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

Rajdeep Hazra Banerjee, Abhinav Ravi, Ujjal Kr Dutta
Myntra Designs, Bangalore, India

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 apparels 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 demonstrate that the captions generated by our approach are close to the actual style information for the evaluated apparels.

Keywords: Transfer Learning, Image Captioning, Attention-mechanism, Encoder-Decoder.

1 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 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 apparels for occasions. A straightforward solution to address this problem would be to annotate a dataset 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 apparels indicate 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. This observation indicates that the lack of style-based ground truth captions for images from a target domain could be addressed by a transfer learning approach. 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 have 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 have 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.

2 Proposed Method

Model architecture: Figure 2 illustrates our proposed encoder-decoder based image captioning model, that consists of the following major components:
An encoder, wherein we make use of a ResNet101 (pretrained on Imagenet) to obtain the latent representations (that help in transfer learning), and ii) A decoder (a LSTM network), 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 [1]). The final encoding produced by our ResNet101 encoder has a size of $14 \times 14$ with 2048 channels.

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.

3 Experiments

To evaluate the proposed method, we made use of images from the 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 distinct subset of 20000 images, and employed an in-house annotator group to provide captions describing the style information of a small subset of these images. 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).

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).

Results: Figure 3a shows the empirical performance of both the approaches, using confusion matrices. The superiority of our proposed AL-model highlights the benefit of transfer learning employed by our model. Figure 3b 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 4. We observed that the maps are more focused to specific regions of an image in our
4 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 apparels, thus improving the overall customer experience, and possibly increasing add-to-cart ratio.

References

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)