From Pixels to Phrases: Enhancing Image Captioning with LSTM Model

Pulkit Dwivedi
Apex Institute of Technology (CSE) Chandigarh University, India

***Abstract*—Generating natural language captions for images is an important task that requires understanding and identifying the objects within an image. However, the effectiveness of image caption generation has not been thoroughly proven. To address this gap, we propose a novel approach that com- bines Convolutional Neural Networks (CNNs) and Long Short- Term Memory (LSTM) models to generate image captions. Our approach comprises two sub-models: an Object Identification model and a Localization model that extract information about objects and their spatial relationships from images. We then use LSTM models to process the extracted text data, encoding the text input sequence as a fixed-length output vector. Finally, we integrate the image vector outputs and the corresponding descriptions to train the image caption generator model. We compare the performance of our LSTM-based model with other dense models, including VGG-16 and Transformer-based models, using the Flickr8k dataset. Our experimental results demonstrate that our LSTM-based approach outperforms previous VGG and Transformer-based models, as well as state-of-the-art image captioning models. By integrating image and text data using LSTM models, our approach provides a new benchmark for image caption generation, advancing the state-of-the-art in this critical area of research.**

***Index Terms*—Image captioning, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Natural language processing (NLP), Deep learning**

I. INTRODUCTION

Image captioning is an area of research that combines computer vision and natural language processing to create a textual description of the contents of an image. The progress in this field has been driven by advances in both computer vision, such as improved convolutional neural networks and object recognition models, and natural language processing, such as attention-based recurrent neural networks that improve caption accuracy.

The challenge of image captioning lies in not only identifying the objects in the image but also their relationships, attributes, and activities, and expressing this information in natural language. Current image captioning systems mainly focus on improving the alignment between visual and linguistic information by gathering fine-grained semantics and using visual attention to enhance this interaction.

However, the semantic comprehension capacity of pre-trained detectors/classifiers is limited by pre-defined semantic/class labels, and the tuning of these models for visually salient semantics in the output sentence is challenging.

Recently, neural networks have become the most popular and successful techniques in computer vision and natural language processing due to easy access to vast amounts of data and parallel processing resources. These networks primarily consist of convolutional and recurrent layers, which were the best performing tools in many tasks in the early stages of neural network development. Language modeling has also advanced, with RNN/Transformer-based sentence decoders used to achieve linguistic coherence in the output sentence.

Overall, image captioning is an exciting field of research that continues to advance through the development of more sophisticated neural network models and the integration of computer vision and natural language processing techniques.

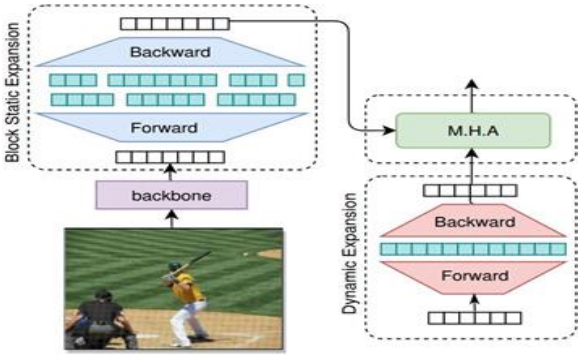

Fig. 1. The expansion mechanism generates the input data into a new one with a variable sequence length during the forward phase and reverses the operation during the backward pass to enable the network to process the input without being confined by the number of components. The Block Static Expansion enables you to perform these operations on a collection of arbitrary and different lengths all at once.

We propose that a combined architecture for computer vision and natural language processing can benefit both fields by enabling the joint modeling of visual and textual signals and facilitating the sharing of knowledge between domains. We expect that the success of the VGG model in various vision tasks will support this idea and strengthen the case for unified modeling of vision and language signals.

The primary contributions of our study are as follows:

1) We introduce a multi-layered deep Convolutional Neural Network (CNN) that detects and extracts distinctive features that are useful for generating image captions.
2) Our proposed model is fine-tuned with existing image captioning models that are based on the VGG architecture.
3) We evaluate and compare our results with state-of-the- art models using BLEU scores.
4) To identify the salient and unique features of objects in images, such as humans, animals, or inanimate objects, our model includes an effective object detection component.

II. LITERATURE REVIEW

Automatically characterising the content of a picture is a fundamental challenge in artificial intelligence that mixes computer vision and natural language processing. In [3,] a generative model based on a deep recurrent architecture is described, which merges recent advances in computer vision and machine translation and may be used to generate meaningful sentences describing an image in this study. The model is trained using the training image to maximise the chance of the target description sentence. Experiments on diverse data sets show that the machine learns the language and its correctness exclusively via visual descriptions. In contrast, recent research has shown that policy-gradient reinforcement learning methods may be used to train deep end-to-end systems directly on non- differentiable task metrics. In [4], Steven J. Rennie et. al. In their research, they investigated the difficulty of enhancing image captioning systems using reinforcement learning, and shown that significant increases in performance may be obtained by carefully tweaking systems using the test metrics of the MSCOCO tasks. They present a system that is built using an unique optimization approach called as self-critical sequence training (SCST). SCST is an implementation of the well-known REINFORCE algorithm. The MSCOCO evaluation results set a new standard for the challenge, increasing the best CIDEr measure with SCST and greedy decoding result from 104.9 to 114.7. Meanwhile, newer models often rely on a pre-trained object detector/classifier to extract semantics from images, leaving untouched the natural language ordering of meanings. In [6], a new Transformer-style architecture called Comprehending and Ordering Semantics Networks (COS-Net) is offered, which integrates an enhanced semantic comprehending and a learnable semantic ordering process into a single design. Empirical evidences reveal that COS-Net clearly outperforms state-of-the-art COCO methods and gets the best CIDEr score of 141.1% on the Karpathy test split to date. However, the majority of previous research has primarily focused on pre-training transformers of moderate sizes (e.g., 12 or 24 layers) on around 4 million photos. In [7], Xiaowei Hu et. al. demonstrated LEMON, a LargE-scale iMage captiONer, and conducted the first empirical investigation on the scaling behaviour of VLP for image captioning in Image captioning. Existing methods try to accomplish Image captioning task by encoding OCR tokens with rich visual and semantic representations. However, with such limited representations, it is possible that strong correlations between OCR tokens will not be established. In [8], Jing et.al. proposed to use the geometrical relationship to increase the connections between OCR tokens They carefully evaluated the height, breadth, distance, IoU, and orientation relationships of the OCR tokens while building the geometrical connection. This research introduced a Long Short-Term Memory + Relation-aware pointer network (LSTM-R) architecture to combine the learnt relation as well as the visual and semantic representations into a single framework. Object detection networks' extracted descriptive region features have played an important role in recent ad- vances in image captioning. However, they are still criticised for a lack of contextual information and fine-grained details, which are traditional grid features' strengths. Presented a novel Dual-Level Collaborative Transformer (DLCT) network in [9], which helped to realise the complementary advantages of the two features. Since texts are ubiquitous in daily life, text- based image captioning (TextCap) [10], which aims to read and reason images with texts, is critical for a machine to understand a detailed and complex scene environment. The task, however, was extremely difficult because an image frequently contains complex texts and visual information that is difficult to

describe in detail. Existing methods attempt to solve this task by extending traditional image captioning methods, which focus on describing the overall scene of images with a single global caption. This was impossible because the complex text and visual information cannot be adequately described in a single caption. To address this issue, Xu et. al. [10] aimed to generate multiple captions that accurately describe various parts of an image in detail. There are three major challenges in achieving the goal: 1) it is difficult to decide which parts of image texts to copy or paraphrase; 2) it is difficult to capture the complex relationship between diverse texts in an image; and 3) it is still unclear how to generate multiple captions with diverse content. To overcome these challenges, a novel idea was proposed in [10] named as Anchor-Captioner method. Specifically, they looked for important tokens that would be given more attention and used them as anchors. Then, for each chosen anchor, they grouped its relevant texts to create the anchor-centered graph (ACG). Finally, they performed multi-view caption generation based on different ACGs to improve the content diversity of generated captions. However, current state-of-the-art image captioning models use auto- regressive decoders, which means that each word is generated by conditioning on previously generated words, resulting in significant latency during inference. To address this issue, non auto regressive image captioning models that generate all words in parallel have recently been proposed to significantly accelerate the speed of inference [11]. However, because they removed words dependence excessively, these non auto- regressive models invariably suffer from significant generation quality degradation. To improve the trade-off between speed and quality, Zhou et. al. presented SATIC, a semi-auto- regressive model for image captioning that retains the auto- regressive property in global but generates words in parallel in local [11]. SATIC can be implemented with only a few minor changes to Transformer. In [12][14], the research on attentive deep learning models for image captioning is done. Rather than providing a complete overview of all existing work on deep image captioning models, they explained distinct forms of attention mechanisms utilised in deep learning models for picture captioning. Although there are variances in how these models exploit attention mechanisms, the most successful deep learning models used for image captioning follow the encoder- decoder architecture [13][18][19]. They uncovered the most successful sorts of attention mechanisms in deep models for image captioning by analysing performance outcomes from different attentive deep models for image captioning.

III. PROPOSED METHODOLOGY

A. Model Architecture

Our image captioning model is based on the CNN + RNN approach, which has seen recent success in image/video captioning, particularly with region-level attention mechanisms. We chose to use the VGG-16 CNN model, which was pro- posed by Kiren Simoyan and Andrew Zisserman in 2014 [24] for large-scale image recognition. VGG-16 performed exceptionally well in object localization and image classification tasks in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), achieving a top-5 tested accuracy of 92.7%. We attempted to improve upon the VGG-16 model by tuning and reworking it. Our dataset was the Flickr8k dataset. Our proposed model uses an R-CNN + RNN approach, where the input image is interpreted as a series of image regions using R-CNN. These regions are uniquely encoded into relation-aware features, conditioned on semantic/spatial graphs, and then decoded into target output words using an attention LSTM decoder. We employed a top-down approach for generating image captions, where the model first processes the input image and generates a vector representing the classification probability for each class corresponding to the image. The softmax function is then used to determine the most accurate class for the image based on the probabilities calculated earlier.

The VGG architecture takes in images in the form of a tensor with image dimensions as inputs. The first two layers consist of 64 tracks with the same buffer and a 33 filter size. This is followed by a max pool layer with a stride of (2, 2), and two layers with convolution layers of 128 filter size and a filter size of (3, 3). Another max-pooling layer with a stride of (2, 2) follows. Two convolution layers with filter sizes of (3, 3) and 256 filters are then added. Two sets of three convolution layers and a max-pooling layer follow, each with 512 filters of the same size (3, 3) and the same buffer. Finally, the image is fed into a stack of two convolution layers. The filters used in the convolution and max-pooling layers are 3*3 in size, and 1*1 pixels are used in some layers to adjust the number of input channels. To prevent the spatial characteristics of the image from being lost, 1-pixel padding (identical padding) is applied after each convolution layer.

B. Model Configuration

The first step in generating a semantic image description is to extract feature vectors with a fixed length that represent

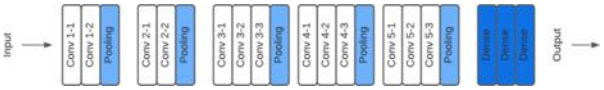

Fig. 2. VGG Architecture

the visual characteristics of the image. This allows the use of recurrent neural networks (RNNs) to focus on capturing linguistic qualities while combining them with visual information in the final output. Word embedding is a technique used to generate compact vector representations of textual data. The VGG architecture includes various models, one of which is the VGG 16, proposed by Kiren Simoyan and Andrew Zisserman in 2014 for large-scale image recognition. The VGG 16 has been successful in object localization and image classification tasks, achieving a top-5 tested accuracy of 92.7% in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The VGG 16 model is available in two versions, with the primary difference being that one uses (1, 1) filter size convolution layers while the other uses (3, 3) filter size convolution layers. These two models have 134 million and 138 million characteristics, respectively, and are otherwise similar except for some convolution layers.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Fig. 3. VGG Configuration

In order to perform object localization, we need to replace the class score with the bounding box position coordinates. The bounding box location is represented by a 4-dimensional vector containing the central coordinates (x,y), height, and width. To generate a dense vector from input text, an empirically integrated sequence is used, with test results reported for the numbers 16, 35, and 40. The dense vector is then fed into the RNN layer of an LSTM network with 256 memory units, which encrypts linguistic properties. The LSTM output is also a 256-element vector. Finally, a dense layer of the same length is applied to the output vectors, and a softmax prediction is made for the next word.

The image captioning architecture used in this study employs a deep neural layer to merge and handle both picture feature vectors and their associated text descriptions. A feature extractor and a textual processor are used to provide a 256- element vector, which is then merged and processed by a dense 256-neuron layer. A dense final output layer anticipates the entire output vocabulary for the following phrase in the sequence. The model can be trained in 256 batches over 20 epochs, using the categorical cross-entropy loss function and the Adam optimizer for improved convergence. The proposed CNN-5 model includes five layers and takes the image as input.

Consider a labelled directional graph, formally $G = (V, E) \in \{Gsem, Gspa\}$ where V is the collection of all detected region vertices, and E is a collection of visual relationship edges. Separate transformation matrices and bias vectors are used for distinct edge directions and labels, aiming to make the modified GCN responsive to both directionality and labels. Accordingly, each vertex vi is encoded via the modified GCN as:

effectiveness, we used the BLEU score, which is a metric that measures the similarity between machine-generated descriptions and human reference translations or descriptions. The BLEU-N score, where N ranges from 1 to 4, was computed by comparing each machine-generated description with a set of high-quality reference phrases, and the average score over the test dataset was used to determine the correctness of the generated descriptions. A score closer to 1 indicates a higher similarity between the generated text and the reference text, whereas a score closer to 0 suggests that the machine's description does not match the reference text. The BLEU score is formally defined as the geometric average of the modified n-gram precision values, and the model is deemed effective if the BLEU scores on the test dataset fall within the specified range of [0, 1]. The range is specified as [21]:

BLEU-1 - [0.401, 0.578]

BLEU-2 - [0.176, 0.390]

BLEU-3 - [0.099, 0.260]
BLEU-4 - [0.059, 0.170]

$$v_i^{(l)} = \rho\Big(\sum_{v_j \in N(v_i)} W_{dir(v_i,v_j)}\, v_j + b_{lab(v_i,v_j)}\Big) \quad (1)$$

B. Comparision of BLEU scores with other models

where $dir(v_i, v_j)$ selects the transformation matrix with re- gard to the directionality of each edge (i.e., W_1 for $v_i - to - v_j$, W_2 for $v_j - to - v_i$, and W_3 for $v_i - to - v_i$). $lab(v_i, v_j)$ represents the label of each edge [25].

The preset length text description is first merged into the input text in this paper before being transformed into a dense vector. This dense vector is used to feed 256 memory devices into the LSTM network. Finally, the LSTM output would be a vector of length 256. As a result, LSTM translates text from the input to a 256-length vector. This is then combined into an image vector of the same length and sent to the deep neural layer for model training.

C. Dataset Description

The Flickr8k dataset was used in this study, which consists of a pre-defined training dataset with 8,000 images, a development dataset with 1,000 images, and a test dataset with 1,000 images. This dataset is a new benchmark collection of 8,000 photos with five distinct captions, which provide clear descriptions of the important objects and events in the photos. The photos were carefully chosen from six different Flickr8k groups and do not typically feature famous individuals or landmarks. Rather, they were selected to represent a diverse range of scenes and situations, including candid shots of people, animals, and objects engaged in various activities. Each image in the dataset measures 28 pixels in height and 28 pixels in width, resulting in a total of 784 pixels per image.

IV. RESULTS

A. Evaluation of the Model

In our study, we compared the performance of our proposed model in image captioning with other state-of-the-art transfer learning approaches like VGG. To evaluate the model's

Table I displays the BLEU-N values derived from the proposed Convolutional and Recurrent neural models as well as existing transfer learning models such as VGG-16 and VGG-20. Our proposed model simply outperforms the BLEU scores of previous VGG-16 and VGG-20 models in each segment.

Evaluation Matrix	Proposed Model	VGG - 16	VGG - 20
BLEU-1 [23]	0.5593	0.5362	0.5240
BLEU-2 [23]	0.29967	0.2566	0.2430
BLEU-3 [23]	0.1992	0.1432	0.1378
BLEU-4 [23]	0.0906	0.0591	0.0511

TABLE I TABLE 1

It is possible to conclude that dense networks, such as VGG- 16, are not necessary to extract features from an input picture. Our proposed five-layer convolution neural network model (CNN-5) outperforms existing state-of-the-art techniques in terms of BLEU- 3 and BLEU-4 scores. It can also be proved that even when the number of layers is greatly increased, the BLEU score remains stable. As a consequence, our proposed deep learning model is less computationally expensive than the VGG-16.

V. CONCLUSION

Generating captions for digital photographs is a challenging and ongoing topic in computer vision research. This study presents a CNN-5 model for image captioning and com- pares it with other models that employ transfer learning, such as VGG-16 and Transformer-based models. The goal is to write captions that contain enough information to be considered human-like. The experimental results suggest that a less dense CNN model can be used for feature extraction in image captioning while still achieving a good BLEU-N score, resulting in faster calculations compared to the state- of-the-art models. It is commonly believed that increasing the depth of a neural network improves its learning rate, but this research shows otherwise. Several factors, including the dataset size, vocabulary used, feature extraction model, and hyperparameter values, influence the image caption generation model's accuracy. The proposed model outperforms existing models, but there is still room for future research. However, the proposed model has some drawbacks, such as slow training time due to the large dataset size, and gradient problems due to the regular measurement and categorization of photo features. These issues can be resolved in future developments.

REFERENCE

[1] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep

learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1-36.

[2] Alzubi, J. A., Jain, R., Nagrath, P., Satapathy, S., Taneja, S., & Gupta, P. (2021). Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. Journal of Intelligent & Fuzzy Systems, 40(4), 5761-5769.

[3] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[4] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7008-7024).

[5] Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. Advances in Neural In- formation Processing Systems, 32.

[6] Li, Y., Pan, Y., Yao, T., & Mei, T. (2022). Comprehending and Ordering Semantics for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17990- 17999).

[7] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., & Wang, L. (2022). Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17980-17989).

[8] Wang, J., Tang, J., Yang, M., Bai, X., & Luo, J. (2021). Improving OCR-based image captioning by incorporating geometrical relationship. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1306-1315).

[9] Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., ... & Ji, R. (2021,
May). Dual-level collaborative transformer for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 3, pp. 2286-2293).

[10] Xu, G., Niu, S., Tan, M., Luo, Y., Du, Q., Wu, Q. (2021). Towards
accurate text-based image captioning with content diversity exploration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12637-12646).

[11] Zhou, Y., Zhang, Y., Hu, Z., & Wang, M. (2021). Semi-autoregressive transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3139-3143).

[12] Zohourianshahzadi, Z., & Kalita, J. K. (2021). Neural attention for image captioning: review of outstanding methods. Artificial Intelligence Review, 1-30.

[13] Yan, C., Hao, Y., Li, L., Yin, J., Liu, A., Mao, Z., ... & Gao, X.
(2021). Task-adaptive attention for image captioning. IEEE Transactions on Circuits and Systems for Video technology, 32(1), 43-51.

[14] Zhao, D., Wang, A., & Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14830- 14840).

[15] Lu, H., Yang, R., Deng, Z., Zhang, Y., Gao, G., & Lan, R. (2021).
Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1s), 1-18.

[16] Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., ... & Zhou, J.
(2022). mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. arXiv preprint arXiv:2205.12005.

[17] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., ... & Wang, L.
(2022). GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv preprint arXiv:2205.14100.

[18] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020, April). Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 13041-13049).

[19] Hsu, T. Y., Giles, C. L., & Huang, T. H. K. (2021). Scicap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624.

[20] Hu, J. C., Cavicchioli, R., & Capotondi, A. (2022). ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning. arXiv preprint arXiv:2208.06551.

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B.

(2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

[22] Tanti, Marc and Gatt, Albert and Camilleri, Kenneth, "Where to put the Image in an Image Caption Generator," Natural Language Engineering.
24. 10.1017/S1351324918000098, 2017

[23] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation," ACL- 2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318,
CiteSeerX 10.1.1.19.9416, 2002

[24] Simonyan, K., & Zisserman, A. (2014). Very deep convolu- tional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[25] Dwivedi, P., & Upadhyaya, A. (2022, January). A Novel Deep Learning Model for Accurate Prediction of Image Captions in Fashion Industry. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 207-212). IEEE.