International Scientific Journal of Engineering and Management
Volume: 02 Issue: 05 | May – 2023 DOI: 10.55041/ISJEM00798
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

METHOD FOR GENERATING DESCRIPTIONS FOR CLOTHING IMAGESTO SUPPORT INTELLIGENT SYSTEM

S Sharmila Shetty
Department of Information Science & Engineering
AJ Institute of Engineering and Technology
Mangalore, Karnataka, India
shamilasshetty27@gmail.com

Ashwija Shetty
Department of Information Science & Engineering
AJ Institute of Engineering and Technology
Mangalore, Karnataka, India
ashwijashetty28@gmail.com

Deeksha Poojary
Department of Information Science & Engineering
AJ Institute of Engineering and Technology
Mangalore, Karnataka, India
poojaryd651@gmail.com

Karthik Kille
Department of Information Science & Engineering
AJ Institute of Engineering and Technology
Mangalore, Karnataka, India
karthikkille2@gmail.com

Mrs. Sivapuram Jayasri
(Project Guide) Assistant Professor,
Department of Information Science and EngineeringA J Institute of Engineering and Technology Mangalore, Karnataka, India
Jayasree@ajiet.edu.in

Abstract— Image captioning aims to generate descriptions of images with natural language sentences automatically. Most methods tackle this problem in an end-to-end fashion in recent years, which generates captions directly from image-level features but ignores high-level semantic information. The method that introduced the attribute concept into the CNN-RNN framework made a considerable improvement while the performance depended on the manually selected attributes heavily. Image captioning, i.e., automatically generating natural language image descriptions, is useful for the visually impaired, and for natural language-based image search. Image captioning has increasingly large domains of application, and fashion is not an exception. Having automatic item descriptions is of great interest for fashion web platforms hosting sometimes hundreds of thousands of images. This system is one of the first to tackle image captioning for fashion images. To contribute to addressing dataset diversity issues, we introduced the InFashAIv1 dataset containing almost 16.000 African fashion item images with their titles, prices, and general descriptions.

I INTRODUCTION

Image captioning is the process aiming to associate a text description with an image in an automatic manner. In this process, a computer is trained to understand the visual content of an image and produce a corresponding descriptive sentence, the caption. This process combines two sub-fields of computer science namely: Natural Language Processing (NLP) and Computer Vision (CV). Automatic caption generation from a given image could have several use of cases Recommended the editing applications, virtual assistants, image indexing, and assisting visually impaired persons to understand the content of an image. Image captioning is a challenging task and it recently drew lots of attention from researchers in CV. Image captioning task which requires machines to have the ability of visual understanding like human beings is a significant component in cross-media semantic analysis. Image captioning, i.e. automatically generating natural language image descriptions, is useful for the visually impaired, and for natural language-based image search.

The main idea of these image captioning models is to encode the given images as vector representations with a convolutional neural network and then use a long short-term memory network as a decoder to generate the text descriptions of the images. The encoder-decoder-based methods have dominated the study of image captioning since these methods can be trained end-to-end and scale to very large training data. Image captioning is the task of providing a natural language description of the content in an image. It lies at the intersection of computer vision and Natural Language Processing. Automatic image captioning is useful to many applications, such as developing image search engines and helping visually impaired people to understand their surroundings. Hence, image captioning has been an active research area. The advent of new convolutional neural networks and object detection architectures have contributed enormously to improving image captioning. With the advancements in deep neural network models, automatic image captioning has become a promising research area. The image captioning research connects both the natural language processing (NLP) and computer vision (CV) communities. The main idea is to enhance machine intelligence by developing multi-modal systems with machine learning techniques. These multi-modal systems should capture the shared semantics between images and text descriptions by jointly learning the multi-modal structure of the image and text data.

ISSN: 2583-6129
DOI: 10.55041/ISJEM00798
www.isjem.com

II PROBLEM STATEMENT

Image captioning is the technique of automatically associating a text description to a picture. A computer is trained in this method to interpret the visual information of an image and produce a corresponding descriptive language, the caption. Automatic caption generation from a given image could have a variety of applications, including recommendations in editing apps, virtual assistants, image indexing, and assisting visually impaired people to grasp the content of an image. This suggested system is all about constructing a deep learning model that recognizes the features of a dress image such as dress material, dress type, the color of the dress, and whether it's for men or women by analyzing the input image using deep learning models that are trained using one of the famous dataset called fashionAIv1.

III OBJECTIVE

The Objectives comprises of the following:

- To develop a system that processes fashion images and classifies them into different categories using a multi-class classification technique.
- To design a system that identifies the dress type, dress material, and dress material using machine learning techniques.
- To apply the concept of machine learning in the fashion industry and to develop an intelligent system that identifies dress-relatedattributes.
- To design a system with well-defined user interfaces for the end users.

IV EXPECTED OUTCOMES

- Proposed system should provide fashion image information such as color, dress type, material, and gender from the uploadedimage.
- Proposed system should prepare image descriptions in sentence form.
- Proposed system should convert textual image descriptions into voice data using text-to- voice conversion techniques.
- System should contain well-defined user interfaces.

V LITERATURE REVIEW

Lun Huang, Wenmin Wang, et. ed [1] in "Image Captioning with Two Cascaded Agents" proposed a pipelined image captioning framework consisting of two cascaded agents. The former is named a "semantic adaptive agent" which generates the input to the decoder by consulting the information from the current decoding process, and the latter as "caption generating agent" which selects a single word of the vocabulary as the output of the decoder by taking into consideration of the input and the current states of the decoder. For the framework of two cascaded agents, they design a multi-stage training procedure to train the two agents with different objectives by fully utilizing reinforcement learning.

Marco Pedersoli et. ed [2] in "Areas of Attention for Image Captioning" propose "Areas of Attention", proposed a novel attention-based model for automatic image captioning. Our approach models the dependencies between image regions, caption words, and the state of an RNN language model, using three pairwise interactions. In contrast to previous attention-based approaches that associate image regions only to the RNN state, our method allows a direct association between caption words and image regions. During training these associations are inferred from image-level captions, akin to weakly-supervised object detector training. These associations help to improve captioning by localizing the corresponding regions during testing.

Feng Chen, Songxian Xie et. ed [3] in "What Topics Do Images Say: A Neural Image Captioning Model With Topic Representation" proposed a topic-guided neural image captioning model which incorporates a topic model into the CNN-RNN framework. Our model represents each image as a set of topics and each topic as various words with relevant distributions. They conduct experiments on the Microsoft COCO dataset. The results show that our model outperforms the baselines.

Yingwei Pan et. ed [4] in "X-Linear Attention Networks for Image Captioning" introduced a unified attention block — X-Linear attention block, that fully employs bilinear pooling to selectively capitalize on visual information or perform multimodal reasoning. Technically, the X-Linear attention block simultaneously exploits both the spatial and channel-wise bilinear attention distributions to capture the 2nd order interactions between the input single-modal or multi-modal features. Higher and even infinity order feature interactions are readily modeled through stacking multiple X-Linear attention blocks and equipping the block with Exponential Linear Unit (ELU) in a parameter-free fashion.

Wei Zhang et. ed [5] in "Reconstruct and Represent Video Contents for Captioning via Reinforcement Learning" addressed the problem of describing the visual contents of a video sequence with natural language. Unlike previous video captioning work mainly exploiting the cues of video contents to make a language description, we propose a reconstruction network (RecNet) in a novel encoder-decoder-reconstructor architecture, which leverages both forward (video to sentence) and backward (sentence to video) flows for video captioning. Specifically, the encoder-decoder component makes use of the forward flow to produce a sentence description based on the encoded video semantic features. Two types of reconstructors are subsequently proposed to employ the backward flow and reproduce the video features from local and global perspectives, respectively, capitalizing on the hidden state sequence generated by the decoder.

Niange Yu et. ed [6] in "Topic-Oriented Image Captioning Based on Order-Embedding" presented an image The captioning framework that generates captions under a given topic. The topic candidates are extracted from the caption corpus. A given image's topics are then selected from these candidates by a CNN-based multilabel classifier. The input to the caption generation model is an image-topic pair, and

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 05 | May – 2023 DOI: 10.55041/ISJEM00798
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
www.isjem.com

the output is a caption of the image. For this purpose, a cross- modal embedding method is learned for the images, topics, and captions. In the proposed framework, the topic, caption, and image are organized in a hierarchical structure which is preserved in the embedding space by using the order embedding method. The caption embedding is upper bounded by the corresponding image embedding and lower bounded by the topic embedding. The lower bound pushes the images and captions about the same topic closer together in the embedding space.

Xueting Zhang et. ed [7] in "Multi-Scale Cropping Mechanism for Remote Sensing Image Captioning" proposed a training mechanism of multi-scale cropping for remote sensing image captioning in this paper, which can extract more fine-grained information from remote sensing images and enhance the generalization performance of the base model.

Min Yang et. ed [8] in "Multitask Learning for Cross-domain Image Captioning" introduced "MLADIC", a novel Multitask Learning Algorithm for cross-Domain Image Captioning. MLADIC is a multitask system that simultaneously optimizes two coupled objectives via a dual learning mechanism: image captioning and text-to-image synthesis, with the hope that by leveraging the correlation of the two dual tasks, we are able to enhance the image captioning performance in the target domain. Concretely, the image captioning task is trained with an encoder-decoder model (i.e., CNN-LSTM) to generate textual descriptions of the input images. The image synthesis task employs the conditional generative adversarial network (CGAN) to synthesize plausible images based on text descriptions.

Md. Zakir Hossain et. ed [9] in "Text to Image Synthesis for Improved Image Captioning" proposed an image captioning method that uses both real and synthetic data for training and testing the model. We use a Generative Adversarial Network (GAN) based text to image generator to generate synthetic images.We use an attention- based image captioning method trained on both real and synthetic images to generate the captions. We demonstrate the results of our models using both qualitative and quantitative analysis on popularly used evaluation metrics.

Seung-Ho et. ed [10] in "Domain-Specific Image Caption Generator with Semantic Ontology" proposes a domain-specific image caption generator, which generates a caption based on attention mechanism with object and attribute information, and reconstruct a generate caption using a semantic ontology to provide natural language description for given specific-domain.

VI METHODOLOGY

Deep learning algorithms and machine learning algorithms are two developing methodologies that have recently piqued the interest of many researchers. Deep learning approaches have also had a lot of success in computer vision. These strategies provide users with a consistent feature extraction classification framework, freeing them from time-consuming handmade feature extraction. Deep learning algorithms also have the potential to improve the accuracy of disease identification. Deep learning techniques, specifically convolutional neural networks and recurrent neural network, are used in the proposed system to propose a model to predict fashion image attributes such as colour, material, type, and gender from the input fashion image. The system use the fashinAIv1 dataset to train and classify the deep learning model.

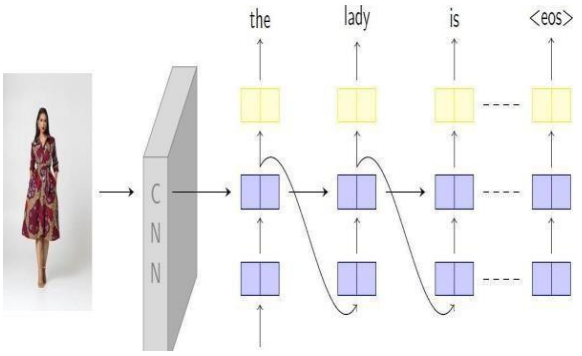

Figure 1.Architecture

The graphic above depicts the proposed system's system architecture. Modules for pre-processing, feature extraction, model construction, and training are included in the system. To forecast fashion-related properties such as gender, colour, material, and type, the system requires a fashion image dataset. To forecast fashion qualities, the system requires a fashion image collection. The system constructs and trains an image model using a Convolution Neural Network, and it classifies attributes using an SVM model.

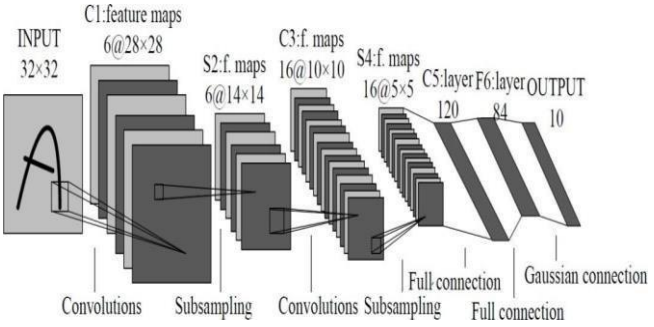

Figure2 Convolution Neural Network

Convolutional neural networks are a type of deep learning algorithm. It is basically a multi-layer perceptron that replicates local perception in order to perform input-to-output mapping. It extracts data characteristics at various scales using multiple convolutions and pooling. The CNN network is distinguished by the method employed in local connections and shared weights. On the one hand, it minimises the number of weights, making the network easier to optimize, while also lowering the risk of overfitting. CNNs are typically made up of three mutually supported layers: a convolutional layer, a pooling layer, a fully connected layer, and a Softmax layer. Local features are obtained during the convolution process. Because one of the convolution layers is made up of multiple convolution units.

ISSN: 2583-6129
DOI: 10.55041/ISJEM00798
www.isjem.com

VII CONCLUSION

We have decided to develop an image captioning technique to extract visual information from an input fashion image. The proposed system uses deep learning techniques, namely - convolutional neural network and recurrent neural network are utilized to propose a model to predict fashion image attributes such as color, material, type and gender from the input fashion image. The system uses fashinAIv1 dataset to train the deep learning model and for classification.

VIII REFERENCES

[1] Yezhou Yang, Ching Lik Teo, Hal Daume III, and Yiannis ´ Aloimonos, "Corpus-guided sentence generation of natural images," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 444–454.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015.

[3] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C.Berg, and Tamara L. Berg, "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 12, pp. 2891– 2903, 2013

[4] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching imageand sentence," in ICCV, 2015, pp. 2623–2631.

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In ECCV, 2016.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[7] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–13,2018.

[8] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B.Plank, "Automatic description generation from images: A survey of models, datasets, and evaluationmeasures," Journal of Artificial Intelligence Research, vol. 55, pp. 409– 442, 2016.

[9] L. White, R. Togneri, W. Liu, and M. Bennamoun, Neural Representations of Natural Language,vol. 783. Singapore: Springer, 2018.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottomup and top-down attention for image captioning and visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–608

