CapStyle - Stylized Image Captioning using Deep Learning Models

P Rama Devi[§], Shylaja S S[¶], Samarth G Vasist[*], Samyak S Sarnayak[†] and Ritik Hariani[‡]
Department of Computer Science, PES University
Bengaluru, India
Email: [§]pramadevi@pes.edu, [§]shylaja.sharath@pes.edu [*]samarthgvashist2000@gmail.com,
[†]samyak201@gmail.com, [‡]ritikhariani@gmail.com

***Abstract*—The development of deep neural networks comprising of CNNs and RNNs has made automatic image captioning a simpler task. However, the written descriptions lack style and a few non-factual aspects. One such style is presenting the description of the image with a set of adjectives which is very common in day-to-day conversations as it effectively and strategically influences one's decisions due to its ability to make the object or the person to stand out either in the affirmative or in a negative manner. We introduce a new dataset of stylized image captions derived from Flickr8k, named CapStyle5k and we design a system to describe an image and present a model that automatically generates captions for the image with styles embedded, using CapStyle5k. CapStyle involves experimentation on the variation of deep learning models which are enhanced by an Attention layer that generates captions with good accuracy. We show experimentally that the performance of CapStyle is competitive with the existing approaches for generating visual captions with styles, as evaluated by various automatic predefined metrics such as BLEU, ROUGE, METEOR, SPICE and CIDEr. Qualitatively, our model generates captions with adjectives embedded into them which describe the image in a natural way.**

***Keywords*—Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory, Gated Recurrent Unit, Image Captioning, Attention, Evaluation Metric.**

I. INTRODUCTION

THE recent years have shown development in the field of computer vision/scene understanding, machine learning and natural language processing. This recent technology can help boost self-driving cars and also serve as an aid to the blind in their day to day activities. Generating automatic captions for an image without any style is called a factual description of the image. However, producing captions with a predefined non-factual style can be more useful in communications, interpersonal relationships, and decision making. These captions can inflict a sense of pride, emotion and feeling. They can also be used in social media platforms to prompt suitable captions with style for an image before posting it online. Therefore, to accomplish this we provide a thorough statistical comparison on the different combinations of models to perform this task.

In the last 5 years, the challenges and complexities faced in image captioning have been reviewed to find the best possible algorithm. For an image captioning model to understand the image, detect features, locations, actions being performed and to finally generate the caption with proper semantics and style has proved to be a difficult task. To accomplish this, CNNs and RNNs have been successful in the past few years.

Fig. 1. Generated Caption: *"brave skilled daring man on motorcycle is driving down track"*

To implement this for the task of stylized image captioning, CapStyle has been introduced.

II. RELATED WORK

In the recent years, there has been advancements in methods to improve automatic captioning/descriptions of an image. This is because images are a strong source for visual interpretation. Image captioning is mostly based on object, action, scene recognition.

According to Christian Szegedy et al. [23] deep learning neural networks play in important role in image classification. Recent accomplishments of using neural networks in image classification [8], [11], [12], [24] instigates strong needs in using neural networks for image captioning [5], [6], [26], [27], [28], [29], [30].

The leading basic process till date for automatic image captioning is to use the encoder-decoder framework which is based on sequence to sequence description generation using neural networks [25]. Andrej Karpathy et al. [6], demonstrated a method to receive state of the art results for image descriptions on Flickr8k, Flickr30k, MSCOCO datasets using multimodal RNNs. Oriol Vinyals et al. [5] extracted global image features using hidden activations of a CNN and then fed them into a LSTM which is trained to generate a sequence of words. Xu et al. [26] took it one step further by proposing the attention mechanism, which selectively attends to different areas of the image when generating words one by one.

The above mentioned image captioning models provide only factual descriptions of an image. However, [31], [32], [33],

[34], [35] are the few that have proposed models that are related to our work of stylised image captioning. Chen et al. [32] have proposed a new style-factual LSTM that uses two groups of matrices to capture the factual and stylized knowledge, respectively, and automatically learns the word-level weights of the two groups based on previous context.

Gan et al. [34] have proposed a novel model component, named factored LSTM, which automatically distils the style factors in the monolingual text corpus. Later at runtime, they control the style in the caption generation process which helps to produce attractive visual captions with the desired style of humour/romance. Mathews et al.[33] developed a model that learns to generate visually relevant styled captions from a large corpus of styled text without aligned images. The core idea of their model was to separate semantics and style. Gella et al. [37] investigated to generate descriptive captions for visually impaired people. Mathews et al.[31] model consists of two parallel RNNs i.e. switching RNNs – one represents a general background language model; another specialises in descriptions with sentiments. Face cap model embed facial expression features in different ways, to generate image captions. Hossain et al. [36] have collected information on the topics related to image captioning and its various methods.

Hence, a model that automatically generates captions for the image with styles embedded, using CNNs as the encoder and RNNs as the decoder is proposed. This model uses different variations of CNNs, LSTMs [15] and GRUs powered by an Attention layer which in-turn generates captions with good accuracy. It is shown experimentally that CapStyle can compete with existing approaches for generating visual captions with styles as evaluated by various automatic pre-defined metrics [17], [19], [20], [21], [22], [38] on the CapStyle5k dataset.

III. Construction of the CapStyle5k Dataset

To accomplish this research in the stylized image captioning, a new dataset called CapStyle5k has been built by modifying the existing Flickr8k image captioning dataset. The dataset consists of 3000 images that have stylized captions which were written manually. Another 2000 image-caption pairs were added from the Flickr8k dataset. In this dataset, each image is paired with 5 stylized captions.

A list of adjectives that were appropriate to the Flickr8k dataset was created. Descriptive adjectives from the list were added to each caption for enhancement of the stylized caption prediction. This was done so that the model could generate captions where the nouns were described adding a sense of common feeling to the images.

IV. Method

A. Data Pre-processing

Each image in the Flickr8k dataset is provided with a unique ID and consists of 5 captions describing the image. In data preprocessing basic data cleaning, which mainly includes lowercasing all the words, removing special symbols and punctuation marks and eliminating all the words consisting of numbers, was performed. A vocabulary which consists of all the unique possible words from the dataset was created. Refining the vocabulary and reducing the size of the unique words was done to increase the occurrences of the words which are more likely to occur or which are common. A threshold was set considering the words which have been repeated a fixed number of times greater than or equal to the threshold.

A dictionary of key-value pairs was formulated whose keys are the image IDs and the values are a list of their respective captions. The contents of this dictionary are then saved as descriptions.

B. Feature Extraction

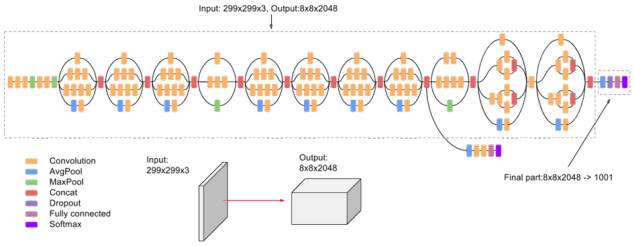


Fig. 2. Feature Vector Extraction (Feature Engineering) from InceptionV3 [1]

Here, the features of the images from the Flickr8k dataset were obtained where the conversion of every image into a fixed sized vector was done in order to be fed as input to the neural network. This is done by feeding the image to a Convolutional Neural Network Model. The size of the image is standardized to 299 x 299. Popular pre-defined CNN models such as InceptionV3 and InceptionResNetV2 were used for feature extraction. However, classification of the image is not done for this task. Instead a fixed-length informative vector for each image is obtained by removing the last classifying layer from the model. Then a 2048-length vector is extracted for every image in the dataset. All the extracted features are serialized into a pickle (.pkl) file.

C. Model Building

In this step, the deep learning model is created using various combinations of different layers. The models have two input layers - one for the image features and another for the previous caption (tokenized). Different layers like Fully Connected layer (Dense layer), Dropout layer, LSTM, GRU and Embedding layers were used. An additional Attention layer was used for few of the models implemented. Different combinations of LSTM, GRU and Attention layers were implemented and evaluated. These different combinations were added either directly to the embedded captions or after concatenating both the inputs.

All captions in the dataset are processed using the tokenizer. The tokenized captions and the features of the corresponding image are used as inputs to train the model. The model is

[1]Feature Vector Extraction (Feature Engineering). *Advance guide to Inception V3 on Cloud TPU*. GOOGLE CLOUD. https://cloud.google.com/tpu/docs/inception-v3-advanced

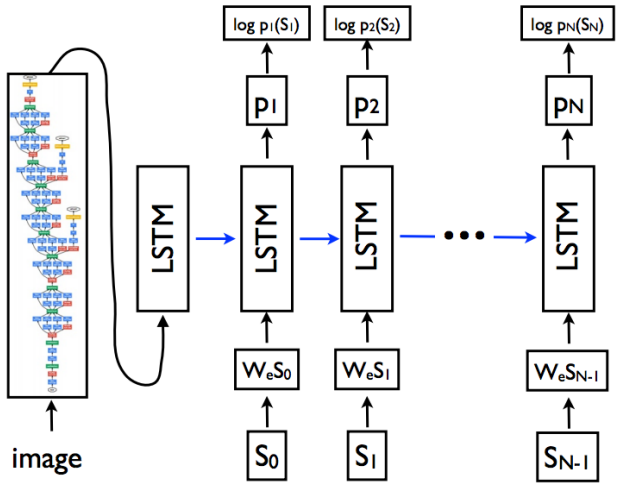


Fig. 3. CNN (encoder) and LSTM (decoder) Architecture. Source: Oriol Vinyals et al. *Show and tell: A neural image caption generator*. In: Proceedings of the IEEE conference on computer vision and patternrecognition. 2015, pp. 3156–3164. [2] [5]

trained for several epochs and in every epoch, the model weights are updated and is tested on the test dataset to calculate the loss and accuracy. The updated model is saved after every epoch.

D. Model Evaluation

The performance of various deep learning models is tested using evaluation metrics to perceive their accuracy in numbers by using the weights generated in the model definition stage. Here, evaluation metrics such as BLEU, METEOR, SPICE, ROUGE, CIDEr [17], [19], [20], [21], [22], [38] respectively are implemented. The model is evaluated after each epoch and the best results are considered. The model generates the captions for the test images and the captions are compared with the pre-defined captions in the dataset and the accuracy is evaluated. The pre-processed captions are tokenized by encoding them and then saving it. Thus, a statistical comparison of the efficiency of every model can be established.

V. Experimentation

A. Dataset

We evaluate our model on the newly constructed CapStyle5k dataset which contains three thousand Flickr images with stylized captions and two thousand Flickr images with factual captions. 4.3K images with their stylized captions were used to train the factual image captioning model. 700 images were then used for validation.

B. Methodology

Different predefined CNN models such as InceptionV3 and ResNetV2 were used as encoder models for feature extraction. We have pre-processed the photos with the InceptionV3 and ResNetV2 model (without the output layer) and have used the extracted features predicted by this model as input. The Feature Extractor model takes as input an image resized to fit the input size (which is 299x299 in case of InceptionV3) and outputs a 2048-length feature vector representing the image.

The Sequence Processor model expects input sequences with a predefined length (32 words) which are fed into an Embedding layer for handling the text input, followed by various combinations of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) neural network layers which consists of 256 memory units, though other sizes were also tested. Single LSTM, Dual LSTM, alternating LSTM and GRU layers and many other variations were used in the decoder model. It was observed that after increasing the number of layers beyond three to four layers of GRU and LSTM, the performance of the model did not show any improvements and even worsened in some cases. Hence, our model was restricted to this number, as the stylized captions generated were pretty acceptable. There is another input model which takes the 2048-length feature as input and reduces it. Both the input models produce a 256 element vector. Further, both input models use regularization in the form of 50% dropout. This is done to reduce overfitting of the training dataset, as the model configuration learns very fast.

The Decoder model merges the vectors from both input models using an addition operation. This is then fed to a Dense 256 neuron layer and then to a final output Dense layer that makes a softmax prediction over the entire output vocabulary for the next word in the sequence.

Hyper parameter tuning was performed to study the performances of various models. Several parameters such as batch size, number of layers, number of units, dropout rate, batch normalization, activation function and optimizers were varied and a comprehensive study was made. Optimizers such as ‘Adam’, ‘Adagrad’, ‘Adadelta’ and ‘SGD’ were experimented out of which ‘Adam’ proved to be the best. Activation functions such as ‘tanh’, ‘sigmoid’, ‘relu’ and ‘softmax’ were varied to obtain the best results.

In the final dense layer where the decoder model merges the vectors of both the inputs we implemented an attention layer as the mechanism which enables the neural network to focus on relevant parts of the input more than the irrelevant parts when doing a prediction task which helped us get better results.

For evaluating the model we used BLEU, ROUGE, METEOR, CIDEr and SPICE scores and a comparative study of various models were made against these evaluation metrics. On performing a comprehensive comparison using the scores obtained as the basis, the model with 4 LSTM layers and an attention layer provided the best results.

VI. Evaluation

The models were evaluated using various evaluation metrics including BLEU, ROUGE, CIDEr, SPICE and METEOR [38].

A. BLEU scores

Bilingual Evaluation Understudy (BLEU) Scores [17] were calculated using ground truths and predictions given by the models. Upto 4-gram BLEU scores were calculated which are

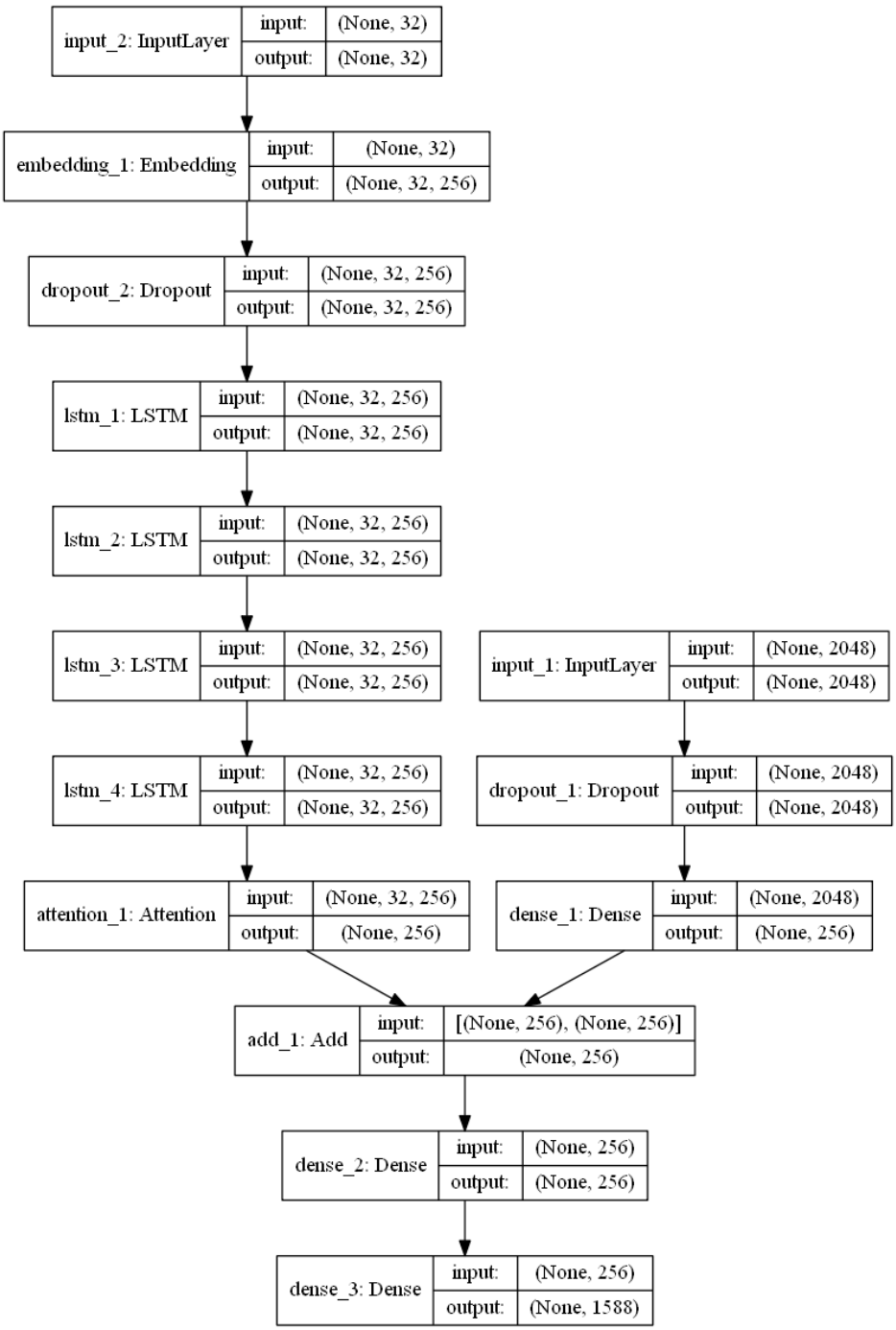


Fig. 4. Final 4 LSTM Model with Attention

denoted as BLEU-1, BLEU-2, BLEU-3 and BLEU-4 for unigram, bigram, trigram and 4-gram respectively. BLEU scores are calculated by finding the precision of n-gram overlaps in all the five captions available per image and hence higher BLEU scores may not represent accurate captions since it measures only precision and not recall. Therefore, we have used other evaluation metrics.

B. ROUGE scores

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [21] is another evaluation metric which uses not only n-gram overlaps but also Longest Common Subsequence (LCS) based statistics. ROUGE-1 and ROUGE-2 use unigram and bigram overlaps respectively. ROUGE-L uses LCS and ROUGE-W uses weighted LCS to find consecutive LCSs.

C. METEOR scores

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [19] is a machine translation evaluation metric which calculates the harmonic mean of precision and recall of unigram matches between sentences and it uses synonyms and paraphrases matching.

D. SPICE scores

Semantic Propositional Image Caption Evaluation (SPICE) [20] is an automated caption evaluation metric which uses scene graphs to better capture human judgement. A set of tuples is generated using the scene graph. SPICE score is found by the F1-score between the ground truth tuples and prediction tuples. Synonym matching is used (as in METEOR) for tuple matching.

E. CIDEr scores

Consensus-based Image Description Evaluation (CIDEr) [22] is an image caption evaluation metric that uses weighting over n-grams. The cosine similarity between n-grams of the predicted and the references is computed for the final score.

VII. RESULTS

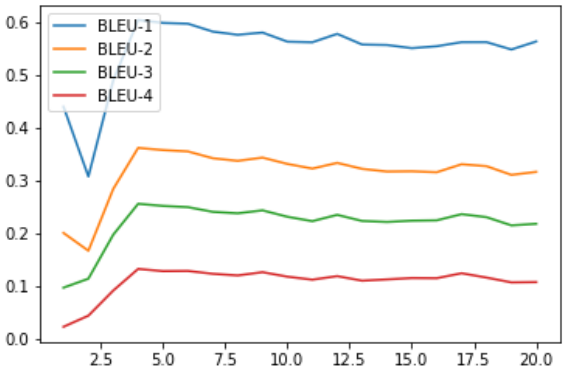


Fig. 5. Variation of BLEU scores with number of epochs

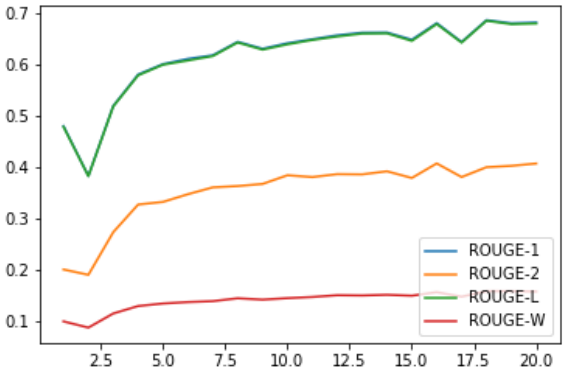


Fig. 6. Variation of ROUGE scores with number of epochs

The scores obtained after evaluation of all models are summarized in the table VII. The encoder models used are InceptionV3 [39] and InceptionResNetV2 [40]. With each of the encoder models, different combinations of LSTM and GRU were used along with an attention layer. All models were evaluated using various evaluation metrics including BLEU, ROUGE, METEOR, SPICE and CIDEr as stated previously.

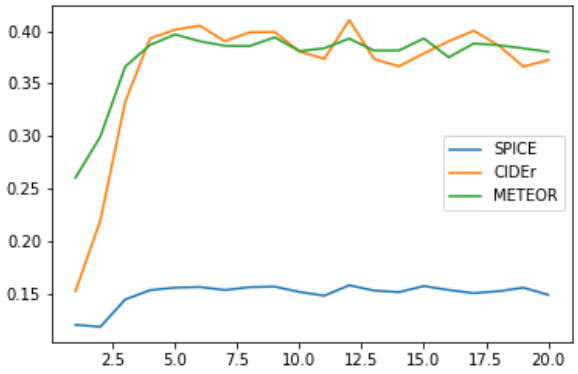


Fig. 7. Variation of SPICE, CIDEr and METEOR scores with number of epochs

It can be observed that InceptionV3 + 4 LSTM model with Attention layer (refer fig. 4) produced the best results as indicated by Figure 9. The captions generated by this model were good. Hence it can be concluded that InceptionV3 + 4 LSTM model with Attention layer is the best model among the models that were tested.

A. Examples of captions generated

Features were extracted from a few images from the Flickr8K dataset and fed as input to the model. Captions were extracted from the tokenized output using the tokenizer. It can be seen in table 8 that the captions have adjectives for certain objects. Hence it can be concluded that we have successfully generated stylized caption.

B. Comparison between captions generated using Flickr8k and CapStyle5k as datasets

The InceptionV3 + 4 LSTM with Attention model was trained using Flickr8k and CapStyle5k as datasets. The same images were fed to both the trained models and the results were compared. It can be observed in table 10 that the captions generated with Flickr8k are only factual and lacked any style. In contrast, the captions generated by CapStyle5k describe some objects using certain adjectives, this gives a sense of style to the caption. Thus it can be concluded that CapStyle5k can be used to generate stylized captions which are better than the factual captions generated using Flickr8k.

C. Variation of scores with number of epochs

During training, one epoch is said to be done when the entire dataset is passed through the model once. Accuracy can be increased by training the model for more epochs. By evaluating the model using the weights generated in each epoch, the variation in scores can be inferred.

The InceptionV3 + 4 LSTM + Attention model was trained for 20 epochs and scores were evaluated and plotted for every

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	METEOR	SPICE	CIDEr
Evaluation Metrics Scores For Various Models with InceptionResNetV2 encoder											
4 LSTM	0.5861	0.3486	0.2469	0.1272	0.6543	0.3831	0.6523	0.1487	0.3955	0.1580	0.4255
2 LSTM 2 GRU	0.5886	0.3469	0.2442	0.1245	0.6452	0.3881	0.6445	0.1465	0.3932	0.1611	0.4128
1 LSTM 1 GRU	0.5919	0.3524	0.2502	0.1256	0.6237	0.3516	0.6230	0.1386	0.3932	0.1568	0.4077
2 LSTM 1 GRU	0.5952	0.3594	0.2544	0.1333	0.6420	0.3775	0.6396	0.1440	0.3962	0.1615	0.4322
Evaluation Metrics Scores For Various Models with InceptionV3 encoder											
4 LSTM	0.6033	0.3615	0.2558	0.1327	0.5814	0.3275	0.5796	0.1293	0.3868	0.1534	0.3931
4 GRU	0.5865	0.3456	0.2395	0.1215	0.6566	0.3891	0.6554	0.1498	0.3864	0.1582	0.4048
2 LSTM	0.5797	0.3416	0.2404	0.1225	0.6308	0.3608	0.6290	0.1418	0.3964	0.1558	0.4221
2 GRU	0.5972	0.3528	0.2493	0.1271	0.6002	0.3355	0.5986	0.1324	0.3934	0.1604	0.4204
1 LSTM 1 GRU	0.5686	0.3335	0.2374	0.1229	0.6500	0.3791	0.6482	0.1488	0.3914	0.1557	0.4003
2 LSTM 2 GRU	0.5729	0.3319	0.2312	0.1181	0.6593	0.3806	0.6585	0.1520	0.3871	0.1531	0.3763
3 LSTM 1 GRU	0.5819	0.3480	0.2465	0.1270	0.6543	0.3804	0.6527	0.1467	0.3971	0.1583	0.4123
2 LSTM 1 GRU	0.5791	0.3367	0.2334	0.1164	0.6497	0.3822	0.6480	0.1472	0.3876	0.1578	0.3947
2 GRU 1 LSTM	0.5534	0.3168	0.2199	0.1052	0.6514	0.3852	0.6496	0.1493	0.3807	0.1538	0.3712
3 LSTM	0.5689	0.3325	0.2337	0.1210	0.6676	0.3915	0.6651	0.1516	0.3856	0.1575	0.4026
3 GRU	0.5907	0.3395	0.2319	0.1120	0.5979	0.3231	0.5960	0.1325	0.3844	0.1541	0.3882
5 LSTM	0.5754	0.3425	0.2390	0.1164	0.6591	0.3716	0.6583	0.1485	0.3905	0.1570	0.3939
1 LSTM	0.5737	0.3411	0.2442	0.1270	0.6605	0.3942	0.6588	0.1509	0.3864	0.1549	0.4099
1 GRU	0.5815	0.3487	0.2475	0.1283	0.6617	0.3747	0.6601	0.1475	0.3910	0.1514	0.4084

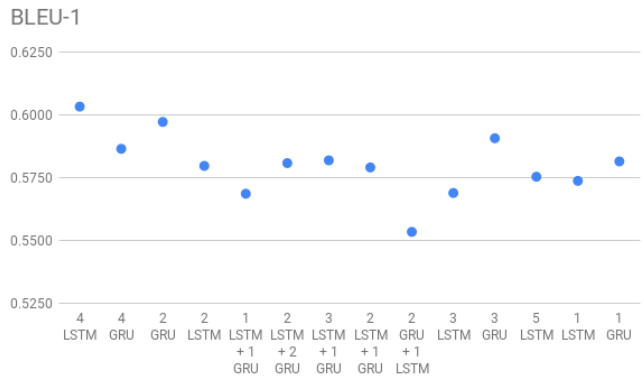
epoch. It can be observed from the graphs in figures 5, 6 and 7 that BLEU scores reach a maximum value after a few epochs and continue to decrease slightly. ROUGE scores, on the whole, keep increasing with the number of epochs. SPICE,

(a) *brave adventurous enthusiastic man climbing rock wall*

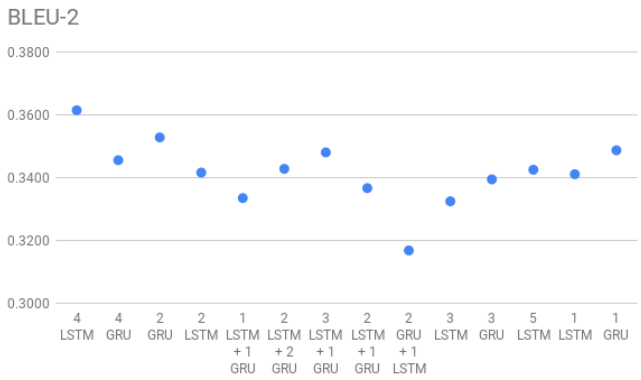
(b) *cute little smiling child in blue coat rides tricycle*

(c) *speedy yellow car is driving down road*

Fig. 8. Captions generated by Capstyle5k using InceptionV3 + 4 LSTM + Attention



(a) Graph depicting BLEU-1 scores for different models.



(b) Graph depicting BLEU-2 scores for different models.

Fig. 9. Variation of BLEU scores for different models

CIDEr and METEOR scores reach a high value after a few epochs and continue to vary by a small margin as the model is trained for more epochs.

VIII. Conclusions and Future work

In this paper, we aimed to generate attractive stylized captions for images. By trying out various combinations on LSTM and GRU layered modules in addition to an Attention layer implemented on our CapStyle dataset and by observing the various evaluation measures,we have arrived at a conclusion that InceptionV3 as the encoder and 4 LSTM layers with an Attention layer as the decoder produces the best results. This is evident from the results obtained above.

Our future work includes generating visually relevant captions with different styles having a mix of emotions visible in the image. We also intend to implement a facial expression recognition model where we can extract the features from the faces present in the image. With this our model could be informed about the emotional quotient present in the image, hence automatically generating suitable captions that convey abstract emotions. Our future work also comprises of implementing a CNN language model as for the decoder architecture to secure better results and also the creation of a better dataset by including more stylised captions which can be scrutinized by the public by conducting the quality check of the captions.

References

[1] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: European conference on computer vision. Springer. 2014, pp. 740–755.
[2] Ranjay Krishna et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations” . In: the International Journal of Computer Vision 123.1 (2017), pp. 32–73.
[3] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics” . In: Journal of Artificial Intelligence Research 47 (2013), pp. 853–899.
[4] Girish Kulkarni et al., “Baby talk: Understanding and generating image descriptions” . In: Proceedings of the 24th CVPR. Citeseer. 2011.
[5] Oriol Vinyals et al. “Show and tell: A neural image caption generator” . In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 3156–3164.
[6] Andrej Karpathy and Li Fei-Fei. “Deep visual semantic alignments for generating image descriptions” . In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 3128–3137.
[7] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. “Im2text: Describing images using 1 million captioned photographs” . In: Advances in neural information processing systems. 2011, pp. 1143–1151.

(a) Flickr8k: *man in red shirt on a bike*
Capstyle5k: *brave adventurous enthusiastic man rides bike on dirt road*

(b) Flickr8k: *two dogs playing in the grass*
Capstyle5k: *two ferocious dogs are playing in the grass*

(c) Flickr8k: *a football player is running with the ball*
Capstyle5k: *fit skilled football player in red jersey is running with football*

Fig. 10. Captions generated by Flickr8k and Capstyle5k using InceptionV3 + 4 LSTM + Attention

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks" . In: Advances in neural information processing systems. 2012, pp. 1097–1105.

[9] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge" . In: the International Journal of Computer Vision 115.3 (2015), pp. 211–252.

[10] C Szegedy et al. "Going deeper with convolutions, CoRR abs/1409.4842" . In: URL http://arxiv. org/abs/1409.4842 (2014).

[11] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large scale image recognition" . In: arXiv preprint arXiv:1409.1556 (2014).

[12] Kaiming He et al., "Deep residual learning for image recognition" . In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770–778.

[13] Ilya Sutskever, James Martens, and Geoffrey E Hinton. "Generating text with recurrent neural networks" . In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011, pp. 1017– 1024.

[14] Zachary C Lipton, John Berkowitz, and Charles Elkan. "A critical review of recurrent neural networks for sequence learning" . In: arXiv preprint arXiv:1506.00019 (2015).

[15] Sepp Hochreiter and Jurgen Schmidhuber. "Long short-term memory" . In: Neural computation 9.8 (1997), pp. 1735–1780.

[16] Junyoung Chung et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling" . In: arXiv preprint arXiv:1412.3555 (2014).

[17] Kishore Papineni et al. "BLEU: a method for automatic evaluation of machine translation" . In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics. 2002, pp. 311–318.

[18] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization" . In: arXiv preprint arXiv:1412.6980 (2014).

[19] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In ACL, 2014.

[20] Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould. "SPICE: Semantic Propositional Image Caption Evaluation" . In: arXiv preprint arXiv:1607.08822

[21] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8, 2004.

[22] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In CVPR, pages 4566–4575, 2015.

[23] Christian Szegedy et al. "Deep Neural Networks for Object Detection" . In Advances in Neural Information Processing Systems 26 (NIPS 2013).

[24] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet. Going deeper with convolutions. CVPR, pages 1–9, 2015.

[25] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, pages 3104–3112, 2014.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, pages 2048–2057, 2015.

[27] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). ICLR, 2015.

[28] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In CVPR, pages 2422–2431, 2015.

[29] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the long-short term memory model for image caption generation. In ICCV, pages 2407–2415, 2015.

[30] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. CVPR, 2016.

[31] A. Mathews, L. Xie, and X. He. Senticap: Generating image descriptions with sentiments. AAAI, 2015.

[32] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, Jiebo Luo. "Factual" ' or "Emotional" ': Stylized Image Captioning with Adaptive Learning and Attention. The European Conference on Computer Vision (ECCV), 2018, pp. 519-535

[33] Alexander Mathews, Lexing Xie, Xuming He. "SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text" . In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8591-8600.

[34] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, Li Deng. "StyleNet: Generating Attractive Visual Captions With Styles" . The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3137-3146.

[35] Omid Mohamad Nezami, Mark Dras, Peter Anderson, Len Hamey.Face-Cap: Image Captioning Using Facial Expression Analysis. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2018.

[36] Hossain, M., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. In Journal, ACM Computing Surveys (CSUR) Volume 51 Issue 6, February 2019, Article No. 118.

[37] S. Gella and M. Mitchell. Residual multiple instance learning for visually impaired image descriptions. NIPS Women in Machine Learning Workshop, 2016.

[38] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In IEEE conference on CVPR, 2016.

[40] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections off Learning. In proceedings of Thirty First Conference on Artificial intelligence, AAAI, 2016.