguyen
Faculty of Information Technology, HCMC University of Technology and Education
01 Vo Van Ngan, Thu Duc, Ho Chi Minh, Vietnam
Email: baont@hcmute.edu.vn

1. INTRODUCTION

Image captioning is the process of automatically producing one or more natural language sentences for a given input static image. Most existing captioning methods utilize the encoder-decoder model, which mainly focuses on recognizing and capturing the relationships between objects in the input, and it has been found effective for this task [1], [2]. Encoder-decoder models utilize a convectional neural network (CNN) to encode the image and represent by a compact hidden feature map. Then, followed by a recurrent neural network (RNN) to decode this representation and generate word-by-word sentence to form one or more descriptive sentences of the image [3].

Although image captioning has gained much encouragement, one limitation of these models is that that they often describe the entire scene rather than paying attention to local features that are relevant to the description. To solve this problem, encoder CNN and decoder RNN are used together with attention mechanisms as discussed in [4], which are able to capture different local parts of the input image when producing the output sentence. According to our human visual system, attention mechanisms are interesting facets as highlighted in [3]. Attention models offer a more dynamic approach to highlighting salient features in an image, rather than representing the whole input image as a static feature map. This is particularly valuable when dealing with images containing multiple objects. Leveraging image representations, such as the top layer output of a CNN condensing information in the image down to its most important objects, is an effective means of capturing information from various regions of the image. By integrating attention mechanisms, the model has ability of incorporating correct visual contexts, resulting in significant empirical improvements in performance.

There are many types of attention model, and most of them are the self-attention in the convolutional layer of CNN [5]. Among of them, spatial dimension attention [4] has got the majority of recent researches, which focuses on the spatial dimension of the image. The approach relies on spatial probabilities that adjust the weights of the final convolutional feature map generated by a CNN encoding an input image, and then using an RNN as a decoder to produce the sentence. The spatial attention mechanism aimed at recognizing and capturing the relationship between objects in the input image [5]. The reason is that, in image captioning task, most of previous used-datasets are the outdoor ones with images containing common objects. Thus, in order to generate the describing sentence, the model is required to identify these objects, their actions and the relationship between them. Due to that, applying spatial attention techniques is sufficient image captioning. However, generating the description for fashion images need to be considered on the different aspects of attention mechanism. Specifically, the generated sentence of a fashion image should represent not only the relationship between objects in the image, but also some other clothes details such as shape, texture, style, and fabric. From that point view, if the model only uses the spatial attention mechanism, which mainly focus on recognizing and capturing the relationships between objects, some main attributes of fashion items will be missed in the description sentence. It is clear that the spatial based captioning model does not meet requirements of fashion-image captioning task.

In this study, a novel artificial neural network (ANN) model which can capture clothing attributes to generate a description for a fashion-image is proposed. The proposed model tries to use both spatial and channel wise attentions [6]. By leveraging the spatial mechanism, it becomes possible to encode the context of every pixel in the image, leading to more precise predictions of density maps at the pixel level, representing for object/item in the input image. While the latter focuses on extracting more distinct features among various channels to enable the model to pay attention to texture or pattern. It is important to note that each filter of a CNN acts as a pattern detector, and every channel of the feature map corresponds to the activation response of its respective convolutional filter. Therefore, implementing an attention mechanism in a channel-wise manner can be seen as a process of choosing semantic attributes. By combining spatial and channel-wise attention, the sentence generation context can be dynamically modulated in multi-layer feature maps, which contain both spatial or location information (i.e., where) and attribute information (i.e., what) [6] that the model should focus on. The more detail about channel attention and the cooperation between channel attention and spatial attention is in section 3.

Moreover, it is important to note that most of the datasets currently available for the fashion image captioning task do not include sufficient information of fashion item attributes. Therefore, we have also refined some datasets related to various fashion problems, such as DeepFashion [7] and Fashion-Gen [8], to make them suitable for use in fashion captioning tasks involving attributes. Finally, we have evaluated the effectiveness of our proposed approach on the Fashion-Gen dataset, which we refined to incorporate the fashion attributes of the DeepFashion dataset. In summary, our paper makes two main contributions.

Firstly, we propose a novel approach that combines channel-wise and spatial attention for fashion image captioning. This approach is designed to capture not only the relationships between items in an image, but also their fashion attribute features such as style, texture, shape, and fabric. Secondly, we refine the DeepFashion [7] and Fashion-Gen [8] datasets to make them suitable for the fashion image captioning task. We then evaluate the performance of our novel method on the Fashion-Gen dataset. By doing so, we demonstrate the usefulness of our approach for generating more accurate and descriptive captions for fashion images.

The paper is structured as follows. Section 1 provides an overview of the image captioning and highlights some specific requirements for the fashion-image captioning problem. Section 2 discusses current trends in image captioning and identifies some drawbacks of each method. Section 3 introduces our proposed method for fashion-image captioning using attention mechanism. Section 4 presents the results of experiments conducted on the Fashion-Gen dataset using our proposed method. Section 5 concludes the paper by summarizing our contributions and outlining some potential directions for research in this area.

2. RELATED WORKS

Image captioning is a difficult problem that involves generating a description of a given input image. A comprehensive review of existing approaches, evaluation measures, and benchmark datasets for image captioning has been provided by Bernardi *et al.* [9]. To perform image captioning, it is necessary to identify the significant objects in the image, their attributes, and their relationships. Additionally, the generated sentences must be syntactically and semantically correct.

Recent advancements in deep learning (DL) have significantly increased the accuracy of image captioning tasks. In deep learning, image features can be automatically learned from training data, allowing for the handling of large set of images. DL algorithms are adept at handling the complexities and challenges of image captioning. Typically, a DL based image captioning model comprises a two component model encoder (CNN) and decoder (RNN) [2]. The CNN extracts a compact representational feature of the whole image, while the RNN can effectively work with sequential data, such as generating a sequence of words. One common approach is to merge the CNN and RNN, as demonstrated in works by Simonyan and Zisserman [10], Karpathy and Fei-Fei [11] and Vinyals *et al.* [3], to identify patterns in the image and then use that information to generate the image descriptions. There has also been increasing interest in integrating the encoder-decoder with reinforcement learning for image captioning, as shown in works such as [12].

The encoder-decoder image captioning model has limitations, such as when the feature vector extracted from the image by the CNN has much detail information for the RNN to decode into descriptive captions. To address this, attention mechanisms have been added to improve the performance of the model, such as the attention-based model proposed by Xu *et al.* [4], which learns where to attend and when generating image captions. Another approach to improving models is the squeeze-and-excitation networks (SENet [5]), which enables networks to construct features by using spatial and channel-wise features within local fields at each layer. CNNs extract hierarchical information from images using filters and each channel in the convolutional layer contains a significant amount of information. This information can be utilized by RNNs to generate captions in a meaningful manner. Li and Yamaguchi [13] and Hu *et al.* [5] achieved encouraging results using this model, with channel attention performing better than spatial attention and CNN-RNN [3].

In recent days, many different aspects of fashion industry have been solved by using computer vision or deep learning techniques. Liu and Lu [14] tried to predict fashion attributes, while [15] proposed an approach for item matching. There are some other works such as virtual try on [16], [17], trend forecasting [18], fashion design [19], [20], fashion style models [21], [22], clothing category classification [23], [24]. A different research project concentrated on the generation of an automated capsule wardrobe [25]. This approach enables the creation of an outfit using clothes already present in the user's wardrobe. Fashion recommendation is currently one of the most widely used applications in the fashion industry. It is goal is to determine whether a particular outfit is well-coordinated or not. It can be fashion item-based recommendation [26], [27], or outfit-based recommendations using criteria such as versatility or compatibility [28], [29].

In this study, we propose a model for fashion image captioning task based on channel attention and spatial attention. Instead of using the channel-wise or spatial attention separately as some previous methods, our approach combines both of them to generate a description sentence containing attributes of fashion products such as texture, fabric, shape, and style. First, we enhance the internal representation feature vectors of an attended category of fashion objects by a CNN network as ResNet50, and then that feature vectors will go through channel-wise attention and spatial attention model to construct informative features within local receptive fields at each layer. Finally, the informative features are put into a decoder RNN, to generalize the description for a fashion input image. By incorporating both channel-wise and spatial attention mechanism, we have a strong belief that it can improve performance of fashion image captioning with attributes. The next section will describe more detail about our proposed method.

3. METHOD

The recent remarkable achievements in computer vision through deep neural networks (DNNs) have led to the development of various neural network-based methods for generating image captions [3], [30]. These methods typically employ the encoder-decoder paradigm [31], wherein CNNs are utilized to encode images as features, which are then fed to RNNs (or their variations such as gated recurrent unit (GRU) and long short-term memory (LSTM)) for generating image captions. Moreover, visual attention mechanism has been widely used in order to enable models to selectively concentrate on objects of interest, as it helps the model determine when and where to focus its attention [32], [33].

Inspired from these researches, we propose an approach also based on the encoder-decoder [3], [11] framework which uses CNN as the encoder to extract the compressed representation of a whole image, and RNN as the decoder to interpret the representation for generating captions. In addition, to capture clothing attributes (such as shape, fabric, style, texture) of the fashion items inside the input image, the combination of channel-wise and spatial attention is also integrated in our method.

The base line of our proposed method is showed in Figure 1. Our model includes three main parts: encoder-decoder, depth attention part, and spatial attention part. In encoder-decoder architecture, the encoder's role is encoding the inputs into internal representation. Then the internal representation goes through deep attention and spatial attention mechanism to compute the attentive features. This final context vector is passed into the decoder to generate the description for the input fashion images. The following will describe more detail about each part of our proposed method. Firstly, we will provide a brief overview of the encoder-decoder image captioning framework [3], [4], and then we describe channel attention mechanism and our combination of channel and spatial attention for fashion image captioning problem.

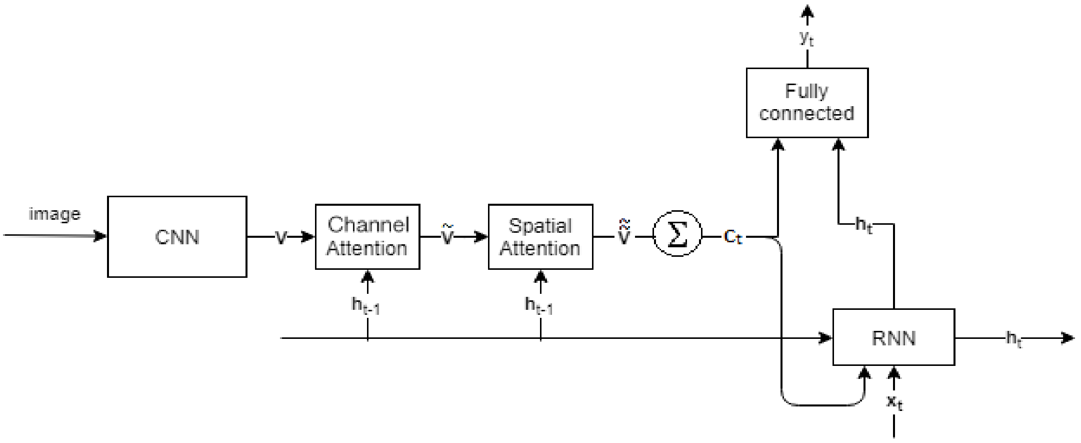


Figure 1. The overview of our proposed model for fashion image description, using two attention mechanisms (spatial and channel-wise) in the encoder-decoder model to interpret caption sentences dynamically in multi-layer feature maps

3.1. Encoder-decoder for fashion image description

In the encoder-decoder framework, the model learns to transform one representation into another. Initially, an encoder tries to encodes the input into a context feature vector, which is later decoded by a decoder network to generate the output. In image captioning, encoders usually base on CNNs, while RNNs or its variants (e.g., GRU, LSTM) are usually used for decoder [3], [34], [35]. The objective of the encoder-decoder is to maximize the similarity of generating the correct caption given an image, as (1),

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum \log p(y|I;\theta) \tag{1}$$

where $\theta$are the model parameters, I is the input image, and $y = [y_1, y_2, \dots, y_t]$ is the corresponding description. The log likelihood of the joint probability distribution can be denoted as (2).

$$\log p(y|I;\theta) = \sum_{t=1}^{T} \log p(y|y_1, \dots, y_t, I) \tag{2}$$

For convenience, the dependency on model parameters is often dropped. In the encoder-decoder framework with RNN, the conditional probability of generating the next word given the previous words and the image features is modeled as (3),

$$\log p(y|y_1, \dots, y_t, I) = f\,(h_t, x_t, c_t) \tag{3}$$

where f is a multilayer perceptron (MLP) that outputs the probability of y_t; and x_t is the embedded input of RNN at time t, and c_t is the context vector extracted from image I at time t. To generate accurate image descriptions, the context vector (c_t) is a crucial component in the neural encoder-decoder framework, as it provides visual information. To compute c_t, our proposed method involves using both spatial and channel-wise attention mechanisms to form a precise representation vector.

3.2. Spatial attention mechanism

Encoder-decoders image captioning model can produce a short description for an input image based on context vector extracted from the whole image. However, using a global feature map to generate captions can lead to suboptimal results because a caption word generally only relates to sub regions of the image. This is because irrelevant regions may be included in the global feature vector. Attention mechanisms have been added to the encoder-decoder model to provide explicit word-level alignment between input image and the output caption [4]. Rather than working on each image region equally, spatial attention mechanism focuses on specific regions in the input given image that are more semantically related to the output caption. At each generation step, the decoder focuses on a specific segment of the encoder, incorporating both the present hidden state in the decoder and the attended hidden states in the encoder to produce the subsequent word. The attention mechanism enables the gradient of the captioning loss to be back-propagated with respect to both the region coordinates and the extracted features, facilitating the model to learn which regions of the image are relevant to the generated caption.

We flatten the width and height of the original image, which has dimensions W and H respectively, to obtain a new feature map with dimensions $W \cdot H \cdot D$. Each location in the feature map can be represented as $v_i \in R^C$. At each time step of the decoder RNN, we employ a single-layer neural network in conjunction with a softmax function to compute the attention distributions α over all the image regions, based on the previous RNN hidden state h_{t-1}. This allows us to selectively attend to relevant image regions when generating the next word in the caption. The (4) define the spatial attention model $\emptyset_s$:

$$\begin{aligned} a &= \tanh W_s V + b_s \oplus W_{hs} h_{t-1} \\ \alpha &= softmax(W_i a + b_i) \end{aligned} \tag{4}$$

where $W_S \in R^{k \times C}$, $W_{hs} \in R^{k \times d}$, $W_i \in R^k$ are used to transform the visual features and the hidden state to a common dimension. Here, $\oplus$ represents the addition by adding each column of the matrix by the vector. The biases vector $b_s \in R^k$, $b_i \in R^k$ are part of the model.

3.3. Channel attention mechanism

As mentioned in [13], the main difference between fashion image captioning and general image captioning problem is that, the describing sentence for general image usually mentions the relationships of different objects appearing in the input image, while fashion product images contain mainly just only one object or fashion item. It makes global attention cannot represent the effective characteristics of fashion product in this case. In addition, [6] represented that using attention mechanisms following a channel-wise one can be considered as the step of choosing meaningful attributes. The channel-wise mechanism [6] focuses on extracting the visual feature V by using CNN filters. Every CNN filter can be considered as a pattern detector when using a response activation of each CNN filter.

Inspired from [6], [13], we adopted a channel attention mechanism for computing the attentive feature map V_t in fashion description problem as (5),

$$V_t = g\ (V, h_t) \tag{5}$$

where g is considered as the function of channel-wise attention, and $V \in R^{w \times h \times d}$ is the image features map extracted from encoder.

Given the features map $V \in R^{w \times h \times d}$ and hidden state $h_t \in R^m$ of the RNN, similar to [6], we calculate the average of the features map along $w \times h$ dimensional and then feed it through a network with the softmax function to compute the attention distribution over the d dimension of the image features map corresponding to each feature. The channel attention model g can be calculated as (6),

$$\begin{aligned} \overline{v} &= \frac{1}{w \cdot h} \textstyle\sum_i^w \sum_j^h V(i,j) \\ a_t^{\overline{v}} &= \tanh\left(W^{\overline{v}} \cdot \overline{v} + b^{\overline{v}}\right) \\ a_t^{h_t} &= \tanh(W^{h_t} \cdot h_t + b^{h_t}) \\ a_t &= a_t^{\overline{v}} + a_t^{h_t} \\ w_t &= \sigma(W^a \cdot a_t) \end{aligned} \tag{6}$$

where v is the mean vector of feature map V over $w \times h$ dimension, $a^v, a^{h_t} \in R^m$ are the compact information of v and h_t respectively, $a_t \in R^m$ acts as interperting infomation of these two a^v, and $a^{h_t}.w_t \in R^d$ is the attention weight. While σ refer to the sigmoid function $W^{v \in R^{m \times d}}$, $W^{h_t} \in R^{m \times n}$, $b^v, b^{h_t} \in R^m$ are neural learnable weights and biases. Based on attention weights, the attentive features map can be obtained by (7):

$$\tilde{V} = V \odot w_t \tag{7}$$

where $\odot$ is denoted as the matrix element-wise multiplication operator.

3.4. Channel-spatial attention mechanism

When incorporating two attention mechanisms, there are two different orders of implementation. It could be to apply channel-wise attention before spatial attention, or in the reverse way. According to [32], exploiting channel-wise attention would significantly improve the model performance, especially when there are a large number of channels. Inspired from that, in this paper, we proposed to use the channel-wise attention before spatial attention.

The illustration of our proposed model for fashion image description with channel attention is showed in Figure 1. The process involves using channel-wise attention to obtain weights α for the initial feature map, followed by obtaining a channel-wise weighted feature map through a linear combination of the weights α and the feature map V. The spatial attention model is then used on the channel-wise weighted feature map V to obtain spatial weights β. The channel-wise weighted feature map V, as well as the two attention weights α and β, are then fed to f for computing the feature map as (8).

$$\begin{aligned} \widetilde{V_t} &= g(V, h_t) \\ \widetilde{\widetilde{V_t}} &= f(\tilde{V}, h_t) \end{aligned} \tag{8}$$

where g is the channel attention function, was described at (5), f is the arbitrarily spatial attention function that return an attentive features map which has weighted features on all h, w and d dimensions. The more detail of this step can be found in Figure 2. Additionally, we can further compute the context vector $c_t \in R^d$ as defined in (9), and later c_t is used for predicting the next word y_{t+1}.

$$c_t = \frac{1}{w \cdot h} \sum_i^w \sum_j^h \widetilde{V_t}(i, j) \tag{9}$$

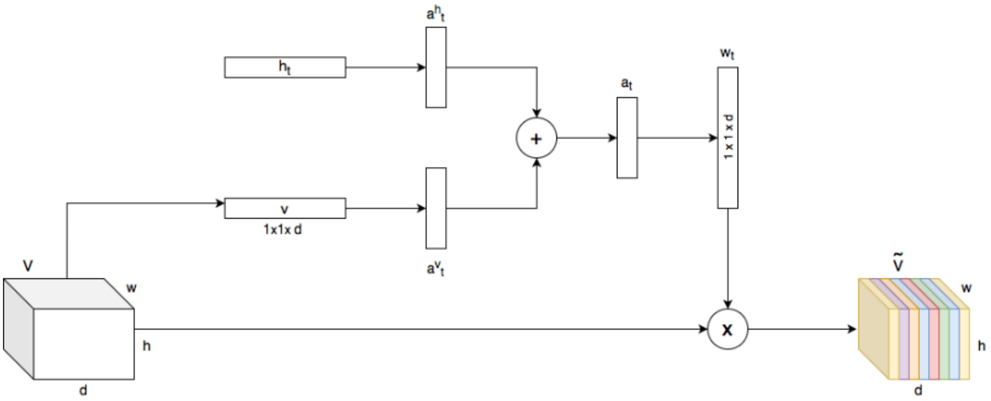


Figure 2. The channel-wise attention model for fashion image captioning. It involves computing the mean vector v of the feature map, and concatenating it to form interpreting information $a_t \in R^m$. The attentive features map is obtained by multiplying attention weight $w_t \in R^m$with the input V

4. EXPERIMENTS

We demonstrate the efficacy of our proposed approach for generating captions for fashion images. by running experiments on two well-known bench-mark dataset: Deep-Fashion and Fashion-Gen. The following is the descriptions of each dataset. After that, details of our experiments are presented. The last part of this section will discuss some quantitative analysis by using three popular measuring metrics (BLEU, ROUGE-L, and CIDEr) [36].

4.1. Datasets

In order to train an appropriate model for the fashion image description, we have utilized two datasets which are known as the benchmarks in fashion field. Based on the characteristics of every dataset, we utilized for corresponding tasks. More particularly, DeepFashion [7] is a dataset which consists of rich fashion attributes as texture, shape, style, fabric, part, categories. It is appropriate to train the fashion image features extractor in the encoder part. On the other hand, Fashion-Gen [8] is a fashion image-description pair dataset which includes both fashion images and the corresponding description. Therefore, it is utilized to train for generating the fashion description.

4.1.1. DeepFashion

DeepFashion is a fashion dataset comprising over 800,000 images. It includes diverse images ranging from well-posed shop images to free consumer pictures. It is annotated with clothing item information such as 50 categories, 1,000 fashion attributes, clothing landmarks, and bounding box. To keep the extracted features related to fashion attributes in the encoder. In the experiment, DeepFashion dataset is selected to train a feature extractor that is utilized to extract only focus on the rich fashion attributes features instead of features of general objects as ImageNet dataset. More particularly, we pre-trained ResNet50 on DeepFashion dataset tolearn 1,000 fashion attributes.

4.1.2. Fashion-gen

FashionGen is comprised of 293,008 images, divided into the training set of 260,480 images, the validation set of 32,528, and the testing set of 32,528 images. The images depict fashion items captured from one to six angles, depending on the item category. It includes expert-written descriptive captions of paragraph length for each fashion item. Although originally created for fashion image generation, this dataset can also be used for fashion image captioning. To facilitate this usage, we conducted pre-processing steps including data cleaning, data normalization, and removal of infrequent words unrelated to fashion attributes. To refine the word dictionary for the Fashion-Gen dataset, we utilized the fashion attribute classes that were collected from the DeepFashion dataset. As a result, the word dictionary was reduced by 80%, but the number of images only decreased by 10%. The obtained attributes description of clothing items was utilized to support the training of the fashion image description task in our framework.

4.2. Implementation

First, the encoder is trained to extract features from fashion images, and we obtained a global feature map of $V^{7\times7\times2048}$. The image feature map of the convolution layer of pretrained ResNet50 model from the DeepFashion dataset is considered as the output of the encoder. The DeepFashion dataset is utilized to learn the weights of encoder part because we would like to gain focus on the fashion attributes instead of the characteristics of generality objects.

In the decoder section, we combine the context vector c_t and the last hidden state vector h_{t-1} to form the input vector to the RNN. Then, RNN will generate a prediction vector h_t for the next word y_{t+1}. Moreover, we also utilize two LSTM layers with 512 hidden nodes. During the training of the decoder, we use the Adam optimizer. The learning rate of $1e-4$ for 10 epochs is initialized as a warm-up phase. Then, we continue training for 100 epochs with a learning rate of $1e-5$, which helps the model's weights to learn more effectively by adjusting the rate of weight updates. By adjusting the learning rate, it helps the weights of model easier to learn. On the other hand, we re-use the weights of encoder from the pre-trained processing on the DeepFashion and then we fine-tune both encoder and decoder to be likely to appropriate with the characteristics of fashion images and the descriptions on Fashion-Gen dataset. The batch size is set to 64, and we apply Beam Search [4] with the beam size of 3 when generating captions for the Fashion-Gen dataset. Figure 3 shows some examples of fashion image captions which are generated from our proposed method which incorporates traditional spatial attention and channel-wise attention mechanism, versus the only spatial-wise attention method.

4.3. Quantitative analysis

The effectiveness of our proposed method for fashion image captioning task was evaluated through experiments conducted on the Fashion-Gen dataset and validating the captions by using three popular measuring metrics: BLEU, ROUGE-L, and CIDEr [36]. Our results are compared with ones from three other SOTA methods: the encoder-decoder model, the channel-wise attention [13], and the spatial attention [4]. The comparison is showed in Table 1. In particular, our proposed model obtains the results of 45.6, 36.4, 28.8, 22.8 for BLEU-1, BLEU-2, BLEU-3, BLEU-4 respectively, while ROUGE-L is 50.6 and CIDEr score is 89.7.

First of all, channel-wise attention got the worst results (28, 20.5, 14.4, 10.9, 34.5, 41.2) for all BLUE, ROUGE-L and CIDEr), due to the fact that it could only focus on the visual feature by selecting semantic attributes through CNN filters. It is not able to represent the regional information such as Spatial-wise attention. Secondly, even encoder-decoder model has been successful for general image captioning task, but when applying for fashion image, it cannot capture some clothing attributes of the fashion items inside the input image. That is why its scores are lower than ones of spatial-wise attention and our proposed methods. Spatial-wise attention mechanism provides the descriptions relating to partial regions of an image, and has consistently outperformed its relative channel-wise attention for all validating metrics. However, spatial-wise attention is still missing the visual features of the fashion items as channel-wise attention can capture. Finally, both spatial-wise and channel-wise attention are integrated as a middle layer between encoder and decoder with the aim of leveraging the semantic attributes for fashion captioning. The new proposed model has the capability of capturing both regional information and visual features of the fashion items. The results demonstrate a significant improvement in the performance of the fashion image captioning task using our proposed model.

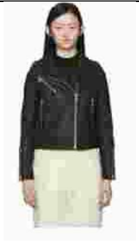
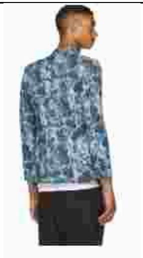
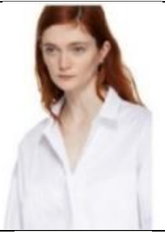
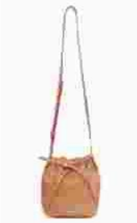
Fashion image	Ground truth captions	Spatial-wise attention	Our proposed method
	Long sleeve supple leather jacket in black, notched lapel collar, off set zip closure at front, zippered pocket at body, epaulet, zippered expansionpanel at cuff, welt pocket at fully lined interior, silver tone hardware, tone stitching.	Long sleeve French terry hoodie in black, draw string at hood, logo printed in white at chest, kangaroo pocket at waist, rib knit cuff and hem, tonal stitching.	Long sleeve grained leather jacket in black, notched lapel collar with press stud fastening, off set zip closure at front, zipped pocket at body, zippered expansion panel at cuff, zippered pocket at fully lined interior, silver tone hard-ware, tonal stitching.
	Long sleeve blazer in tone of blue, abstract pattern printed throughout, the three-button closure at front, notched lapel collar, four button surgeon cut,belt pocket at front and at fullylined interior, single vent at back, tonal stitching.	Long sleeve quilted down filled jacket in black, multi color graphic printed throughout, rib knit stand collar, cuff and hem, two-way zip closure at front, welt pocket at waist, utility pocket at upper sleeve, weltpocket at fully lined interior, silver tone hair wear, tonalstitching, brushed 90% goose down, 10% feather.	Long sleeve shirt in tone of blue, blue and white, button closure at front, single button barrel cuff, tonal stitching.
	Sateen shirt in white, open spread collar, concealed button closure atfront, single button barrel cuff, press stud cinch strap at back collar, press stud cinch strap atback collar, vented side seam, drop tail hem, silver tone hardware, tonal stitching.	Long sleeve cotton poplin shirtin white, spread collar, but- ton closure at front, patchpocket at chest, single button barrel cuff, tonal stitching	Long sleeve cotton poplin shirt in white, spread collar, button closure at front, tonal stitching.
	Structured vegetable tanned leather bucket bag in cammellotan, adjustable shoulder strap with post stud fastening, logo stamp in gold stone at face, self-tie draw string at throat, leatherinterior in rose pink, gold tone hardware, tonal stitching.	Grained calfskin shoulder bag in pink, detachable and adjustable shoulder strap withlanyard clasp fastening, logo stamp in gold tone at face, zipclosure, leather logo patch and patch pocket at interior, textile lining in beige, silver tone hardware, tonal stitching	Structured vegetable tanned leather bucket bag in cammello tan, adjustable shoulder strap, logo stamp in gold tone at face, self-tie drawstring at throat, tonal leather interior, gold tone hardware, tonal stitching.

Figure 3. Examples of captions generated for Fashion-Gen dataset

Table 1. The comparison between our proposed method and other methods for fashion image captioning task on fashion-gen dataset

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
Encoder-Decoder [3]	29.2	23.5	16. 4	12.4	45.5	512.6
Channel-wise [13]	28	20.5	14.4	10.9	34.5	41.2
Spatial-wise [4]	40.8	33.3	26.7	22.1	50.2	91.3
Our proposed	45.6	36.4	28.8	22.8	50.6	89.7

5. CONCLUSION

In summary, we introduce a novel model for fashion image captioning, drawing inspiration from the channel attention mechanism combined with spatial attention, with the aim of enhancing the attention on the details of fashion attributes. The channel attention is used to gain insights into the representation of deep features rather than only the spatial dimension. By incorporating between channel attention and spatial attention, our proposed model is capable of capturing both the relationships between items within the input fashion image and the specific attributes of the clothes. The experiments show the effectiveness of channel attention combined with spatial attention, when compared with the obtained results only applying spatial attention in case of the fashion image captioning with attributes on the Fashion-Gen dataset. In the future, we plan to compare our proposed model with different aspects of attention mechanism on various fashion image datasets to demonstrate its effectiveness.

ACKNOWLEDGMENT

We sincerely acknowledge that this work was supported by the research project T2022-135, funded by Ho Chi Minh City University of Technology and Education. We are grateful for the university's support, which has been instrumental in the successful progress and completion of this project.

REFERENCES

[1] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10971–10980.

[2] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," *Prepr. arXiv.2004.14231*, Apr. 2020.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *Prepr. arXiv.1411.4555*, Nov. 2014.

[4] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *Prepr. arXiv.1502.03044*, Feb. 2015.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[6] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6298–6306, doi: 10.1109/CVPR.2017.667.

[7] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1096–1104, doi: 10.1109/CVPR.2016.124.

[8] N. Rostamzadeh *et al.*, "Fashion-gen: The generative fashion dataset and challenge," *Prepr. arXiv.1806.08317*, Jun. 2018.

[9] R. Bernardi *et al.*, "Automatic description generation from images: a survey of models, datasets, and evaluation measures," *Prepr. arXiv.1601.03896*, Jan. 2016.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Prepr. arXiv.1409.1556*, Sep. 2014.

[11] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, Apr. 2017, doi: 10.1109/TPAMI.2016.2598339.

[12] W. Zhao *et al.*, "Dual learning for cross-domain image captioning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov. 2017, pp. 29–38, doi: 10.1145/3132847.3132920.

[13] S. Li and K. Yamaguchi, "Attention to describe products with attributes," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, May 2017, pp. 215–218, doi: 10.23919/MVA.2017.7986839.

[14] J. Liu and H. Lu, "Deep fashion analysis with feature Map upsampling and landmark-driven attention," in *Computer Vision-ECCV 2018 Workshops*, vol. 11131, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 30–36.

[15] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, "NeuroStylist: neuro compatibility modeling for clothing matching," in *Proceedings of the 25th ACM international conference on Multimedia*, Oct. 2017, pp. 753–761, doi: 10.1145/3123266.3123314.

[16] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 8481–8489, doi: 10.1109/CVPR46437.2021.00838.

[17] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 14126–14135, doi: 10.1109/CVPR46437.2021.01391.

[18] Z. Al-Halah, R. Stiefelhagen, and K. Grauman, "Fashion forward: forecasting visual style in fashion," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 388–397, doi: 10.1109/ICCV.2017.50.

[19] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie, "DesIGN: Design inspiration from generative networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11131, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 37–44.

[20] C. Yu, Y. Hu, Y. Chen, and B. Zeng, "Personalized fashion design," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 9045–9054, doi: 10.1109/ICCV.2019.00914.

[21] M. Takagi, E. Simo-Serra, S. Iizuka, and H. Ishikawa, "What Makes a Style: experimental Analysis of Fashion Prediction," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2017, pp. 2247–2253, doi: 10.1109/ICCVW.2017.263.

[22] W.-L. Hsiao and K. Grauman, "Learning the latent 'Look': unsupervised discovery of a style-coherent embedding from fashion images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 4213–4222, doi: 10.1109/ICCV.2017.451.

[23] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4271–4280, doi: 10.1109/CVPR.2018.00449.

[24] W.-L. Hsiao and K. Grauman, "Creating capsule wardrobes from fashion images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7161–7170, doi: 10.1109/CVPR.2018.00748.

[25] C. Panagiotakis, H. Papadakis, A. Papagrigoriou, and P. Fragopoulou, "Improving recommender systems via a dual training error based correction approach," *Expert Systems with Applications*, vol. 183, Nov. 2021, doi: 10.1016/j.eswa.2021.115386.

[26] Y. Hou, E. Vig, M. Donoser, and L. Bazzani, "Learning attribute-driven disentangled representations for interactive fashion retrieval," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 12127–12137, doi: 10.1109/ICCV48922.2021.01193.

[27] P. Tangseng and T. Okatani, "Toward explainable fashion recommendation," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2142–2151, doi: 10.1109/WACV45572.2020.9093367.

[28] S. Lu, X. Zhu, Y. Wu, X. Wan, and F. Gao, "Outfit compatibility prediction with multi-layered feature fusion network," *Pattern Recognition Letters*, vol. 147, pp. 150–156, Jul. 2021, doi: 10.1016/j.patrec.2021.04.009.

[29] B. T. Nguyen, O. Prakash, and A. H. Vo, "Attention mechanism for fashion image captioning," in *Advances in Intelligent Systems and Computing*, vol. 1284, Y.-P. Huang, W.-J. Wang, H. A. Quoc, L. H. Giang, and N.-L. Hung, Eds. Cham: Springer International Publishing, 2021, pp. 93–104.

[30] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” *Prepr. arXiv.1506.03099*, Jun. 2015.

[31] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: adaptive attention via a visual sentinel for image captioning,” *Prepr. arXiv.1612.01887*, Dec. 2016.

[32] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *NIPS’14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, vol. 2, pp. 3104–3112.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL ’02*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[35] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Barcelona, Spain: Association for Computational Linguistics*, 2004, pp. 74–81.

[36] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

BIOGRAPHIES OF AUTHORS

Bao T. Nguyen received the B.Sc. degree in information technology from University of Sciences, Ho Chi Minh City, Vietnam, in 2002 and the M.Sc. and Ph.D. degrees in computer science from Ritsumeikan University, Japan and Trento University, Italy, in 2011 and 2015, respectively. Currently he is a senior lecturer at HCMC University of Technology and Education, Ho Chi Minh City, Vietnam. Before that, he worked as a researcher at Surgical Planning Laboratory (SPL), a laboratory for mathematics in imaging at Harvard University, and CiMec (Center of Mind and Brain) at Fondazione Bruno Kessler Institution (FBK), Trento, Italy. His research interests include the applications of artificial intelligence (AI), image processing, computer vision, machine learning, medical imaging, neuroinformatic and fashion learning. He is also a research affiliate of IEEE and serves on the Editorial Board for many international journals and conferences of computer science. He can be contacted at email: baont@hcmute.edu.vn.

Son T. Nguyen is the head of Information System Division at Faculty of Information Technology, University of Technology and Education, HCM Vietnam. He got Ph.D. degree from University of Technology, HCM, Vietnam. His research interests include artificial intelligence (AI), machine learning, data mining, and time series. He can be contacted at email: sonnt@hcmute.edu.vn.

Anh H. Vo received the M.S. degree in computer science from University of Sciences, Ho Chi Minh City, Vietnam in 2015, and is currently a Ph.D. candidate. Since 2012, she has been a lecturer and researcher at Information Technology Faculty, Ton Duc Thang University, Vietnam. Her main research interests include image processing, pattern recognition, computer vision, machine learning and data mining. She can be contacted at email: vohoanganh@tdt.edu.vn.