

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Inception Models for Fashion Image Captioning: An Extensive Study on Multiple Datasets

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Del Moro, M., Tudosie, S.C., Vannoni, F., Galassi, A., Ruggeri, F. (2023). Inception Models for Fashion Image Captioning: An Extensive Study on Multiple Datasets. Cham : Springer [10.1007/978-3-031-42448-9_1].

Availability:

This version is available at: <https://hdl.handle.net/11585/941313> since: 2023-09-11

Published:

DOI: http://doi.org/10.1007/978-3-031-42448-9_1

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Del Moro, M., Tudosie, S.C., Vannoni, F., Galassi, A., Ruggeri, F. (2023). Inception Models for Fashion Image Captioning: An Extensive Study on Multiple Datasets. In: Arampatzis, A., et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023. Lecture Notes in Computer Science, vol 14163. Springer, Cham.

The final published version is available online at: https://doi.org/10.1007/978-3-031-42448-9_1

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Inception Models for Fashion Image Captioning: an Extensive Study on Multiple Datasets

Mirko Del Moro, Serban Cristian Tudosie, Francesco Vannoni, Andrea Galassi^[0000–0001–9711–7042], and Federico Ruggeri^[0000–0002–1697–8586]

Department of Computer Science and Engineering, University of Bologna, Italy
a.galassi@unibo.it federico.ruggeri6@unibo.it

Abstract. Fashion e-commerce platforms are becoming increasingly popular. However, scanning, rendering, and captioning fashion items are still done mostly manually. In this work, we address the task of generating a textual description of a fashion item from an image portraying it. We carry out an extensive study with several neural architectures based on InceptionV3. We consider two existing fashion image captioning datasets, FACAD and InFashion. We also curate a novel dataset, Fashion-Cap, that contains more than 290,000 images and 40,000 corresponding captions. In our analysis, we observe significant differences between the three datasets’ captions, with Fashion-Cap having higher quality captions. To the best of our knowledge, this is the most extensive experimental study in fashion image captioning to date. Our experimental results show that our dataset is less challenging than FACAD but more than InFashion, which confirms our insights, suggesting that it could be a valuable benchmark for this domain.

Keywords: Fashion · Dataset · Image Captioning · NLP

1 Introduction

In the last few years, the e-commerce fashion industry has witnessed significant growth. Major worldwide events like the recent COVID-19 pandemic defined a valuable playground for e-commerce sales platforms, whose growth has greatly exceeded even the most generous predictions. As a result, many fashion consumers are progressively adopting e-commerce platforms as their default shopping solution [24]. This phenomenon has led to the definition of e-commerce platforms that cover a wide variety of fashion items and services, which pose a challenge due to the great human effort that they require. Indeed, the definition of an autonomous pipeline for scanning, rendering, and captioning fashion items is still in its infancy, consequently most of the effort is still attributed to human workers. Current research mainly addresses the consumer perspective by defining adequate recommender systems [32]. However, a complete pipeline should contain other components designed to capture a consumer’s attention and provide them with the necessary information in an effective way. For instance,

captions should be short with minimal but relevant details, to be compatible with smartphone screens and voice-based searches.¹

In this work, we discuss the definition of generative models for automatically defining captions for fashion items. Despite the growth of e-commerce fashion platforms, this problem is still scarcely addressed in the literature. To the best of our knowledge, only two datasets designed for fashion image captioning have been released so far: the FACAD dataset [30] and the InFashAI [12] project. We propose an extensive study on these datasets and release a novel one called Fashion-Cap, which we obtain by adapting an image generation dataset to the task of image captioning. We evaluate a well-known generative architecture for image captioning [29], experimenting with different configuration settings and variants to assess the task’s difficulty. Compared to existing contributions, our method relies on the input fashion image and does not leverage additional domain knowledge like fashion attributes [30]. This design choice reflects the purpose of reducing human effort when defining fashion e-commerce platforms. Our contribution is twofold: (i) we release Fashion-Cap, a new dataset for the task of fashion image captioning, which is obtained by adapting and curating a dataset for image generation; (ii) we provide a reproducible and extensive study on three datasets for the fashion image captioning task using several encoder-decoder neural architectures. To the best of our knowledge, we are the first to propose a study on as many datasets in this domain. We make our code and data publicly available.³

2 Related Work

Model pre-training has become the default approach in the image captioning domain, especially for the encoder module [27, 29], as well as in the image-text understanding domain for the vision and language multitask [2, 15, 4]. Yang et al. [30] were the first to propose large pre-trained models for image captioning by proposing the FASHion CAPtioning Dataset (FACAD). In their study, the authors use an encoder-decoder neural architecture, as in [27], but they also integrate task-specific attribute embeddings trained via reinforcement learning. They rely on a set of fashion-related attributes extracted from the input image to regularize model training. More precisely, they introduce attribute-level semantic (ALS) and sentence-level semantic (SLS) rewards as metrics to improve the quality of generated image captions. In contrast, our proposed solution doesn’t require the identification of domain-specific attributes to generate an image caption. Indeed, we speculate that acquiring domain knowledge can become a bottleneck for defining efficient image captioning tools for the fashion industry. In particular, the absence of a standardized set of fashion attributes can lead to a time-consuming attribute identification annotation step.

¹ <https://content26.com/blog/product-description-word-counts-length-matters-2/>

² <https://www.bigcommerce.com/blog/perfect-product-description-formula/>

³ Publicly available repository: <https://www.github.com/NoLogicPlease/Visionizer>

Table 1: Source datasets statistics.

Dataset	Images	Max Resolution	Categories	Captions	Avg. Caption Length	Poses	Task
FACAD [30]	993,000	1560x2392	78	130,000	21	multiple	I. Captioning
InFashAI+DeepFashion [12]	87,821	800x1070	n/a	87,821	9	single	I. Captioning
Fashion-Gen [21]	325,536	1360x1360	48	78,850	30	multiple	I. Generation

Table 2: Composition of datasets used in our study.

Dataset	Images	Train Images	Val Images	Test Images	Resolution	Images per Caption	Avg. Caption Length
Reduced-FACAD	55,021	44,016	5,502	5,503	299x299	1	17
Reduced-InFashAI	86,763	69,410	8,676	8,677	299x299	1	9
Fashion-Cap	290,441	232,352	29,044	29,045	299x299	up to 8	10

Fashion image captioning has been taken into consideration also by Hacheme and Sayouti [12]. They implemented a model based on the *Show and tell* approach [27]: an encoder-decoder architecture in which the encoder is a Convolutional Neural Network (CNN), and the decoder is a Recurrent Neural Network (RNN). They initialized the encoder using a pre-trained ResNet152 [13] and used a Long Short-Term Memory (LSTM) as decoder. In their work, they jointly train their model on two datasets, one of Western-style items and one of African-style items, with the purpose of transferring knowledge between the two. With respect to their work, we add more recent techniques, namely Beam Search and Bahdanau attention [1]. The former is used to improve the decoder performance in the caption generation, while the latter is introduced to make the model more interpretable [28]. Another layer of controllability and interpretability could be added by using a framework for generating controllable and grounded captions through regions, as proposed by [6]. Lastly, differently from them, we do not rely on an index-based representation of words but employ Glove embeddings [19].

Beyond image captioning, artificial intelligence has been applied to the fashion domain for several other purposes, such as generating synthetic images from items description [21], assessing the similarity between two images of fashion items [8], recognizing items characteristics [17], and providing specialized and tailored recommendations [31, 9]. Additional information can be found in the following surveys: [3], [16], and [22].

3 Data

In this study, we consider three sources: the FACAD dataset [30], a collection presented in [12] containing two datasets (InFashAI and DeepFashion), and Fashion-Gen [21]. We select only a subset of the data available in these sources, according to the following principles: (i) the images must be publicly available; (ii) all the images related to the same item must have the same quality; (iii) the captions must be concise. In particular, we implement the last principle by measuring the

average length of the captions across the three sources, which is 20 words, and filtering out any data with a longer caption. Table 1 provides a summary of the original sources, whereas Table 2 shows the datasets used in our study.

3.1 Reduced-FACAD

The FASHIONING CAPTIONING Dataset (FACAD) [30] is a collection of 993,000 high-resolution fashion images. The dataset contains images of fashion items targeting different seasons, ages (kids and adults), and categories (clothing, shoes, bag, accessories, etc.). Each fashion item is collected from different angles (front, back, side, etc.). Figure 1a shows an example. FACAD is the first large dataset built specifically for the image captioning task in the fashion domain. In particular, the dataset contains 130K image captions, each one corresponding to a single clothing item represented in 6-7 images. The average length of the captions is 21 words and each of them contains a single sentence that often includes also information that can be considered subjective (e.g., “so-simple yet so-chic”, “retro flair”). Concerning the image captioning task, FACAD is a challenging dataset since the images and captions were collected from the web through web scraping of fashion websites, and therefore there are cases where captions contain linguistic or format errors.

For each item, there is only one image with a proper background, object position, and image quality that properly represents the fashion item. The other ones, as shown in Figure 1a, are less consistent and they contain noisy elements, e.g., the background. For this reason, we consider only such image for each item and ignore the remaining ones. Additionally, we filter out images with a corresponding caption of more than 20 words. Eventually, we obtain a dataset comprising 55,021 images with corresponding captions. We label this dataset subset as Reduced-FACAD hereafter.

3.2 Reduced-InFashAI

We consider the work of Hacheme and Sayouti [12] as the second source of data. They present a novel dataset, Inclusive Fashion AI (InFashAI), which contains 8,842 clothing images with corresponding captions targeting the African fashion culture. The images were collected from Afrikrea,⁴ a well-known marketplace specializing in fashion items. They also use the DeepFashion dataset [34,17], which contains 78,979 images of Western culture items collected from Pinterest.⁵ Instead of using the original captions, Hacheme and Sayouti constructed new ones through crowdsourcing, instructing a team of volunteers that followed a template-based approach such as: *The (man|woman|lady) is wearing (a|an) (western|african) *item description**. For this reason, image captions are relatively short, with an average length of 9 words. Figure 1b shows an example.

⁴ <https://www.afrikrea.com/>

⁵ <https://www.pinterest.com/>

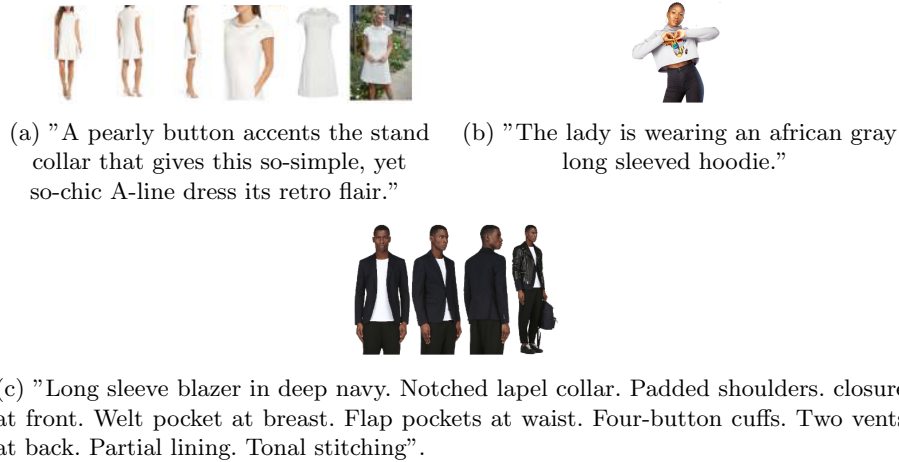


Fig. 1: Examples of fashion item images and corresponding caption in (a) FACAD, (b) InFashAI+DeepFashion, and (c) FashionGen, respectively.

Overall, the resulting dataset contains 87,821 images with corresponding captions. We consider the publicly available version of this dataset, which contains 86,763 images with corresponding captions. We denote this version as Reduced-InFashAI hereafter.

3.3 Fashion-Cap

The Fashion-Gen dataset [21] was originally proposed for the task of image generation. It contains 325,536 high-definition fashion images, but the publicly available version of the dataset only features images in 256x256 resolution. The items were photographed under consistent studio conditions, and the photos are paired with item captions provided by professional stylists. Similarly to FACAD, for each fashion item, multiple images taken from different angles were collected depending on the item category. Figure 1c shows an example. Overall, the dataset contains 78,850 image captions, whose average length is 30 words. This length is due to the fact that captions are articulated and verbose, usually spanning through multiple sentences. The first one typically describes the fashion item with the most relevant characteristics, while the following ones are shorter and contain minor details.

Starting from Fashion-Gen data, we curate a novel dataset for the task of image captioning. We consider the publicly available version of this dataset, which contains 293,018 image-captions pairs. Since all the images associated with a fashion item (and its caption) have the same quality and there are no relevant inconsistencies between them, we do not discard any of them, in contrast to what we have done with FACAD. However, to address the verbosity of the

captions, we filter them by considering only those having 20 or fewer words to obtain concise textual descriptions comparable in length to the ones reported in Reduced-FACAD and Reduced-InFashAI datasets. Furthermore, we consider a text normalization preprocessing step based on regular expressions to remove impurities like excessive blank spaces and special characters. The resulting dataset contains 290,441 images paired with 42,172 unique captions, with an average length of 8 words. We denote the obtained dataset as Fashion-Cap hereafter.

4 Experimental Setting

4.1 Models

We experiment with several models based on a general encoder-decoder architecture. Each step of the captioning process generates a new token of the caption following this scheme:

1. An input image X is encoded by the encoder: $\tilde{X} = ENC(X)$;
2. The encoded image $\tilde{X}$ and the embedding of the $y_0 = \langle \text{start} \rangle$ token for generation are concatenated and fed as input to the decoder;
3. The decoder generates the first token $y_1 = \text{softmax}\left(DEC([\tilde{X} || y_0], h_0)\right)$, where h_0 is the decoder initial hidden state;
4. The decoder iteratively generates the following caption tokens:

$$y_t = \text{softmax}\left(DEC([\tilde{X} || y_{t-1}], h_{t-1})\right)$$

The simplest model, which we address as Baseline, follows a popular encoder-decoder architecture for image captioning and is represented in Figure 2 (top). This architecture was first introduced in [27] and is itself inspired by previous work on sequence-to-sequence translation [25]. The encoder is based on a pre-trained InceptionV3 architecture [26], a popular convolutional neural network for assisting in image analysis and object detection, followed by a single fully connected layer. The decoder comprises a recurrent layer and a stack of two fully connected layers for caption generation. The textual inputs are encoded through trainable embeddings of size 300. Differently from [27], to generate y_t , we use greedy search, which we denote as Max Search. Max Search concerns selecting the token with the highest probability as output at each generation step. We experiment with two variations of the Baseline that differ for the recurrent layer: one uses a GRU [5], the other one an LSTM [14]. This approach is similar to the one used in [12], except that we follow the original model of the decoder, while they replace it with a pre-trained ResNet152 [13].

We enhance the Baseline with more recent techniques, obtaining a model that we call Visionizer, as shown in Figure 2 (bottom). Inspired by [29], we add an attention layer in the decoder, before the concatenation step. Specifically, we employ Bahdanau attention [1], using the hidden state of the recurrent layer as query element [10]. The introduction of this module is motivated by its many successes in Natural Language Processing and Computer Vision tasks, but also

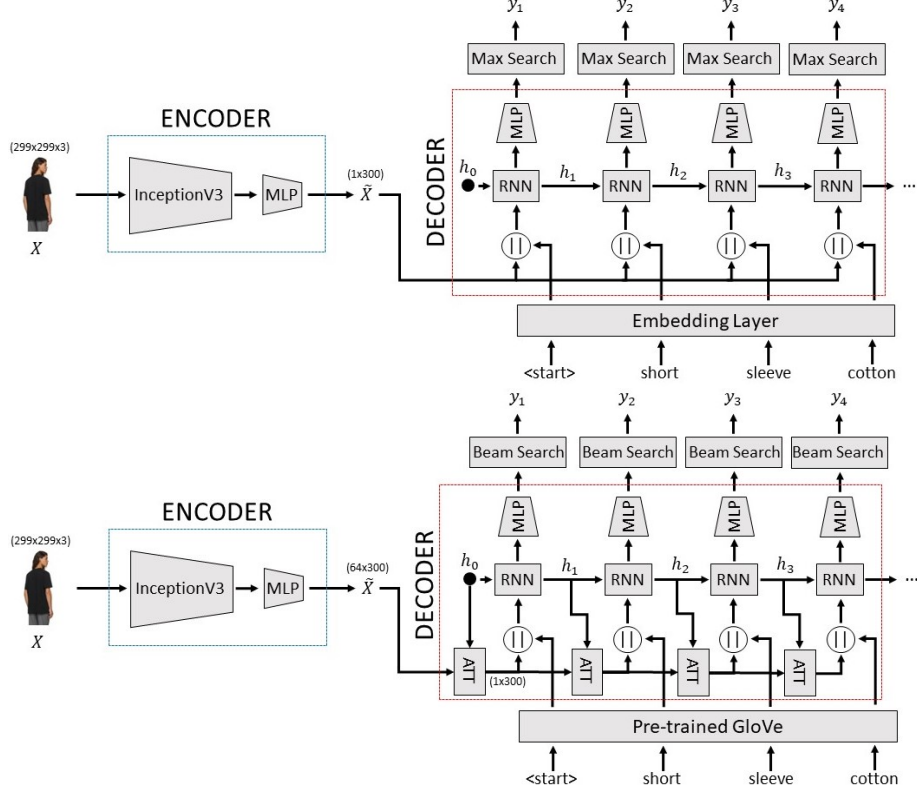


Fig. 2: The architecture of the Baseline approach (top) and Visionizer (bottom).

because it allows interpreting the output of the model [28]. In addition, we encode the textual input using 300-dimensional GloVe embeddings [19], but we keep the encoding layer trainable to also learn out of vocabulary terms (OOV) and fine-tuning the embeddings. As for Baseline, we experiment Visionizer with GRU and LSTM for the recurrent layer. Finally, we also add the possibility with Visionizer to generate captions through beam search. The Beam Search algorithm selects multiple tokens for a position in a given sequence based on conditional probability. Unlike the decoder with max search, on each step of the decoder, beam search keeps track of the top k most probable partial translations (hypotheses). The beam size parameter is used to determine how large is the space of hypothesis.

4.2 Setup

We split each described dataset into train (80%), validation (10%), and test (10%) splits (see Table 2). We train our models with Adam optimizer, using teacher forcing [11] as an additional regularization at training time. Teacher

Table 3: Model performance for fashion image captioning.

Model	Reduced-FACAD			Reduced-InFashAI			Fashion-Cap		
	BLEU	CHRF	BERT	BLEU	CHRF	BERT	BLEU	CHRF	BERT
Baseline (GRU)	0.056	0.105	0.846	0.849	0.822	0.977	0.402	0.395	0.903
Baseline (LSTM)	0.050	0.101	0.846	0.852	0.827	0.978	0.405	0.397	0.905
Visionizer (GRU)	0.086	0.123	0.848	0.897	0.882	0.984	0.509	0.483	0.923
-Beam Search	0.142	0.157	0.827	0.864	0.842	0.979	0.421	0.409	0.905
-Attention	0.097	0.141	0.789	0.847	0.831	0.967	0.412	0.399	0.895
Visionizer (LSTM)	0.087	0.121	0.849	0.898	0.880	0.985	0.520	0.494	0.926
-Beam Search	0.125	0.153	0.826	0.865	0.843	0.979	0.423	0.409	0.907
-Attention	0.083	0.112	0.788	0.848	0.820	0.972	0.391	0.388	0.894

forcing is a strategy for training recurrent neural networks that uses ground truth as input, instead of model output from a prior time step as an input. Training with teacher forcing allows to converge faster, but it leads to exposure bias problems at inference time, because of the unavailability of the ground truth. We fix the resolution of input images to 299x299 resolution to account for different input formats across datasets.

For what concerns hyper-parameters, we train for 2 epochs because the perplexity of the model on the validation set started to degenerate after. We chose the textual embedding size of 300 as suggested in [19]. The batch size is chosen as 64 to match the approach in [12]. We set the learning rate to 0.001 and we use 512 units in the fully connected layer. Lastly, we chose 2 for the beam size and 2 for the k beam parameter to evaluate the impact of the Beam Search using the minimum possible values. Model capacity is an important factor in deep learning and image captioning as shown in [23] and [15], thus suggesting a future study on the model size. Due to computational resource limitations, we did not perform an extensive hyper-parameter calibration search. We leave this as future work. As evaluation metrics, we consider two syntactic-oriented metrics, namely BLEU [18], CHRF [20]. Additionally, we consider BERTScore [33], a recent metric that is based on neural networks and is semantic-oriented. More in detail, BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence, encoding them using BERT [7].

5 Results

Table 3 reports evaluation metrics regarding the image captioning task on the three discussed datasets. In particular, we evaluate each model when the recurrent layer is defined by a GRU and by an LSTM architecture. We also perform an ablation study on Visionizer by removing the Beam Search and the Attention module. Overall, all the models have similar behavior on the three datasets. Reduced-FACAD is clearly the most challenging one, and the best models achieve only a score of ~ 0.14 in BLEU and ~ 0.84 in BERT. On Fashion-Cap the best models reach about ~ 0.52 in BLEU and ~ 0.93 in BERT. Reduced-InFashAI is

clearly the easier dataset: the best models obtain an almost perfect BERT score (~ 0.99) and a considerably high BLEU score (~ 0.90). This is probably due to the fact that the captions follow a template structure, and therefore the generation of many tokens (e.g., the first half of the caption) is quite easy. In all the considered cases, the CHRF score is similar to the BLEU one. We observe that the Visionizer models using the beam search outperform their Baseline counterparts in all datasets and metrics. In particular, the Visionizer with LSTM and beam search performs best across all datasets. The model achieves an improvement in the BLEU score over its baseline counterpart of ~ 3 , ~ 5 , and ~ 12 percentage points on, respectively, Reduced-FACAD, Reduced-InFashAI, and Fashion-Cap. We observe a similar improvement for the same model regarding CHRF.

The alignment between the BLEU and CHRF metrics is expected as both metrics capture syntactic and lexical similarities. In contrast, we observe fewer improvements for BERTScore, possibly motivated by the limited length of the captions. This is particularly evident in Reduced-InFashAI, where captions are shorter and follow a template-based construction.

For what concerns the model without the beam search, we observe inconsistent results across datasets. In particular, the Visionizer model with beam search outperforms its counterpart in Fashion-Cap and Reduced-InFashAI. We observe this performance improvement in all the reported evaluation metrics. In contrast, removing the beam search leads to improved results in Reduced-FACAD. However, it is worth noting that model performance is notably lower compared to the other datasets. Indeed, FACAD is a challenging dataset containing noisy image captions. Therefore, syntactic-oriented metrics like BLEU and CHRF might favor noisy captions similar to the original ones. We speculate that this characteristic of the dataset is responsible for the observed experimental results.

6 Qualitative Analysis

We carry out a qualitative analysis of Visionizer results considering two cases for Reduced-FACAD and Fashion-Cap. For each test set, we analyze the image for which the Visionizer with Max Search obtained the best BLEU score and the one for which it obtained the worst score, to highlight the contribution of the Beam search.

Figure 3 shows examples from Reduced-FACAD dataset. In particular, in Figure 3 (top), the Visionizer with Beam Search successfully captures part of the ground-truth caption concerning the 'soft and stretchy blend'. In contrast, Visionizer with Max Search fails at capturing these details, while we observe that the baseline model repeats this pattern with different adjectives. Concerning worst-generation performance cases, in Figure 3 (bottom), we observe that all models fail at capturing the fashion details described in the ground-truth caption, which are particularly challenging since they involve domain-specific knowledge that may not be retrievable solely from the input image.

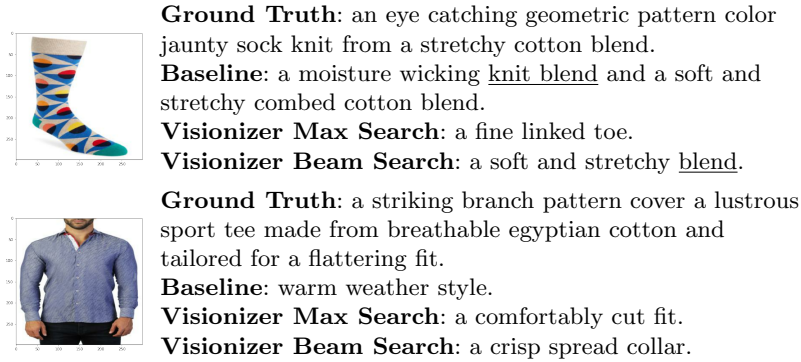


Fig. 3: Examples of generated captions on Reduced-FACAD test set, chosen considering the best BLEU score (top) and worst BLEU score (bottom), with respect to Visionizer Max Search. We underline the main differences.

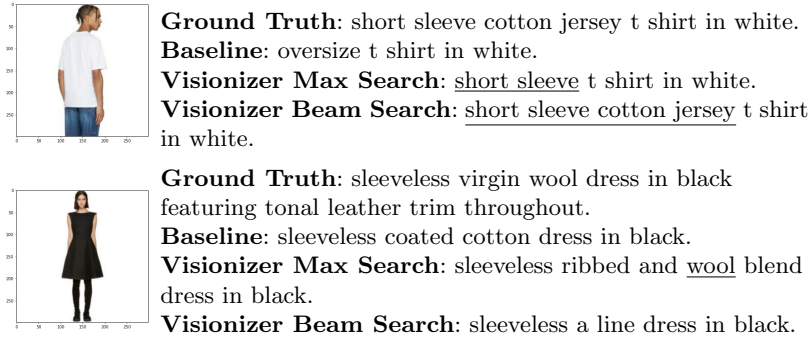


Fig. 4: Examples of generated captions on Fashion-Cap test set, chosen considering the best BLEU score (top) and worst BLEU score (bottom), with respect to Visionizer Max Search. We underline the main differences between the captions.

For what concerns Fashion-Cap, in Figure 4 (top) we observe that Visionizer models recognize an additional characteristic of the item ("short sleeve") compared to the baseline model. Furthermore, the Visionizer model with Beam Search also correctly generates the term "cotton jersey", while its Max Search counterpart fails. In Figure 4 (bottom), we observe that all the models perform similar errors (e.g., missing the second part of the ground-truth caption). However, it is worth noticing, that Visionizer with Beam Search is able to correctly recognize the material of the item ("wool").

7 Conclusions

We have presented an extensive study concerning fashion image captioning. We have provided background and motivation for the definition of efficient genera-

tive models oriented to online application scenarios in the fashion domain. We have released a novel dataset for this task, and we have experimentally assessed its difficulty and compared it to two existing ones. To the best of our knowledge, this is the first study that investigates this problem by covering multiple datasets. Our experiments suggest our dataset can be tackled with popular architectures for image captioning, obtaining satisfactory results. Nonetheless, it can be considered challenging and leaves room for future improvements with more advanced techniques. In future work, we want to integrate semantic-based metrics such as BERTscore during the training, as part of the loss function. Moreover, the use of *professor forcing* [11] regularization instead of *teacher forcing* would reduce the discrepancy between the inputs received by the networks at training and test time, potentially leading to a performance improvement.

Acknowledgments

This work was partially supported by the European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-“FAIR” - Spoke 8.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
2. Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: universal image-text representation learning. In: ECCV. Lecture Notes in Computer Science, vol. 12375, pp. 104–120. Springer (2020)
3. Cheng, W.H., Song, S., Chen, C.Y., Hidayati, S.C., Liu, J.: Fashion meets computer vision: A survey. ACM Comput. Surv. **54**(4) (jul 2021)
4. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: ICML. pp. 1931–1942. PMLR (2021)
5. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: SSST@EMNLP. pp. 103–111 (2014)
6. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: A framework for generating controllable and grounded captions. In: CVPR. pp. 8307–8316 (2019)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (1). pp. 4171–4186 (2019)
8. Dong, J., Ma, Z., Mao, X., Yang, X., He, Y., Hong, R., Ji, S.: Fine-grained fashion similarity prediction by attribute-specific embedding learning. IEEE Transactions on Image Processing **30**, 8410–8425 (2021)
9. Dong, M., Zeng, X., Koehl, L., Zhang, J.: An interactive knowledge-based recommender system for fashion product design in the big data environment. Information Sciences **540**, 469–488 (2020)
10. Galassi, A., Lippi, M., Torroni, P.: Attention in natural language processing. IEEE Trans. Neural Networks Learn. Syst. **32**(10), 4291–4308 (2021)
11. Goyal, A., Lamb, A., Zhang, Y., Zhang, S., Courville, A.C., Bengio, Y.: Professor forcing: A new algorithm for training recurrent networks. In: NIPS. pp. 4601–4609 (2016)

12. Hacheme, G., Sayouti, N.: Neural fashion image captioning : Accounting for data diversity. CoRR **abs/2106.12154** (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (June 2016)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
15. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. CoRR **abs/2111.12233** (2021)
16. Liu, S., Liu, L., Yan, S.: Fashion analysis: Current techniques and future directions. IEEE MultiMedia **21**(2), 72–79 (2014)
17. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (June 2016)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
20. Popovic, M.: chrF: character n-gram f-score for automatic MT evaluation. In: WMT@EMNLP. pp. 392–395 (2015)
21. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: The generative fashion dataset and challenge. CoRR (2018)
22. Song, S., Mei, T.: When multimedia meets fashion. IEEE MultiMedia **25**(3), 102–108 (2018)
23. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 539–559 (2023)
24. Sumarliah, E., Usmanova, K., Mousa, K., Indriya, I.: E-commerce in the fashion business: the roles of the covid-19 situational factors, hedonic and utilitarian motives on consumers’ intention to purchase online. Int. J. of Fashion Design, Technology and Education **15**(2), 167–177 (2022)
25. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS. pp. 3104–3112 (2014)
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164 (2015)
28. Wiegrefe, S., Pinter, Y.: Attention is not not explanation. In: EMNLP/IJCNLP (1). pp. 11–20. Association for Computational Linguistics (2019)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. vol. 37, pp. 2048–2057 (2015)
30. Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C., Tan, J., Xie, D., Wang, J., Wang, X.: Fashion captioning: Towards generating accurate descriptions with semantic rewards. In: ECCV (13). vol. 12358, pp. 1–17 (2020)
31. Yin, R., Li, K., Lu, J., Zhang, G.: Enhancing fashion recommendation with visual compatibility relationship. In: WWW. p. 3434–3440. New York, NY, USA (2019)
32. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning based recommender system: A survey and new perspectives. ACM Comput. Surv. **52**(1), 5:1–5:38 (2019)
33. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: ICLR (2020)
34. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. In: ICCV. pp. 1689–1697 (2017)