Afro SpecDetect: A multimodal Transformer-based attributes retrieval system for African fashion images

Gilles Hacheme
Ai4Innov
Nairobi, Kenya
gilles.hacheme@ai4innov.com

Nouréini Sayouti
Ai4Innov
Paris, France
noureini.sayouti@ai4innov.com

Abstract

Despite advancements in Artificial Intelligence for fashion, African fashion remains underrepresented. This paper presents Afro SpecDetect, a dataset tailored for African fashion multimodal captioning, incorporating attributes like color, material, and fabric. Experiments demonstrate enhanced performance in Bleu and F1 scores when the items' typologies are provided as a context in addition to the image.

1 Introduction

With its vibrant patterns, diverse fabrics, and multifaceted designs, African fashion is a fundamental part of African culture and heritage, profoundly influencing global fashion trends. However, this rich aspect of African culture has been significantly underrepresented in technological research, particularly in image recognition and multimodal captioning.

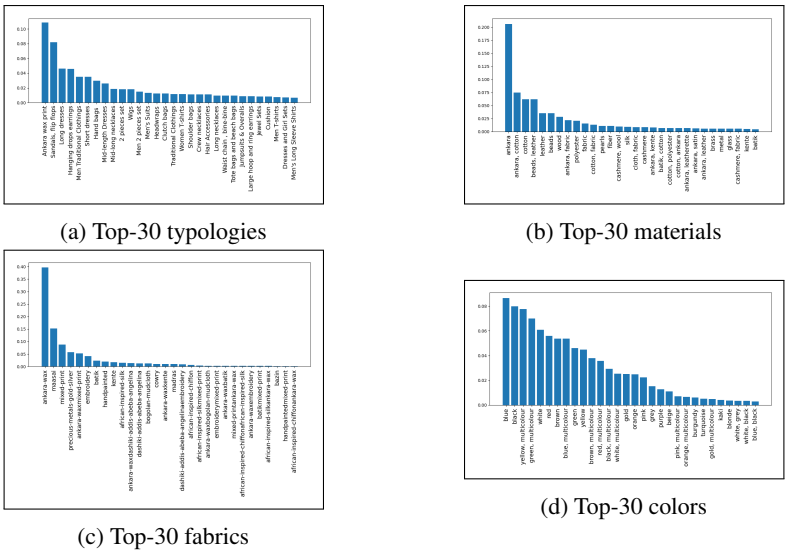
Existing datasets predominantly focusing on Western styles fail to capture the complexities and nuances of African fashion. Models trained on these datasets perform inadequately with African fashion images, not adequately reflecting their vibrant diversity. This limitation also extends to neglecting the multimodal aspect of fashion captioning, involving generating descriptive captions that consider visual cues and textual and contextual information.

Recognizing this gap, this paper introduces the Afro SpecDetect dataset, designed explicitly for African fashion and facilitating multimodal captioning. This is a crucial step towards advancing machine understanding of fashion images.

To understand their influence on image recognition models, the experiments conducted using this dataset explore various factors, such as model size, the inclusion of typologies, and different types of embeddings. The ultimate aim is to provide valuable insights to researchers and guide future work in African fashion image recognition and multimodal captioning.

Figure 1: The diversity of African fashion

5th Deep Learning Indaba Conference (DLI 2023).



(a) Top-30 typologies

(b) Top-30 materials

(c) Top-30 fabrics

(d) Top-30 colors

Figure 2: Summary of top-30 features

2 Dataset

Existing fashion datasets like DeepFashion [1] and Fashion-MNIST [2] primarily focus on Western styles. African fashion datasets are still largely under-represented. However, these recent years, an increased interest in building African-specific fashion datasets has been noticed. We can mention InFashAIv1, the first African-centered fashion AI dataset containing almost 16.000 African fashion item images with titles, prices, and general descriptions [3]. The second version of this dataset, InFashAIv2, contains 60,636 images with their generic descriptions. Another effort in that direction is the AFRIFASHION1600, a dataset that contains 1600 samples of contemporary African fashion images. These images are labeled into eight classes which represent different African fashion styles [4].

This paper introduces the *Afro SpecDetect* dataset, primarily designed to empower the building of multimodal captioning systems for African art and fashion items. The dataset comprises 100,000 items with typologies (broad categories), materials, fabrics, colors, and general descriptions. This unique dataset has been acquired through our partnership with Afrikrea. We are open to sharing this dataset for research purposes after a review by us and our partner. Some statistics about the attributes in the dataset can be found in figure 2.

In this paper, we showcase how items' typologies provided in this new dataset can be merged with images to generate better captions.

3 Model architecture

3.1 Images and text preprocessing

We implement a data preparation pipeline to combine images and text in the Seq2Seq Transformer model. We implement different processing steps, including horizontal flips, cropping, color jittering, and normalization. Additionally, we used the 'Img2Vec' module [1] that extracts the feature representations from the input images based on a pre-trained ResNet-18. We extract a given layer from the ResNet-18 and use the pixels, including all the channels, as features. For instance, as layer 4's output size is 7x7x512, we transform it into a sequence of 49x512 vectors. In other words, we have 49 512-dimensional vectors. In the Seq2Seq Transformer model, we added a linear layer before embedding the image into the model's hidden dimension. This is helpful as we experiment with different hidden dimensions for the captioning model.

[1] From https://github.com/christiansafka/img2vec

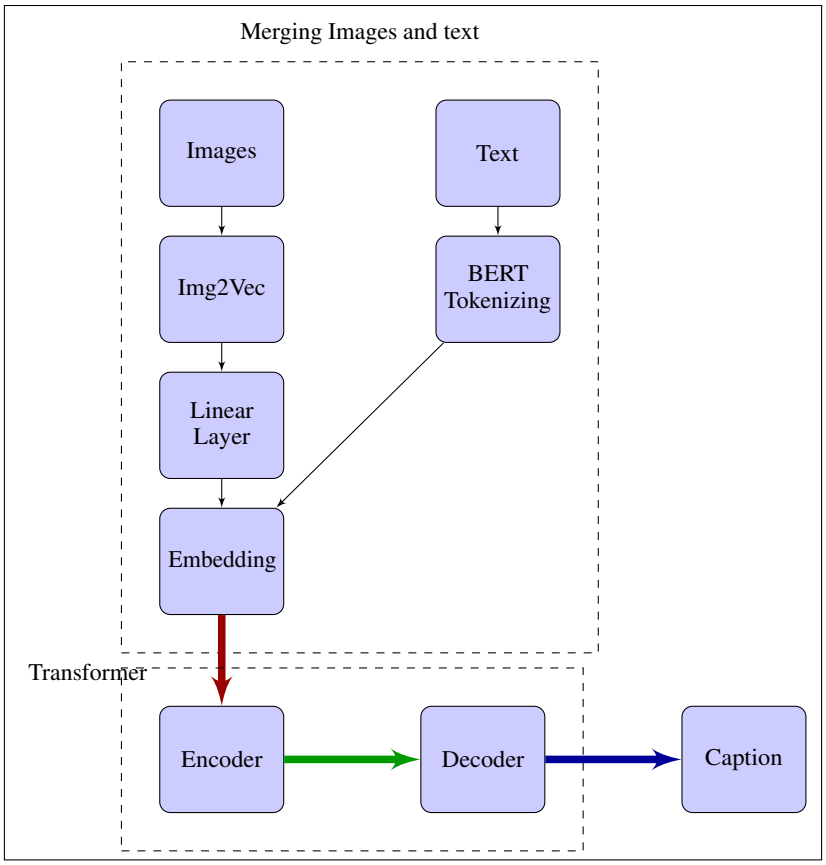


Figure 3: Description of the architecture

The typology[2] and the caption are tokenized using the Bert tokenizer and then padded to the specified maximum length. The maximum length varies according to the ResNet layer used to encode the images. When using layer 4, the maximum length is 100, and when using layer 3, it is 246.

3.2 Multimodal Seq2Seq Transformer model

The proposed model is a Seq2Seq Transformer model for image captioning, composed of an Encoder and a Decoder, and utilizes Multi-Head Self-Attention and Multi-Head Encoder-Decoder Attention mechanisms. The model incorporates the attention mechanisms to efficiently encode the input image and language sequences and generate high-quality image captions. This model allows us to test different hypotheses regarding multimodal image captioning.

Indeed, the Encoder takes an image vector and a source language sequence as input and generates an encoded image sequence using Multi-Head Self-Attention and Positionwise Feedforward Layers. The Decoder takes a target language sequence and the encoded image sequence as input and generates a sequence of tokens representing the image caption using Multi-Head Self-Attention and Multi-Head Encoder-Decoder Attention Layers. The output of the Decoder is passed through a linear layer to generate the output sequence.

4 Experiments

Our experiments investigate two key hypotheses. We first hypothesize that providing the model with the item's typology in addition to the image should help the model disambiguate between different items on the same picture and focus more on the item of interest. This is expected to make the

[2] A category representing the most important object in an image

captioning more accurate and contextually more relevant. We conducted experiments comparing the model's performance with and without typology to validate our approach.

The second hypothesis assesses the effect of data annotation on the model's performance. Indeed, some items in our dataset have missing or incorrect attributes. As a result, we annotate half of the data to correct these discrepancies and improve the overall quality of the training set. We then use the labeled set to train an initial model to impute the missing attributes. And we evaluate the model's performance both with and without the annotation (weak labeling) strategy. When the annotation strategy is not used, only the items without missing attributes are used to train the captioning model. The evaluation metrics include the BLEU scores to assess the overall captioning quality and attribute-specific F1 scores for colors, materials, and fabrics.

4.1 BLEU Scores for all attributes combined

The models employing typology consistently outperformed those lacking typological considerations. This holds across both categories— with and without annotations— for every model size (tiny, small, medium) and both embeddings (Scratch and BERT). The models trained with Scratch embeddings tended to outperform those using BERT embeddings. Interestingly, the performance disparity between the 'with typology' and 'without typology' models was less pronounced when annotations were present. Furthermore, the BLEU scores indicate that model complexity, as suggested by size, contributes to performance improvement.

Table 1: Bleu scores for all the attributes combined

Bleu score for all		**Without Annotations**				**With Annotations**			
		With Typology		**Without Typology**		**With Typology**		**Without Typology**	
Model size	**Embedding**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**
Tiny	Scratch	0.204	0.209	0.100	0.100	0.203	0.207	0.086	0.088
	BERT	0.161	0.185	0.052	0.060	0.174	0.202	0.052	0.069
Small	Scratch	0.209	0.223	0.087	0.108	0.210	0.224	0.101	0.092
	BERT	0.192	0.188	0.108	0.064	0.176	0.204	0.071	0.082
Medium	Scratch	0.215	0.226	0.105	0.117	0.221	0.226	0.104	0.103
	BERT	0.179	0.186	0.091	0.065	0.172	0.210	0.064	0.095

4.2 F1 Scores for Specific Attributes

Color Attribute: Typological consideration proves critical for performance in the Color attribute as well, with models 'with typology' consistently scoring higher than those 'without typology.' As seen in the BLEU score evaluation, the Scratch embedding performs superiorly to the BERT embedding across all model sizes and conditions. In terms of model size, larger models generally yield better results. However, the usage of annotations does not consistently enhance the F1 scores for this attribute.

Table 2: F1 scores for Color attribute

		Without Annotations				**With Annotations**			
		With Typology		**Without Typology**		**With Typology**		**Without Typology**	
Model size	**Embedding**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**
Tiny	Scratch	0.270	0.282	0.196	0.208	0.263	0.279	0.165	0.195
	BERT	0.186	0.214	0.116	0.103	0.185	0.266	0.059	0.168
Small	Scratch	0.288	0.306	0.178	0.223	0.268	0.311	0.200	0.218
	BERT	0.240	0.232	0.191	0.146	0.196	0.272	0.107	0.176
Medium	Scratch	0.288	0.313	0.208	0.229	0.299	0.317	0.207	0.216
	BERT	0.229	0.226	0.136	0.103	0.188	0.295	0.046	0.182

Material Attribute: The trend of the 'with typology' models performing better continues for the Material attribute. Especially noteworthy is the performance of the Scratch embedding over BERT in the 'without annotations' category, which narrows when annotations are included. In line with the previous results, model size continues to have a positive correlation with performance.

Table 3: F1 scores for the Material attributes

		Without Annotations				With Annotations			
		With Typology		Without Typology		With Typology		Without Typology	
Model size	**Embedding**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**
Tiny	Scratch	0.340	0.346	0.152	0.150	0.348	0.348	0.150	0.157
	BERT	0.251	0.310	0.021	0.104	0.298	0.333	0.057	0.091
Small	Scratch	0.354	0.381	0.147	0.192	0.370	0.376	0.142	0.168
	BERT	0.320	0.305	0.153	0.086	0.288	0.338	0.103	0.060
Medium	Scratch	0.373	0.388	0.179	0.193	0.368	0.388	0.156	0.188
	BERT	0.275	0.307	0.115	0.101	0.286	0.351	0.076	0.108

Fabric Attribute: Finally, for the Fabric attribute, 'with typology' models consistently outperform their 'without typology' counterparts, underscoring the importance of typology. The Scratch embedding shows stronger performance than BERT across most conditions. As with the other attributes, larger models perform better. However, the effectiveness of annotations varies significantly, appearing to contribute positively in the 'with typology' scenario while less so in the 'without typology' one.

Table 4: F1 scores for the Fabric attribute

		Without Annotations				With Annotations			
		With Typology		Without Typology		With Typology		Without Typology	
Model size	**Embedding**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**	**Layer 4**	**Layer 3**
Tiny	Scratch	0.456	0.467	0.343	0.367	0.464	0.476	0.342	0.375
	BERT	0.399	0.432	0.270	0.240	0.426	0.444	0.250	0.222
Small	Scratch	0.466	0.476	0.322	0.382	0.475	0.491	0.328	0.364
	BERT	0.433	0.436	0.345	0.225	0.406	0.456	0.250	0.246
Medium	Scratch	0.490	0.486	0.367	0.386	0.480	0.491	0.355	0.377
	BERT	0.418	0.442	0.318	0.240	0.410	0.472	0.194	0.254

These results underline typology's value in enhancing models' performance across different attributes and metrics. Larger models yield better results, and the Scratch embedding consistently delivers superior performance over BERT. However, the impact of annotations appears to be variable, warranting further investigation for specific attribute optimization.

5 Applications, Future Work, and Conclusion

The Afro SpecDetect dataset uniquely focuses on African fashion recognition and multimodal captioning. It presents opportunities for practical applications like enhancing recommendation systems in e-commerce for African fashion items, contributing to more culturally representative AI, and aiding in fashion design and trend analysis.

Future research could expand the dataset to include more diverse data types like tactile information or 3D modeling. Exploring different model architectures, training techniques, and transfer learning opportunities can also lead to further improvements. There is potential for including broader African fashion styles and subcultures to enhance generalizability and representativeness.

Our experimentation showed that Afro SpecDetect aids models in recognizing key attributes of African fashion, improving performance in both Bleu and F1 scores. Incorporating typology and annotations and employing various models and architectures, including different embedding types and sizes, enhanced captioning performance. Notably, even smaller models delivered satisfactory performance, highlighting implications for resource-efficient applications.

In conclusion, Afro SpecDetect fills a crucial gap in African fashion recognition and multimodal captioning. It represents a significant contribution, paving the way for more culturally diverse AI applications and new opportunities in e-commerce, design, and trend analysis. Further improvements and expansions to the dataset will continue to advance the field, but as it stands, Afro SpecDetect provides a strong foundation for future research.

References

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 06 2016.

[2] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[3] Gilles Hacheme and Noureini Sayouti. Neural fashion image captioning: Accounting for data diversity. *arXiv preprint arXiv:2106.12154*, 2021.

[4] Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Sharon Ibejih, Opeyemi Osakuade, Ifeoma Okoh, and Mary Salami. Afrifashion1600: a contemporary african fashion dataset for computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3968–3972, 2021.