Attr2Style: A Transfer Learning Approach for Inferring Fashion Styles via Apparel Attributes

Rajdeep H Banerjee, Abhinav Ravi, Ujjal Kr Dutta
Data Sciences, Myntra, India
{rajdeep.banerjee,abhinav.ravi,ujjal.dutta}@myntra.com

Abstract

Popular fashion e-commerce platforms mostly provide details about low-level *attributes* of an apparel (for example, neck type, dress length, collar type, print etc) on their product detail pages. However, customers usually prefer to buy apparel based on their *style information*, or simply put, *occasion* (for example, party wear, sports wear, casual wear etc). Application of a supervised image-captioning model to generate style-based image captions is limited because obtaining ground-truth annotations in the form of style-based captions is difficult. This is because annotating style-based captions requires a certain amount of fashion domain expertise, and also adds to the costs and manual effort. On the contrary, low-level attribute based annotations are much more easily available. To address this issue, we propose a transfer-learning based image captioning model that is trained on a source dataset with sufficient attribute-based ground-truth captions, and used to predict style-based captions on a target dataset. The target dataset has only a limited amount of images with style-based ground-truth captions. The main motivation of our approach comes from the fact that most often there are correlations among the low-level attributes and the higher-level styles for an apparel. We leverage this fact and train our model in an encoder-decoder based framework using attention mechanism. In particular, the encoder of the model is first trained on the source dataset to obtain latent representations capturing the low-level attributes. The trained model is fine-tuned to generate style-based captions for the target dataset. To highlight the effectiveness of our method, we qualitatively and quantitatively demonstrate that the captions generated by our approach are close to the actual style information for the evaluated apparel. A Proof Of Concept (POC) for our model is under pilot at Myntra (www.myntra.com) where it is exposed to some internal users for feedback.

Introduction

Catalog images of fashion e-commerce websites are mostly annotated with captions providing details about the low-level attributes of an apparel (for example, neck type, dress length, collar type, print etc). Such captions are easier to annotate as low-level attributes being generic in nature are easier to obtain. Often, apparel manufacturers themselves provide such information. However, captions providing *nature* or *style* (or *looks*) information of an apparel (Figure 1a) are relatively less common (for example, party wear, sports wear, casual looks etc). This is despite the fact that users have a higher preference for *style* information over *low-level attributes* while buying apparel for occasions. A straightforward solution to address this problem would be to annotate a dataset

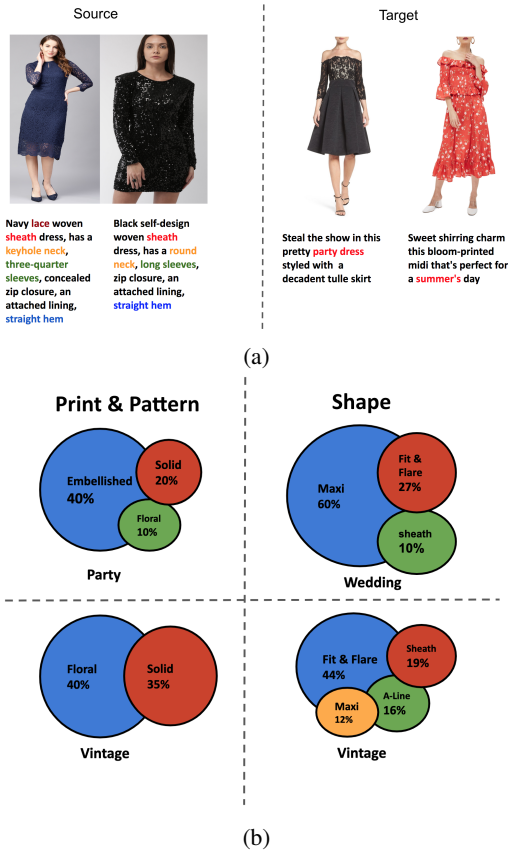
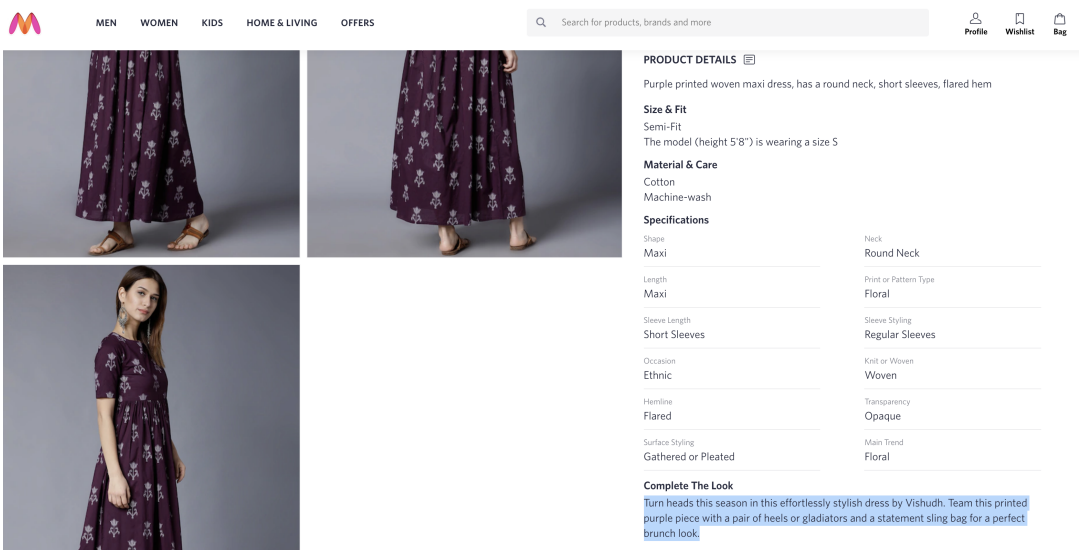
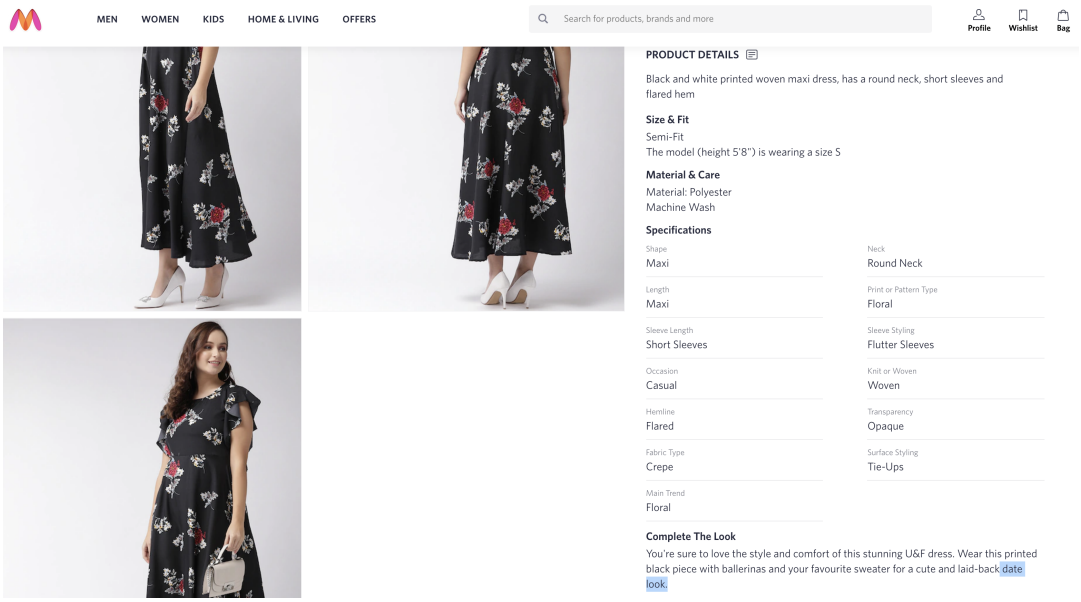


Figure 1: (a) Illustration of attribute and style based captions from source and target domains respectively, and (b) Correlation among low-level attributes and higher-level style information (for the *vintage* style, left quadrant shows the *print and pattern* based attributes, and the right quadrant shows the *shape* based attributes).



(a)



(b)

Figure 2: Sample product display pages on our platform. a) "Complete the look" shows a *looks* based caption, with *attribute*-based caption information in "Product Details", b) An example of style based caption for the *date look*.

with *style* information based captions and train an image captioning model. However, annotating style-based captions is not trivial, and requires a certain amount of fashion domain knowledge, in addition to economic expenses and manual efforts.

A clear look at apparel indicates that *attributes* and *styles* often have correlations among them. For example, as shown in Figure 1b, a high percentage of *party* dresses have *embellished prints* as the dominant attribute, with *floral prints* as the minor one. On the other hand the *vintage* style has the *floral prints* as the dominant attribute. Due to this observation, we conjecture that the lack of style-based ground truth captions for images could be addressed by a transfer learning approach via attribute-based information.

In this paper, we propose an innovative AI system that addresses the above issue by transfer learning. In particular, we apply our method to overcome the scarcity of style-based image captions on a typical fashion e-commerce platform. Figure 2 shows sample images of Product Display Pages (PDP) on our platform, showing both types of captions, and Figure 3 shows common types of *looks*. However, we would like to note that the style-based captions are available for only a handful of images, whereas there is an abundance of images with attribute-based captions. To learn the style-based

Figure 3: Different types of looks.

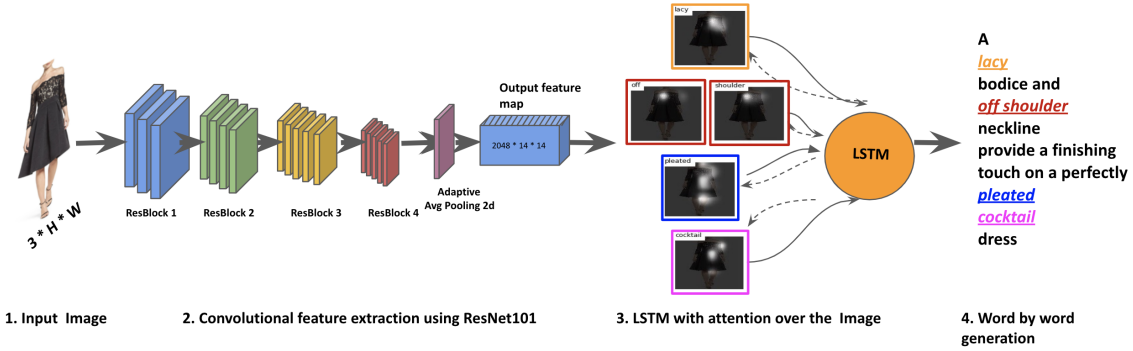


Figure 4: Our image caption network

captions via images with attribute-based captions, we cast our problem into a transfer learning setting. Specifically, we consider the large set of images with attribute-based captions as a *source domain dataset*, and the handful of images with style-based captions as a *target domain dataset*.

Firstly, we train a machine learning model on the *source domain dataset* to learn latent information corresponding to the attributes, and then transfer this knowledge to that of the *target domain dataset* to learn information corresponding to the styles (i.e., looks). To this end, we propose an attention-based image captioning model with an encoder-decoder that leverages transfer learning to obtain style-based captions for target domain images. We first train our model on a source dataset which has abundant attribute-based ground truth captions. The encoder of the trained model now captures the attribute information in the form of latent embeddings. The model is then fine-tuned on the target dataset, which has only a limited number of images with ground truth style-based captions. By virtue of the latent representations learned by the encoder, we are able to *transfer* knowledge of the attributes from the source domain and learn better style-based captions for images from the target domain.

Discussion on the transfer method: The transfer mechanism of our proposed approach is similar in spirit to that of the Domain-Adversarial training of Neural Networks (DANN) method (Ganin et al. 2016), that studied a representation learning approach for domain adaptation. Their approach is directly inspired by the theory on domain adaptation suggesting that, for effective domain transfer to be achieved, predictions must be made based on *features* (that cannot discriminate between the source and target domains). They make use of an adversarial loss to learn *latent features*. Their features are very generic in nature. However, in our case, as shown in Figure 1b, there is a well-known correlation between lower level attributes and higher level styles. Thus, the encoder of the network trained on attribute level captions produces *latent representations* which are agnostic to the domain knowledge (thus, satisfying the theory on domain adaptation). Now this same latent representation helps in *transferring* the low-level attribute information to the higher level styles, and hence act as good representations for caption generation. Despite the *two-phase like training*, in principle, transfer learning is achieved by virtue of the latent features learned. Our approach is indeed a crafty and subtle way of performing transfer learning.

Proposed Method

Model architecture: Figure 4 illustrates our proposed encoder-decoder based image captioning model, that consists of the following major components: i) An encoder, wherein we make use of a ResNet101 (He et al. 2016a) (pretrained on Imagenet (Deng et al. 2009)) to obtain the *latent representations* (that help in *transfer learning*), and ii) A decoder (an LSTM network (Hochreiter and Schmidhuber 1997)), that makes use of the latent features to provide image captions. We incorporate an attention mechanism in the decoder to obtain a correspondence between the feature vectors and portions of the 2-D image. For this, we extract features from a

lower convolution layer of the network, hence allowing the decoder to selectively focus on certain parts of an image (soft attention as in (Xu et al. 2015)). The ResNet based encoder has the option of fine-tuning convolution blocks 2 through 4. The final encoding produced by our ResNet101 encoder has a size of 14×14 with 2048 channels.

Attention-based LSTM model For a pair of sequences $\boldsymbol{A} = \{w_1^a, \cdots, w_m^a\}$ and $\boldsymbol{B} = \{w_1^b, \cdots, w_n^b\}$, let their LSTM encoding be denoted as:

$$\begin{aligned} \boldsymbol{h}_t^a &= \text{lstm}(e(w_t^a), \boldsymbol{h}_{t-1}^a), \forall t \in [1, m] \\ \boldsymbol{h}_t^b &= \text{lstm}(e(w_t^b), \boldsymbol{h}_{t-1}^b), \forall t \in [1, n] \end{aligned} \tag{1}$$

Here, $e(w)$ is the corresponding embedding for the word w. The last hidden state $\boldsymbol{h}_n^b \in \mathbb{R}^d$ is used to *attend* the intermediate representations $\boldsymbol{H}^a = \{\boldsymbol{h}_1^a, \cdots, \boldsymbol{h}_m^a\} \in \mathbb{R}^{m \times d}$ of $\boldsymbol{A}$, using the attention-mechanism (Bahdanau, Cho, and Bengio 2014):

$$\begin{aligned} \tilde{\alpha}_t &= \boldsymbol{v}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_t^a + \boldsymbol{W}_2 \boldsymbol{h}_n^b + \boldsymbol{b}), \forall t \in [1, m] \\ \alpha_t &= \text{softmax}(\tilde{\alpha}_t) \\ \boldsymbol{c}_\alpha &= \sum_{t=1}^{m} \alpha_t \boldsymbol{h}_t^a \end{aligned} \tag{2}$$

Here, $\boldsymbol{W}_1 \in \mathbb{R}^{d1 \times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d1 \times d}$, $\boldsymbol{b} \in \mathbb{R}^{d1}$ and $\boldsymbol{v} \in \mathbb{R}^{d1}$ are parameters to be learned, and $\boldsymbol{c}_\alpha$ is called the *context vector*.

In our case, the LSTM network generates captions one word at every time step t conditioned on the context vector $\boldsymbol{c}_\alpha$, the previous hidden state $\boldsymbol{h}_{t-1}$ and the previously generated words. The context vector $\boldsymbol{c}_\alpha$ is a dynamic representation of the relevant part of the image input at time t.

Let, $\mathbf{a}_i$, $i = 1, ..., L$ denote annotation vectors that are features extracted at different image locations. Using the soft-attention mechanism discussed as above, we can redefine our context vector $\boldsymbol{z}_t$ as:

$$\boldsymbol{z}_t = \sum_i \alpha_{t,i} \mathbf{a}_i \tag{3}$$

Here, we have,

$$\alpha_{t,i} = \frac{exp(e_{ti})}{\sum_{k=1}^{L} exp(e_{tk})} \tag{4}$$

$$e_{ti} = f_{att}(\mathbf{a}_i, \mathbf{h}_{t-1}) \tag{5}$$

where f_{att} is a multilayer perceptron.

Training: We first train the caption generator network using attention to generate *attribute captions* using the source dataset. We use the learned image encoder weights to fine tune the network for generating *style-based captions* using the target dataset.

Experiments

To evaluate the proposed method, we made use of images from our fashion e-commerce website *www.myntra.com*. As the source dataset, we collected a subset of 20000 images that have captions providing *low-level attribute* information (but no *style-based captions*). We collected another limited small subset of 2500 images for which we already had in-house annotated captions describing the *style information*. This second subset is considered as the target dataset, for which we do not make use of the attribute based annotations. We make use of a distinct set of 430 test images (with ground truth style-captions) for evaluating the generated captions. The test data has the following *styles*: party, cocktail, feminine, summer, winter, and none (for rest of images).

	Precision		Recall	
Look	**Ours**	Baseline	**Ours**	Baseline
Party	**72.1**	40.3	**44.0**	23.0
Cocktail	**98.0**	76.1	**50.0**	16.0
Feminine	**85.7**	16.6	**30.0**	1.0
Summer	**100.0**	**100.0**	**16.0**	1.0

Table 1: Quantitative comparison of our method against the baseline in terms of precision and recall for various looks.

	BLEU		**Accuracy**	
	Ours	Baseline	**Ours**	Baseline
Overall	**0.29**	0.26	**0.32**	0.08

Table 2: Quantitative comparison of our method against the baseline in terms of BLEU score and overall accuracy.

Models compared: To demonstrate the effectiveness of the proposed *transfer learning* based approach, we conduct an experiment comparing two models: i) **Baseline Model:** We directly use the available labeled data from the target dataset (with style-based annotations) and train an image captioning model end-to-end using the same architecture as ours (with ImageNet based pretrained weights), for 30 epochs. ii) **AL-model:** The Attribute-Looks model(AL-model) refers to our proposed method. Essentially, we first train our model using the labeled data from the source dataset (with attribute based annotations) in an end-to-end fashion for 30 epochs. Now, we use the same weights for the trained encoder, and fine-tune the model using the limited labeled data from the target dataset (with style-based annotations) for 30 epochs.

Results: Figure 5a shows the empirical performance of both the approaches, using confusion matrices, and the corresponding per *look* precision and recall values are reported in Table 1. In Table 2, we also report the BLEU score to quantify the quality of image captions, and the accuracy over all the looks. A higher value of both these metrics indicates a better captioning performance, and both of these are computed in the range of $0 - 1$.

For a particular test image, we have a ground truth caption corresponding to one of the ground truth *styles* (eg, party, cocktail etc). We make use of our model to predict caption for a test image, and obtain the predicted *style*. Please note that the predicted style is inferred from the generated caption based on the presence of style key words in the caption (eg, party, cocktail etc). Using the ground truth and predicted styles for the set of test images, we can compute the performance metrics like precision, recall, and accuracy using standard definitions in a multi-class classification setting. For eg, to calculate accuracy, we can use the following formula: (TP+TN)/(TP+TN+FP+FN). Here, TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

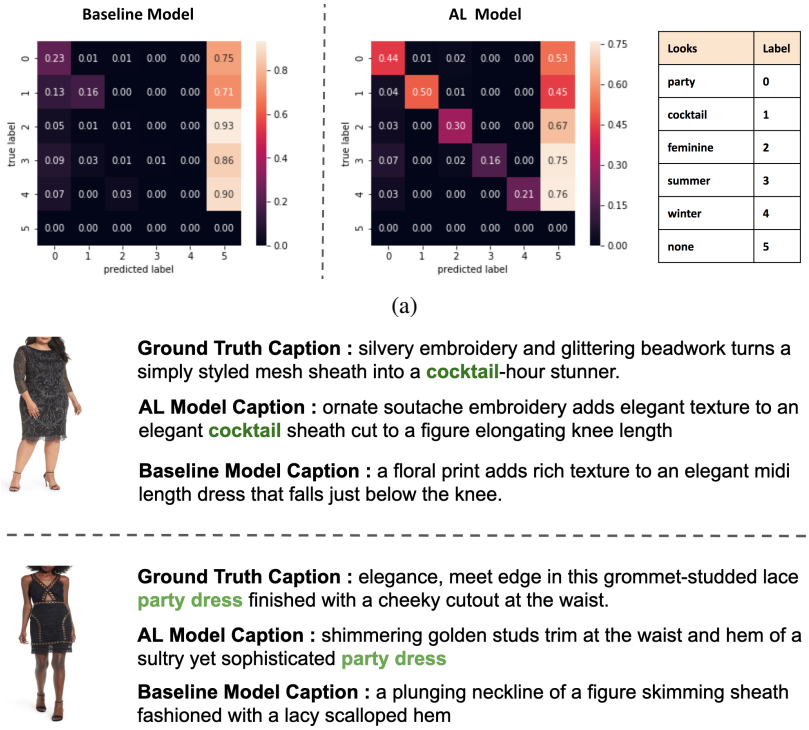


Figure 5: (a) Confusion matrices for the baseline and our method, and (b) Comparison of generated captions using the baseline method and our method.

Note that in Table 1 we report the *class-wise* Precision and Recall values (in this case a class refers to a *style*). In Table 2, we report the accuracy across all the classes.

The best performing method is shown in **bold**. The superiority of our proposed AL-model highlights the benefit of transfer learning employed by our model. Figure 5b compares the captions generated by the baseline and our proposed method. The generated captions by our method are closer to the ground truth. We also show the attention maps corresponding to both the baseline and our approach, in Figure 6. We observed that the maps are more focused to specific regions of an image in our method. This eventually leads to the generation of better quality image captions by our method.

Further details of the system: Following are some of the brief details of the proposed system: i) The hardware used was a Nvidia Tesla V-100 GPU, where the training time was around 2 days. ii) Our model was built for the *dresses* article type on our platform. iii) A pilot was run for around 8-9 months, based on which we figured out that this is a category of focus / emerging category in our geography. Also, studies have revealed that women consumers engage more when styling details are present in the Product Display Page (PDP).

Related Work

Before concluding the paper, we would also like to highlight some important developments in the probem of image captioning. It is a challenging problem that spans concepts from across image understanding as well as natural language generation. Encoder-decoder based deep models (as used in our paper) have shown state-of-the-art performance in image captioning [Yao et al., 2018]. In general, the de-facto approach is to exploit a Convolutional Neural Network (CNN) (eg., ResNet (He et al. 2016b)) to encode image features, followed by a Recurrent Neural Network (RNN) (eg., LSTM (Hochreiter and Schmidhuber 1997)) to generate the caption statements with attention mechanism (Xu et al. 2015).

Considerable efforts have been made to obtain specific improvements. (Anderson et al. 2018) incorporated object-oriented representations. By making use of textual attributes and image regions, some recent works have addressed the problem from a cross-modal point of view (Liu et al. 2019, 2018). However, these approaches instead of comprehending general correlations among image parts, consider an image as unrelated parts (Brendel and Bethge 2018; Geirhos et al. 2018).

Some recent works learn a visual relationship and context-aware attention for image captioning (Wang et al. 2020). Recently (Pan et al. 2020), presented a novel model to capture the second order interactions with channel-wise and spatial bilinear attention. (Cornia et al. 2020) proposed a novel Transformer-based image captioning architecture.

However, as observed, the discussed approaches aim at solving specific problems pertaining to the general image captioning problem by following sophisticated schemes. On the other hand, in our method we wanted to emphasize the

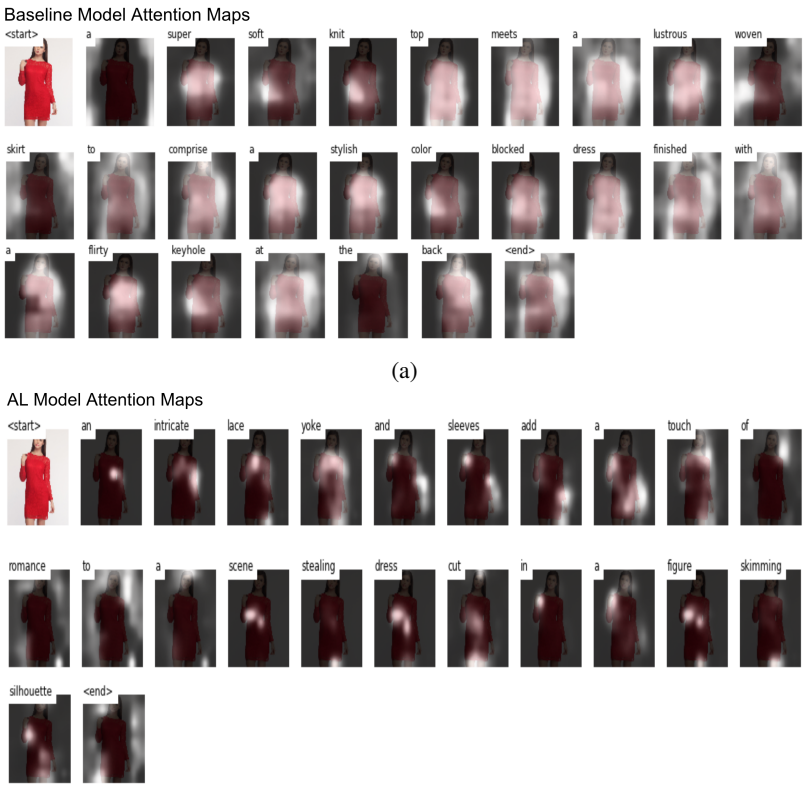


Figure 6: (a) Attention maps of captions using the baseline, and (b) Attention maps of captions by our AL-model.

transfer learning component, rather than addressing the advancement of image captioning. For this reason, we do not use a sophisticated image captioning model. To the best of our knowledge, our application of the image captioning model to perform transfer learning in the fashion application discussed in our paper is novel on its own. **The major contribution of the paper can be seen as the study of the *correlation among the low-level attributes and higher level styles*, and its empirical establishment, by means of both qualitative as well as quantitative observations.** We believe, this could open up newer studies to further leverage such correlations present within fashion article information.

Conclusions

In this paper, we propose a simple, yet effective, transfer learning based approach to address the issue of style-based image captioning for a target dataset. We employ an attention-based image captioning model using an encoder-decoder to obtain style-based captions for an apparel. Because of the correlation among low-level attributes and higher-level style of an apparel, we first train the model on a source dataset with attribute-based ground truth captions. The latent representations obtained by the encoder helps in transfer learning of attribute information to the higher level style-based caption generation. We establish this fact by comparing our model with another version of our model with the same architecture, but without pretraining on the source dataset with attribute information. The captions generated by our model are closer to the actual ground truths, thus showing the benefit of transfer learning. Our method could be used to provide additional style-based captions for fashion apparel, thus improving the overall customer experience, and possibly increasing add-to-cart ratio.

Acknowledgements

We would like to thank our colleague Anoop KR for his contribution in reviewing the core of our work and sharing his insights. We are also grateful to our manager Ravindra Babu for his support and in being a source of inspiration and encouragement throughout the project.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Brendel, W.; and Bethge, M. 2018. Approximating CNNs with Bag-of-local-Features models works surprisingly well

on ImageNet. In *International Conference on Learning Representations (ICLR)*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10578–10587.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. Ieee.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1): 2096–2030.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. LSTM can solve hard long time lag problems. In *Neural Computation*, 473–479.

Liu, F.; Liu, Y.; Ren, X.; Lei, K.; and Sun, X. 2019. Aligning visual regions and textual concepts: Learning fine-grained image representations for image captioning. *arXiv preprint arXiv:1905.06139* .

Liu, F.; Ren, X.; Liu, Y.; Wang, H.; and Sun, X. 2018. simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 137–149.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-Linear Attention Networks for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10971–10980.

Wang, J.; Wang, W.; Wang, L.; Wang, Z.; Feng, D. D.; and Tan, T. 2020. Learning visual relationship and context-aware attention for image captioning. *Pattern Recognition (PR)* 98: 107075.

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