Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards

XuewenYang[1](✉), Heming Zhang[2], Di Jin[3], Yingru Liu[1], Chi-Hao Wu[2], Jianchao Tan[4], Dongliang Xie[5], Jue Wang[6], and Xin Wang[1]

[1] Stony Brook University, Stony Brook, USA
xuewen.yang@stonybrook.edu
[2] USC,Los Angeles, USA
[3] MIT, Cambridge, USA
[4] Kwai Inc., Washington, USA
[5] BUPT, Beijing, China
[6] Megvii, Beijing, China

Abstract. Generating accurate descriptions for online fashion items is important not only for enhancing customers' shopping experiences, but also for the increase of online sales. Besides the need of correctly presenting the attributes of items, the expressions in an enchanting style could better attract customer interests. The goal of this work is to develop a novel learning framework for accurate and expressive fashion captioning. Different from popular work on image captioning, it is hard to identify and describe the rich attributes of fashion items. We seed the description of an item by first identifying its attributes, and introduce *attribute-level semantic* (ALS) reward and *sentence-level semantic* (SLS) reward as metrics to improve the quality of text descriptions. We further integrate the training of our model with maximum likelihood estimation (MLE), attribute embedding, and Reinforcement Learning (RL). To facilitate the learning, we build a new FAshion CAptioning Dataset (FACAD), which contains 993K images and 130K corresponding enchanting and diverse descriptions. Experiments on FACAD demonstrate the effectiveness of our model (Code and data: https://github.com/xuewyang/Fashion_Captioning).

Keywords: Fashion · Captioning · Reinforcement Learning · Semantics

1 Introduction

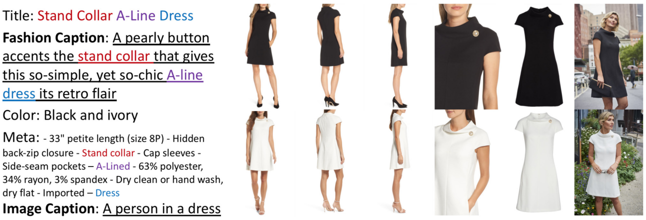
Motivated by the quick global growth of the fashion industry, which is worth trillions of dollars, extensive efforts have been devoted to fashion related research over the last few years. Those research directions include clothing attribute

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58601-0_1) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12358, pp. 1–17, 2020.
https://doi.org/10.1007/978-3-030-58601-0_1

prediction and landmark detection [24,36], fashion recommendation [40], item retrieval [23,37], clothing parsing [7,14], and outfit recommendation [5,11,25].

Accurate and enchanting descriptions of clothes on shopping websites can help customers without fashion knowledge to better understand the features (attributes, style, functionality, benefits to buy, etc.) of the items and increase online sales by enticing more customers. However, manually writing the descriptions is a non-trivial and highly expensive task. Thus, the automatic generation of descriptions is in urgent need. Since there exist no studies on generating fashion related descriptions, in this paper, we propose specific schemes on *Fashion Captioning*. Our design is built upon our newly created *FAshion CAptioning Dataset* (FACAD), the first fashion captioning dataset consisting of over 993K images and 130K descriptions with massive attributes and categories. Compared with general image captioning datasets (e.g. MS COCO [4]), the descriptions of fashion items have three unique features (as can be seen from Fig. 1), which makes the automatic generation of captions a challenging task. First, fashion captioning needs to describe the fine-grained attributes of a single item, while image captioning generally narrates the objects and their relations in the image (e.g., a person in a dress). Second, the expressions to describe the clothes tend to be long so as to present the rich attributes of fashion items. The average length of captions in FACAD is 21 words while a sentence in the MS COCO caption dataset contains 10.4 words in average. Third, FACAD has a more enchanting expression style than MS COCO to arouse greater customer interests. Sentences like "pearly", "so-simple yet so-chic", "retro flair" are more attractive than the plain or "undecorated" MS COCO descriptions.



Title: Stand Collar A-Line Dress

Fashion Caption: A pearly button accents the stand collar that gives this so-simple, yet so-chic A-line dress its retro flair

Color: Black and ivory

Meta: - 33" petite length (size 8P) - Hidden back-zip closure - Stand collar - Cap sleeves - Side-seam pockets – A-Lined - 63% polyester, 34% rayon, 3% spandex - Dry clean or hand wash, dry flat - Imported – Dress

Image Caption: A person in a dress

Fig. 1. An example for Fashion Captioning. The images are of different perspectives, colors and scenarios (shop-street). Other information contained include a title, a description (caption) from a fashion expert, the color info and the meta info. Words in color denotes the attributes used in sentence.

The image captioning problem has been widely studied and achieved great progress in recent years. An encoder-decoder paradigm is generally followed with a deep convolutional neural network (CNN) to encode the input images and a Long Short Term Memory (LSTM) decoder to generate the descriptions [2,15,17,18,39]. The encoder-decoder model is trained via maximum likelihood estimation (MLE), which aims to maximize the likelihood of the next word given the previous words. However, MLE-based methods will cause the model to generate "unmatched" descriptions for the fashion items, where sentences cannot precisely describe the attributes of items. This is due to two reasons.

First, MLE treats the attribute and non-attribute words equally. Attribute words are not emphasized and directly optimized in the training process, however, they are more important and should be considered as the key parts in the evaluation. Second, MLE maximizes its objective word-by-word without considering the global semantic meaning of the sentence. This shortcoming may lead to generating a caption that wrongly describes the category of the item.

To generate better descriptions for fashion items, we propose two semantic rewards as the objective to optimize and train our model using Reinforcement Learning (RL). Specifically, we propose an *attribute-level semantic* (ALS) reward with an attribute-matching algorithm to measure the consistency level of attributes between the generated sentences and ground-truth. By incorporating the semantic metric of attributes into our objective, we increase the quality of sentence generation from the semantic perspective. As a second procedure, we propose a *sentence-level semantic* (SLS) reward to capture the semantic meaning of the whole sentence. Given a text classifier pretrained on the sentence category classification task, the high level features of the generated description, i.e., the category feature, should stay the same as the ground-truth sentence. In this paper, we use the output probability of the generated sentence as the groundtruth category as the SLS reward. Since both ALS reward and SLS reward are non-differentiable, we seek RL to optimize them.

In addition, to guarantee that the image features extracted from the CNN encoder are meaningful and correct, we design a visual attribute predictor to make sure that the predicted attributes match the ground-truth ones. Then attributes extracted are used as the condition in the LSTM decoder to produce the words of description. This work has three main contributions.

1. We build a large-scale fashion captioning dataset FACAD of over 993K images which are comprehensively annotated with categories, attributes and descriptions. To the best of our knowledge, it is the first fashion captioning dataset available. We expect that this dataset will greatly benefit the research community, in not only developing various fashion related algorithms and applications, but also helping visual language related studies.
2. We introduce two novel rewards (ALS and SLS) into the Reinforcement Learning framework to capture the semantics at both the attribute level and the sentence level to largely increase the accuracy of fashion captioning.
3. We introduce a visual attribute predictor to better capture the attributes of the image. The generated description seeded on the attribute information can more accurately describe the item.

2 Related Work

Fashion Studies. Most of the fashion related studies [5,7,23,24,33,36,40] involve images. For outfit recommendation, Cucurull *et al.* [5] used a graph convolutional neural network to model the relations between items in a outfit set, while Vasileva *et al.* [33] used a triplet-net to integrate the type information into the recommendation. Wang *et al.* [36] used an attentive fashion grammar network for landmark detection and clothing category classification. Yu *et al.* [40]

introduced the aesthetic information, which is highly relevant with user preference, into clothing recommending systems. Text information has also been exploited. Han *et al.* [11] used title features to regularize the image features learned. Similar techniques were used in [33]. But no previous studies focus on fashion captioning.

Image Captioning. Image captioning helps machine understand visual information and express it in natural language, and has attracted increasingly interests in computer vision. State-of-the-art approaches [2,15,17,39] mainly use encoder-decoder frameworks with attention to generate captions for images. Xu *et al.* [39] developed soft and hard attention mechanisms to focus on different regions in the image when generating different words. Johnson *et al.* [17] proposed a fully convolutional localization network to generate dense regions of interest and use the generated regions to generate captions. Similarly, Anderson *et al.* [2] and Ma *et al.* [26] used an object detector like Faster R-CNN [29] or Mask R-CNN [12] to extract regions of interests over which an attention mechanism is defined. Regardless of the methods used, image captioning generally describes the contents based on the relative positions and relations of objects in an image. Fashion Captioning, however, needs to describe the implicit attributes of the item which cannot be easily localized by object detectors.

Recently, policy-gradient methods for Reinforcement Learning (RL) have been utilized to train deep end-to-end systems directly on non-differentiable metrics [38]. Commonly the output of the inference is applied to normalize the rewards of RL. Ren *et al.* [30] introduced a decision-making framework utilizing a *policy network* and a *value network* to collaboratively generate captions with reward driven by visual-semantic embedding. Rennie *et al.* [31] used self-critical sequence training for image captioning. The reward is provided using CIDEr [35] metric. Gao *et al.* [8] extended [31] by running a n-step self-critical training. The specific metrics used in RL approach are hard to generalize to other applications, and optimizing specific metrics often impact other metrics severely. However, the semantic rewards we introduce are general and effective in improving the quality of caption generation.

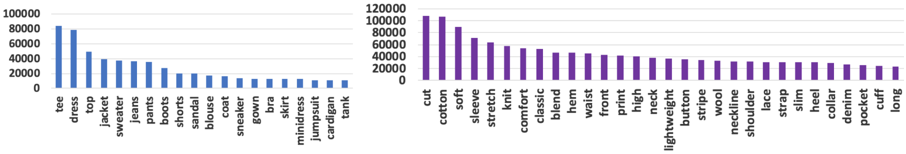
3 The FAshion CAptioning Dataset

We introduce a new dataset - FAshion CAptioning Dataset (FACAD) - to study captioning for fashion items. In this section, we will describe how FACAD is built and what are its special properties.

3.1 Data Collection, Labeling and Pre-Processing

We mainly crawl fashion images with detailed information using Google Chrome, which can be exploited for the fashion captioning task. Each clothing item has on average 6–7 images of various colors and poses. The resolution of the images is 1560 × 2392, much higher than other fashion datasets.

In order to better understand fashion items, we label them with rich categories and attributes. An example category of clothes can be “dress” or “T-shirt”, while an attribute such as “pink” or “lace” provides some detailed information about a specific item. The list of the categories is generated by picking the last word of the item titles. After manual selection and filtering, there are 472 total valuable categories left. We then merge similar categories and only keep ones that contain over 200 items, resulting in 78 unique categories. Each item belongs to only one category. The number of items contained by the top-20 categories are shown in Fig. 2a.



(a)

(b)

Fig. 2. (a) Number of items in the top-20 categories. (b) Number of items in the top-30 attributes.

Since there are a large number of attributes and each image can have several attributes, manual labeling is non-trivial. We utilize the title, description and meta data to help label attributes for the items. Specifically, we first extract the nouns and adjectives in the title using Stanford Parser [32], and then select a noun or adjective as the attribute word if it also appears in the caption and meta data. The total number of attributes we extracted is over 3000 and we only keep those that appear in more than 10 items, resulting in a list of 990 attributes. Each item owns approximately 7.3 attributes. We show the number of items that are associated with the top-30 attributes in Fig. 2b.

To have clean captions, we tokenize the descriptions using NLTK tokenizer[1] and remove the non-alphanumeric words. We lowercase all caption words.

3.2 Comparison with Other Datasets

The statistics of our FACAD is shown in Table 1. Compared with other fashion datasets such as [9,10,24,42,43], FACAD has two outstanding properties. First, it is the biggest fashion datasets, with over 993K diverse fashion images of all four seasons, ages (kids and adults), categories (clothing, shoes, bag, accessories, etc.), angles of human body (front, back, side, etc.). Second, it is the first dataset to tackle captioning problem for fashion items. 130K descriptions with average length of 21 words was pre-processed for future researches.

Compared with MS COCO [4] image captioning dataset, FACAD is different in three aspects. First, FACAD contains the fine-grained descriptions of attributes of fashion-related items, while MS COCO narrates the objects and

[1] https://www.nltk.org/api/nltk.tokenize.html.

Table 1. Comparison of different datasets. $*$ Image sizes are approximate values. CAT: category, AT: attribute, CAP: caption, FC: fashion captioning, IC: image captioning, CLS: fashion classification, SEG: segmentation, RET: retrieval.

Datasets	# img	Img size$*$	# CAT	# AT	# CAP	Avg len	Style	Task
FACAD	993K	1560×2392	78	990	130K	21	enchanting	FC
MS COCO [4]	123K	640×480	–	–	616K	10.4	Plain	IC
VG [21]	108K	500×500	-	-	5040K	5.7	Plain	IC
DFashion [9,24]	800K	700×1000	50	1000	-	-	–	CLS
Moda [42]	55K	-	13	-	-	-	-	SEG
Fashion AI [43]	357K	512×512	6	41	-	-	-	CLS
Fashion IQ [10]	77K	300×400	3	1000	-	-	-	RET

their relations in general images. Second, FACAD has longer captions (21 words per sentence on average) compared with 10.4 words per sentence of the MS COCO caption dataset, imposing more difficulty for text generation. Third, the expression style of FACAD is enchanting, while that of MS COCO is plain without rich expressions. As illustrated in Fig. 1, words like “pearly”, “so-simple yet so-chic”, “retro flair” are more attractive than the plain MS COCO descriptions, like “a person in a dress”. This special enchanting style is important in better describing an item and attracting more customers, but also imposes another challenge for building the caption models.

4 Respecting Semantics for Fashion Captioning

In this section, we first formulate the basic fashion captioning problem and its general solution using Maximum Likelihood Estimation (MLE). We then propose a set of strategies to increase the performance of fashion captions: 1) learning specific fashion attributes from the image; 2) establishing attribute-level and sentence-level semantic rewards so that the caption can be generated to be more similar to the ground truth through Reinforcement Learning (RL); 3) alternative training with MLE and RL to optimize the model.

4.1 Basic Problem Formulation

We define a dataset of image-sentence pairs as $\mathcal{D} = \{(X, Y)\}$. Given an item image X, the objective of Fashion Captioning is to generate a description $Y = \{y_1, \ldots, y_T\}$ with a sequence of T words, $y_i \in V^K$ being the i-th word, V^K being the vocabulary of K words. The beginning of each sentence is marked with a special <BOS> token, and the end with an <EOS> token. We denote $\mathbf{y}_i$ as the embedding for word y_i. To generate a caption, the objective of our model is to minimize the negative log-likelihood of the correct caption using maximum likelihood estimation (MLE):

$$\mathcal{L}_{MLE} = -\sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, X). \tag{1}$$

As shown in Fig. 3, we use an encoder-decoder architecture to achieve this objective. The encoder is a pre-trained CNN, which takes an image as the input and extracts B image features, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_B\}$. We dynamically re-weight the input image features $\mathbf{X}$ with an attention matrix γ to focus on specific regions of the image at each time step t [39], which results in a weighted image feature $\mathbf{x}_t = \sum_{i=1}^{B} \gamma_t^i \mathbf{x}_i$. The weighted image feature is then fed into a decoder which is a Long Short-Term Memory (LSTM) network for sentence generation. The decoder predicts one word at a time and controls the fluency of the generated sentence. More specifically, when predicting the word at the t-th step, the decoder takes as input the embedding of the generated word y_{t-1}, the weighted image feature $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$. The initial memory state and hidden state of the LSTM are initialized by an average of the image features fed through two feed-forward networks f_c and f_h which are trained together with the whole model: $\mathbf{c}_0 = f_c(\frac{1}{B}\sum_{i=1}^{B}\mathbf{x}_i)$, $\mathbf{h}_0 = f_h(\frac{1}{B}\sum_{i=1}^{B}\mathbf{x}_i)$. The decoder then outputs a hidden state $\mathbf{h}_t$ (Eq. 2) and applies a linear layer f and a *softmax* layer to get the probability of the next word (Eq. 3):

$$\mathbf{h}_t = LSTM([\mathbf{y}_{t-1}; \mathbf{x}_t], \mathbf{h}_{t-1}) \tag{2}$$

$$p_\theta(y_t | y_{1:t-1}, \mathbf{x}_t) = softmax(f(\mathbf{h}_t)) \tag{3}$$

where $[;]$ denotes vector concatenation.

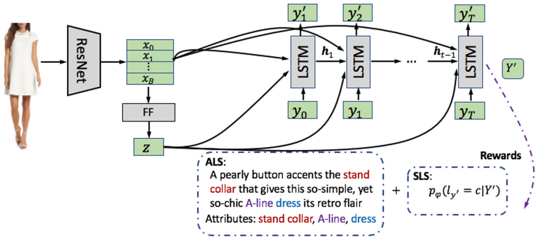

ResNet
x_0
x_1
⋮
x_B
FF
z
y'_1
y'_2
y'_T
LSTM
h_1
LSTM
…
h_{t-1}
LSTM
Y'
y_0
y_1
y_T
Rewards
ALS:
A pearly button accents the stand collar that gives this so-simple, yet so-chic A-line dress its retro flair
Attributes: stand collar, A-line, dress
+
SLS:
$p_\varphi(l_{y'} = c|Y')$

Fig. 3. The proposed model architecture and rewards.

4.2 Attribute Embedding

To make sure that the caption correctly describes the item attributes, we introduce an attribute feature $\mathbf{z}$ into the model, which modifies Eq. 1 into:

$$\mathcal{L}_{MLE} = -\sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, \mathbf{z}, X). \tag{4}$$

This objective aims at seeding sentence generation with the attribute feature of the image. To regularize the encoder to output attribute-correct features, we add a visual attribute predictor to the encoder-decoder model. As each item in the FACAD has its attributes shown in the captions, the predictor can be trained by solving the problem of multi-label classification. The trained model can be applied to extract the attributes of an image to produce the caption.

Figure 3 illustrates the attribute prediction network. We attach a feed-forward (FF) network to the CNN feature extractor, and its output is fed into a sigmoid layer to produce a probability vector and calculate multi-class multi-label loss. We can then modify Eq. 2 and Eq. 3 to include the attribute embedding as:

$$\mathbf{h}_t = LSTM([\mathbf{y}_{t-1}; \mathbf{x}_t; \mathbf{z}], \mathbf{h}_{t-1}) \tag{5}$$

$$p_\theta(y_t|y_{1:t-1}, \mathbf{x}_t, \mathbf{z}) = softmax(f_h(\mathbf{h}_t)) \tag{6}$$

where $\mathbf{z}$ is the attribute features before the output layer, $[;]$ denotes vector concatenation.

4.3 Increasing the Accuracy of Captioning with Semantic Rewards

Simply training with MLE can force the model to generate most likely words in the vocabulary, but not help decode the attributes that are crucial to the fashion captioning. To solve this issue, we propose to exploit two semantic metrics to increase the accuracy of fashion captioning: an attribute-level semantic reward to encourage our model to generate a sentence with more attributes in the image, and a sentence-level semantic reward to encourage the generated sentence to more accurately describe the category of a fashion item. Because optimizing the two rewards is a non-differentiable process, during the MLE training, we supplement fashion captioning with a Reinforcement Learning (RL) process.

In the RL process, our encoder-decoder network with attribute predictor can be viewed as an *agent* that interacts with an external *environment* (words and image features) and takes the *action* to predict the next word. After each action, the agent updates its internal *state* (cells and hidden states of the LSTM, attention weights, etc.). Upon generating the *end-of-sequence* (<EOS>) token, the agent observes a *reward* r as a judgement of how good the overall decision is. We have designed two levels of rewards, as defined below:

Attribute-Level Semantic (ALS) Reward. We propose the use of *attribute-level semantic* (ALS) reward to encourage our model to locally generate as many correct attributes as possible in a caption. First, we need to represent an attribute with a phrase. We denote a contiguous sequence of n words as an n-gram, and we only consider $n = 1, 2$ since nearly all the attributes contain 1 or 2 words. We call an n-gram that contains a correct attribute a tuple t_n. That is, a tuple t_n in the generated sentence contains the attribute in the groundtruth sentence and results in an attribute "Match". We define the proportion of "Matching" for attributes of n words in a generated sentence as: $P(n) = \frac{Match(n)}{H(n)}$, where $H(n)$ is the total number of n-grams contained by a sentence generated. An n-gram may or may not contain an attribute. For a generated sentence with M words, $H(n) = M + 1 - n$. The total number of "Matches" is defined as:

$$Match(n) = \sum_{t_n} \min(C_g(t_n), C_r(t_n)) \tag{7}$$

where $C_g(t_n)$ is the number of times a tuple t_n occurs in the generated sentence, and $C_r(t_n)$ is the number of times the same tuple t_n occurs in the groundtruth caption. We use min() to make sure that the generated sentence does not contain more repeated attributes than the groundtruth. We then define the ALS reward as:

$$r_{ALS} = \beta\{\prod_{n=1}^{2} P(n)\}^{\frac{1}{n}} \tag{8}$$

where β is used to penalize short sentences which is defined as:

$$\beta = \exp\{\min(0, \frac{l-L}{l})\} \tag{9}$$

where L is the length of the groundtruth and l is the length of the generated sentence. When the generated sentence is much shorter than the groundtruth, although the model can decode the correct attributes with a high reward, the sentence may not be expressive with an enchanting style. We thus leverage a penalization factor to discourage this.

Sentence-Level Semantic (SLS) Reward. The use of attribute-level semantic score can help generate a sentence with more correct attributes, which thus increases the similarity of the generated sentence with the groundtruth one at the local level. To further increase the similarity between the generated sentence and groundtruth caption at the global level, we consider enforcing a generated sentence to describe an item with the correct category. This design principle is derived based on our observation that items of the same category share many attributes, while those of different categories often have totally different sets of attributes. Thus, a sentence generally contains more correct attributes if it can describe an item with a correct category.

To achieve the goal, we pretrain a text category classifier p_ϕ, which is a 3-layer text CNN, using captions as data and their categories as labels (ϕ denotes the parameters of the classifier). Taking the generated sentence $Y' = \{y'_1, \ldots, y'_T\}$ as inputs, the text category classifier will output a probability distribution $p_\phi(l_{Y'}|Y')$, where $l_{Y'}$ is the category label for Y'. The sentence-level semantic reward is defined as:

$$r_{SLS} = p_\phi(l_{Y'} = c|Y') \tag{10}$$

where c is the target category of the sentence.

Overall Semantic Rewards. To encourage our model to improve both the ALS reward and the SLS reward, we use an overall semantic reward which is a weighted sum of the two:

$$r = \alpha_1 r_{ALS} + \alpha_2 r_{SLS} \tag{11}$$

where α_1 and α_1 are two hyper-parameters.

Computing Gradient with REINFORCE. The goal of RL training is to minimize the negative expected reward:

$$\mathcal{L}_r = -\mathbb{E}_{Y' \sim p_\theta}[r(Y')] \tag{12}$$

To compute the gradient $\nabla_\theta \mathcal{L}_r(\theta)$, we use the REINFORCE algorithm [38] to calculate the expected gradient of a non-differentiable reward function. To reduce the variance of the expected rewards, the gradient can be generalized by incorporating a *baseline* b:

$$\nabla_\theta \mathcal{L}_r(\theta) = -\mathbb{E}_{Y' \sim p_\theta}[(r(Y') - b)\nabla_\theta \log p_\theta(Y')] \tag{13}$$

In our experiments, the expected gradient is approximated using H samples from p_θ and the baseline is the average reward of all the H sampled sentences:

$$\nabla_\theta \mathcal{L}_r(\theta) \simeq -\frac{1}{H}\sum_{j=1}^{H}[(r_j(Y'_j) - b)\nabla_\theta \log p_\theta(Y'_j)] \tag{14}$$

where $b = \frac{1}{H}\sum_{j=1}^{H} r(Y'_j)$, $Y'_j \sim p_\theta$ is the j-th sampled sentence from model p_θ and $r_j(Y'_j)$ is its corresponding reward.

4.4 Joint Training of MLE and RL.

In practice, rather than starting RL training from a random policy model, we warm-up our model using MLE and attribute embedding objective till converge. We then integrate the pre-trained MLE, attribute embedding, and RL into one model to retrain until it converges again, following the overall loss function:

$$\mathcal{L} = \mathcal{L}_{MLE} + \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_a \tag{15}$$

with λ_1 and λ_2 being two hyper-parameters.

5 Experiments

5.1 Basic Setting

Dataset and Metrics. We run all methods over FACAD. It contains 993K images and 130K descriptions, and we split the whole dataset, with approximately 794K image-description pairs for training, 99K for validation, and the remaining 100K for test. Images for the same item share the same description. The number of images associated with one item varies, ranging from 2 to 12. As several images in FACAD (e.g., clothes shown in different angles) share the same description, instead of randomly splitting the dataset, we ensure that the images with the same caption are contained in the same data split. We lowercase all sentences and discard non-alphanumeric characters. For words in the training set, we keep the ones that appear at least 5 times, making a vocabulary of 15807 words.

For fair and thorough performance measure, we report results under the commonly used metrics for image captioning, including BLEU [27], METEOR [6], ROUGEL [22], CIDEr [35], SPICE [1]. In addition, we compare the attributes in the generated captions with those in the test set as ground truth to find the average precision rate for each attribute using mean average precision (mAP). To evaluate whether the generated captions belong to the correct category, we report the category prediction accuracy (ACC). We pre-train a 3-layer text CNN [19] as the category classifier p_ϕ, achieving a classification accuracy of 90% on testset.

Network Architecture. As shown in Fig. 3, we use a ResNet-101 [13], pretrained on ImageNet to encode each image feature. Since there is a large domain shift from ImageNet to FACAD, we fine tune the conv4_x and the conv5_x layers to get better image features. The features output from the final convolutional layer are used to further train over FACAD. We use LSTM [16] as our decoder. The input node dimension and the hidden state dimension of LSTM are both set to 512. The word embeddings of size 512 are uniformly initialized within $[-0.1, 0.1]$. After testing with several combinations of the hyper-parameters, we set the $\alpha_1 = \alpha_2 = 1$ to assign equal weights to both rewards, and $\lambda_1 = \lambda_2 = 1$ to balance MLE, attribute prediction and RL objectives during training. The number of samplings in RL training is $H = 5$.

Training Details. All the models are trained according to the following procedure, unless otherwise specified. We initialize all models by training using MLE objective with cross entropy loss with ADAM [20] optimizer at an initial learning rate of 1×10^{-4}. We anneal the learning rate by a factor of 0.9 every two epochs. After the model training converges on the MLE objective, if RL training is further needed in a method, we switch to MLE + RL training till another converge. The overall process takes about 4 days on two NVIDIA 1080 Ti GPUs.

Baseline Methods. To make fair comparisons, we take image captioning models based both on MLE training and training with MLE+RL. For all the baselines, we use their published codes to run the model, performing a hyperparameter search based on the original author's guidelines. We follow their own training schemes to train the models.

MLE-Based Methods. **CNN-C**[3] is a CNN-based image captioning model which uses a masked convolutional decoder for sentence generation. **SAT** [39] applies CNN-LSTM with attention, and we use its hard attention method. **BUTD** [2] combines the bottom-up and the top-down attention, with the bottom-up part containing a set of salient image regions, each is represented by a pooled convolutional feature vector. **LBPF** [28] uses a look back (LB) approach to introduce attention value from the previous time step into the current attention generation and a predict forward (PF) approach to predict the next two words in one time step. **TRANS** [15] proposes the use of geometric attention for image objects based on Transformer [34].

MLE + RL Based Methods. **AC** [41] uses actor-critic Reinforcement Learning algorithm to directly optimize on CIDEr metric. **Embed-RL** [30] utilizes a "policy" and a "value" network to jointly determine the next best word. **SCST** [31] is a self-critical sequence training algorithm. **SCNST** [8] is a n-step self-critical training algorithm extended from [31]. We use 1-2-2-step-maxpro variant which achieved best performance in the paper.

5.2 Performance Evaluations

Results on Fashion Captioning. Our Semantic Rewards guided Fashion Captioning (SRFC) model achieves the highest scores on all seven metrics. Specifically, it provides 1.7, 1.4, 3.5, 7.4, 1.2, 0.054 and 0.042 points of improvement over the best baseline SCNST on BLEU4, METEOR, ROUGEL, CIDEr, SPICE, mAP and ACC respectively, demonstrating the effectiveness of our proposed model in providing fashion captions. The improvement mainly comes from 3 parts, attribute embedding training, ALS reward and SLS reward. To evaluate how much contribution each part provides to the final results, we remove different components from SRFC and see how the performance degrades. For SRFC without attribute embedding, our model experiences the performance drops of 0.8, 0.6, 1.0, 3.0, 0.3, 0.011 and 0.021 points. After removing ALS, the performance of SRFC drops 1.3, 0.8, 1.5, 4.6 and 0.6 points on the first five metrics. For the same five metrics, the removing of SLS results in higher performance degradation, which indicates that the global semantic reward plays a more important role in ensuring accurate description generation. More interestingly, removing ALS produces a larger drop in mAP, while removing SLS impacts more on ACC. This means that ALS focuses more on producing correct attributes locally, while SLS helps ensure the global semantic accuracy of the generated sentence. Removing both ALS and SLS leads to a large decrease of the performance on all metrics, which suggests that most of the improvement is gained by the proposed two semantic rewards. Finally, with the removal of all three components, the performance of our model is similar to that of the baselines without using any proposed techniques. This demonstrates that all three components are necessary to have a good performance on fashion captioning.

Results with Subjective Evaluation As fashion captioning is used for online shopping systems, attracting customers is a very important goal. Automatically evaluating the ability to attract customers is infeasible. Thus, we perform human evaluation on the attraction of generated captions from different models. 5 human judges of different genders and age groups are presented with 200 samples each. Among five participants, two are below 30, two are from 40 to 50 years old, one is over 60. They all have online shopping experiences. Each sample contains an image, 10 generated captions from all 10 models, with the sequence randomly shuffled. Then they are asked to choose the most attractive caption for each sample. To show the agreement rate, we calculate Fleiss' kappa based on our existing experimental results, with the rate is in the range of [0.6,0.8] indicating consistent agreement, while the range [0.4, 0.6] showing moderate agreement. The agreement rates for different models are SRFC (ours) (0.63), SCNST (0.61), SCST (0.62), Embed-RL (0.54), AC (0.56), TRANS (0.52), LBPF (0.55), BUTD (0.53), SAT (0.55), CNN-C (0.54). The results in Table 3 show that our model produces the most attractive captioning (Table 2).

Table 2. Fashion captioning results - scores of different baseline models as well as different variants of our proposed method. **A:** attribute embedding learning. We highlight the **best** model in bold.

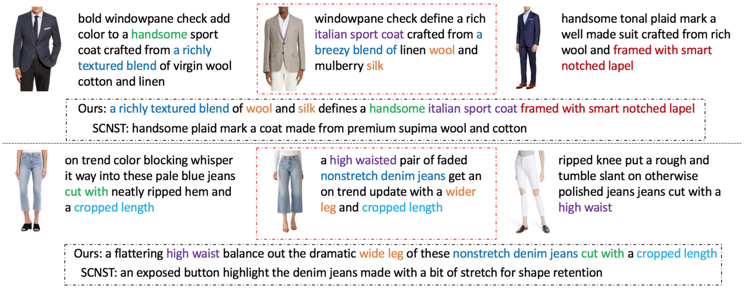
Model	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	mAP	ACC
CNN-C [3]	18.7	18.3	37.8	97.5	16.9	0.133	0.430
SAT [39]	19.1	18.5	38.6	98.4	17.0	0.144	0.433
BUTD [2]	19.9	19.7	39.7	100.1	17.7	0.162	0.439
LBPF [28]	22.2	21.3	43.2	105.3	20.6	0.173	0.471
TRANS [15]	21.2	20.8	42.3	104.5	19.8	0.167	0.455
AC [41]	21.5	20.1	42.8	106.1	19.9	0.166	0.443
Embed-RL [30]	20.9	20.4	42.1	104.7	19.0	0.170	0.459
SCST [31]	22.0	21.2	42.9	106.2	20.5	0.184	0.467
SCNST [8]	22.5	21.8	43.7	107.4	20.7	0.186	0.470
SRFC	**24.2**	**23.2**	**47.2**	**114.8**	**21.9**	**0.240**	**0.512**
SRFC−A	23.4	22.6	46.2	111.8	21.6	0.239	0.491
SRFC−ALS	22.9	22.4	45.7	110.2	21.3	0.233	0.487
SRFC−SLS	22.6	22.2	45.3	109.7	21.1	0.234	0.463
SRFC−ALS−SLS	20.2	19.9	41.5	102.1	18.1	0.178	0.448
SRFC−A−ALS−SLS	19.9	18.7	38.2	98.5	17.1	0.146	0.434

Table 3. Human evaluation on captioning attraction. We highlight the **best** model in bold.

Model	CNN-C	SAT	BUTD	LBPF	TRANS	AC	Embed-RL	SCST	SCNST	SRFC
% best	7.7	7.9	8.1	10.0	8.8	8.4	8.5	10.2	10.7	**19.7**

Qualitative Results and Analysis. Figure 4 shows two qualitative results of our model against SCNST and ground truth. In general, our model can generate more reasonable descriptions compared with SCNST for the target image in the middle column. In the first example, we can see that our model generates a description with more details than SCNST, which only correctly predicted the category and some attributes of the target item.

By providing two other items of the same category and their corresponding captions, we have two interesting observations. First, our model generates descriptions in two steps, it starts learning valuable expressions from similar items (in the same category) based on attributes extracted, and then applies these expressions to describe the target one. Taking the first item (top row of Fig. 4) as an example, our model first gets the correct attributes of the image, i.e., *italian sport coat*, *wool*, *silk*. Then it tries to complete a diverse description by learning from the captions of those items with similar attributes. Specifically, it uses *a richly textured blend* and *handsome* from the first item (left column) and *framed with smart notched lapel* (right column) from the second item to make a new description for the target image. The second observation is that our model can enrich description generation by focusing on the attributes identified even if they are not presented in the groundtrue caption. Even though the *notched lapel* is not described by the ground-truth caption, our model correctly discovers this attribute and generates *framed with smart notched lapel* for it. This is because that *notched lapel* is a frequently referred attribute for items of the category *coat*, and this attribute appears in 11.4% descriptions. Similar phenomena can be found for the second result. The capability of extracting the correct attributes owes to the *Attribute Embedding Learning* and *ALS* modules. The *SLS* can help our model generate diverse captions by referring to those from other items with the same category and similar attributes.



bold windowpane check add color to a handsome sport coat crafted from a richly textured blend of virgin wool cotton and linen

windowpane check define a rich italian sport coat crafted from a breezy blend of linen wool and mulberry silk

handsome tonal plaid mark a well made suit crafted from rich wool and framed with smart notched lapel

Ours: a richly textured blend of wool and silk defines a handsome italian sport coat framed with smart notched lapel

SCNST: handsome plaid mark a coat made from premium supima wool and cotton

on trend color blocking whisper it way into these pale blue jeans cut with neatly ripped hem and a cropped length

a high waisted pair of faded nonstretch denim jeans get an on trend update with a wider leg and cropped length

ripped knee put a rough and tumble slant on otherwise polished jeans jeans cut with a high waist

Ours: a flattering high waist balance out the dramatic wide leg of these nonstretch denim jeans cut with a cropped length

SCNST: an exposed button highlight the denim jeans made with a bit of stretch for shape retention

Fig. 4. Two qualitative results of SRFC compared with the groundtruth and SCNST. Two target items and their corresponding groundtruth are shown in the red dash-dotted boxes in the middle column. The black dash-dotted boxes contain the captions generated by our model and SCNST. Our model diversely learns different expressions from the other items (on the first and third columns) to describe the target item.

6 Conclusion

In this work, we propose a novel learning framework for *fashion captioning* and create the first fashion captioning dataset FACAD. In light of describing fashion items in a correct and expressive manner, we define two novel metrics ALS and SLS, based on which we concurrently train our model with MLE, attribute embedding and RL training. Since this is the first work on fashion captioning, we apply the evaluation metrics commonly used in the general image captioning. Further research is needed to develop better evaluation metrics.

Acknowledgements. This work is supported in part by the National Science Foundation under Grants NSF ECCS 1731238 and NSF CCF 2007313.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24

2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
4. Chen, X., et al.: Microsoft COCO captions: Data collection and evaluation server (2015)
5. Cucurull, G., Taslakian, P., Vazquez, D.: Context-aware visual compatibility prediction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
6. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the 9th Workshop on Statistical Machine Translation (2014)
7. Gabale, V., Prabhu Subramanian, A.: How to Extract Fashion Trends from Social Media? A Robust Object Detector With Support For Unsupervised Learning. arXiv e-prints (2018)
8. Gao, J., Wang, S., Wang, S., Ma, S., Gao, W.: Self-critical n-step training for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
9. Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P.: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: CVPR (2019)
10. Guo, X., Wu, H., Gao, Y., Rennie, S., Feris, R.: The fashion IQ dataset: Retrieving images by combining side information and relative natural language feedback. arXiv preprint arXiv:1905.12794 (2019)
11. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional LSTMs. In: ACM Multimedia (2017)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. He, Y., Yang, L., Chen, L.: Real-time fashion-guided clothing semantic parsing: a lightweight multi-scale inception neural network and benchmark. In: AAAI Workshops (2017)
15. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: transforming objects into words. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
17. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
18. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 664–676 (2017)
19. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2015)

21. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7
22. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
23. Liu, S., et al.: Hi, magic closet, tell me what to wear! In: Proceedings of the 20th ACM International Conference on Multimedia (2012)
24. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Lu, Z., Hu, Y., Jiang, Y., Chen, Y., Zeng, B.: Learning binary code for personalized fashion recommendation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
26. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., Alregib, G., Graf, H.: Attend and interact: Higher-order object interactions for video understanding (2017)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002)
28. Qin, Y., Du, J., Zhang, Y., Lu, H.: Look back and predict forward in image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
30. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward (2017)
31. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
32. Socher, R., Bauer, J., Manning, C.D., Ng, A.Y.: Parsing with compositional vector grammars. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2013)
33. Vasileva, M.I., Plummer, B.A., Dusad, K., Rajpal, S., Kumar, R., Forsyth, D.: Learning type-aware embeddings for fashion compatibility. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 405–421. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_24
34. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
35. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR (2015)
36. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
37. Wang, Z., Gu, Y., Zhang, Y., Zhou, J., Gu, X.: Clothing retrieval with visual attention model. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 (2017)
38. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**, 229–256 (1992). https://doi.org/10.1007/BF00992696

39. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
40. Yu, W., Zhang, H., He, X., Chen, X., Xiong, L., Qin, Z.: Aesthetic-based clothing recommendation. In: Proceedings of the 2018 World Wide Web Conference (2018)
41. Zhang, L., et al.: Actor-critic sequence training for image captioning. In: NIPS workshop (2017)
42. Zheng, S., Yang, F., Kiapour, M.H., Piramuthu., R.: ModaNet: a large-scale street fashion dataset with polygon annotations. In: ACM Multimedia (2018)
43. Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: FashionAI: a hierarchical dataset for fashion understanding. In: CVPRW (2019)