Improving Diversity and Reducing Redundancy in Paragraph Captions

Chandresh S. Kanani
Department of CSE
Indian Institute of Technology, Patna
Patna, India
cskanani@gmail.com

Sriparna Saha
Department of CSE
Indian Institute of Technology, Patna
Patna, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Department of CSE
Indian Institute of Technology, Patna
Patna, India
pb@iitp.ac.in

***Abstract*—The purpose of an image paragraph captioning model is to produce detailed descriptions of the source images. Generally, paragraph captioning models use encoder-decoder based architectures similar to the standard image captioning models. The encoder is a CNN based model, and the decoder is a LSTM or GRU. The standard image captioning models produce unsatisfactory results for the paragraph captioning task due to the lack of diversity in the generated outputs [9]. The paragraphs generated from standard image captioning models lack in language diversity and contain redundant information. In this work, we have proposed an approach with language discriminator for increasing the diversity in language, and dissimilarity score using word mover's distance [4] for reducing redundant information. Using this approach with a state-of-the-art model at testing time, we have improved the METEOR score from 13.63 to 19.01 for the Visual Genome dataset.**

***Index Terms*—paragraph captions, diversity, redundancy, METEOR score**

I. Introduction

The image captioning models aim to describe objects and their relationships from a given image. Majority of earlier works have focused on generating single sentence descriptions. Single sentence descriptions are small and do not capture all the details of an image. Some of the recent works generate paragraph captions instead of single sentence captions; a paragraph caption generally contains a 5-8 sentence description of an image [9].

In comparison to single sentence captioning, paragraph captioning is relatively new. Krause et al. [3] introduced the first significant image paragraph captioning dataset, a subset of the Visual Genome dataset. The models for single sentence captioning generate repetitive and monotonous captions when trained for paragraph captioning. Some of the prior works have tried to tackle this challenge by architectural changes, such as hierarchical LSTM for generating sentences and words separately [3].

In this work, our main objective is to increase language diversity and reduce redundancy in generated paragraph captions. This work is based on the self-critical sequence training (SCST) [12], [13], a technique using policy gradients to optimize a target metric directly. SCST has been successfully applied to the single sentence captioning but not in paragraph captioning. When trained for paragraph captioning, SCST produces repetitive and monotonous captions. We have addressed these issues by using language discriminator and dissimilarity score.

Our experiments show that using language discriminator for selecting sentences for paragraph generation, and removing repetitive sentences using dissimilarity score, increases model performance on the METEOR score. This approach outperforms complex implementations with hierarchical LSTMs [3] and customized adversarial losses [5]. With the use of language discriminator and dissimilarity score, we have improved the METEOR score for the Visual Genome dataset from 13.63 (for up-down SCST model [1]) to 19.01 (with our approach).

II. Background and Related Works

All recent works on image captioning are based on encoder-decoder architecture, introduced by Vinyals et al. [16]. In such models, the encoder is generally a CNN trained for classification, and decoder is a LSTM or GRU. Anderson et al. [1] improved single sentence captions by using object detection on the encoder side.

Krause et al. [3] introduced the first paragraph captioning dataset containing 19,561 images, Table I shows the number of images in train, test and validation sets. They also showed that paragraph captions are more diverse and contain more pronouns, verbs and co-references. Krause et al. [3] also discussed that paragraph captions contain more information and describe more objects in comparison to the single sentence captions of MSCOCO [7].

TABLE I
Dataset statistics: number of samples in train, test and validation sets of Visual Genome dataset.

Data	Number of Samples
Train	14,574
Validation	2,486
Test	2,489

Krause et al. [3] also proposed models for paragraph captioning: one template-based model and two encoder-decoder based models. In both encoder-decoder models, the encoder is an object detection model pre-trained for dense captioning. The first model, called the flat model, treats a paragraph as a single sequence and generates whole paragraph word by word.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

On the other hand, the hierarchical model uses two LSTMs, one for sentence-level generation and one for word level.

Recently Melas-Kyriazi et al. [9] introduced a penalty score for tri-gram repetition while training. With this modification, they have achieved an improvement over SCST. To the best of our knowledge, their model achieved a METEOR score of 17.86 for the Visual Genome dataset.

For this work, we have used the up-down SCST model from Anderson et al. [1]. This model is similar to the flat model from Krause et al. [3], except Anderson et al. [1] has used attention with the top-down mechanism.

III. Approach

The major contributions of our proposed approach are the following:

- We have improved paragraph captions without any major architectural changes.
- Existing paragraph caption generation systems suffer from redundancy and lack in language diversity. In order to improve on these aspects, we have introduced language score and dissimilarity score. We have selected sentences based on the introduced scores and finally top-scoring sentences are selected for the generating the final output.
- The effect of language discriminator and dissimilarity score in improving paragraph captions is shown on a latest paragraph caption generation model, named up-down SCST model.
- To the best of our knowledge, SCST is the best technique in the field of paragraph captions. For that reason, this model is utilized for generating initial set of captions from the given image.

We have used up-down SCST model from Anderson et al. [1] as our base paragraph captioning model. No changes are incorporated in the SCST model architecture. The encoder in this model is trained for object detection and extracts between 10 to 100 objects per image. For each object, the encoder outputs a vector of dimension 2048 after spatial max-pooling. The decoder is a single layer LSTM with hidden dimension 512 and top-down attention.

The main idea behind this work is to generate diversified captions. Previous image captioning works generate captions by choosing the word with maximum probability at each step. Here for increasing the diversity, we sample words according to their probabilities. With this approach, a word which does not have maximum probability can also be selected for caption generation. This induces diversity in the generated captions. With this method, grammatically wrong sentences also get generated, to overcome this we have used language discriminator. Language discriminator provides a score in the range of 0-1; language score indicates language diversity and correctness of generated caption sentences.

We sample 10 captions for an input image. Then we have selected best sentences from the generated captions based on the language score and dissimilarity score.

TABLE II
SCORES PROVIDED BY LANGUAGE DISCRIMINATOR TO DIFFERENT SENTENCES.

Sentence	Language Score
Good Sentences	
There 's people walking on the platform	0.99978906
Four women are standing on a sidewalk	0.99978644
Six people are standing on a field	0.9997836
A cat is laying on the couch	0.9997826
There are decorative frosting on the cake	0.99959093
There are tall green trees on the ground	0.9995908
Bad Sentences	
The brown is brown	0.17665803
The is a mechanical on top of the book	0.17665455
This is a picture of a street in This picture	0.17663047
The shirt has a and mustache	0.07031095
The persons is standing on the surfboard of The kid in front of the boy has a yellow and white wave	0.07029981
A pile of fruits fruit is eating on a apples	0.06998764

A. Language Discriminator

The language discriminator encodes each input sentence with a bidirectional LSTM having 512 hidden dimensions. The last layer of discriminator is fully-connected layer with one neuron and sigmoid activation function.

Let, s be input sentence to the language discriminator. Then, $LS(s)$ is the output of language discriminator. $LS(s)$ is a score in the range of 0-1, 1 being grammatically most accurate and diverse sentence.

For training the discriminator, we have considered sentences of all the ground truth captions from train set as grammatically diverse and correct, and have assigned a score of 1 to them. For negative samples, we have modified the ground truth sentences by repeating or swapping randomly selected words, we have assigned a score of 0 to the negative samples.

The discriminator is trained with binary cross-entropy loss for 30 epochs. The discriminator achieves highest classification accuracy of 95.03% for validation set on epoch 23, after that accuracy becomes stable. We have achieved a classification accuracy of 96.11% on the test data. Fig. 1. shows the discriminator architecture and Fig. 2. shows performance of discriminator on validation data. Table II reports sentences and language scores provided by the discriminator.

B. Dissimilarity Score Calculation

Word Mover's Distance (WMD) We have used word mover's distance for calculating dissimilarity between generated sentences [4]. Word mover's distance between two sentences, S1 and S2, represents minimum cumulative distance which words of S1 needs to travel to reach to the words of S2. Word mover's distance uses pre-trained word embeddings for generating word clouds of sentences S1 and S2.

Each sentence is represented by the relative frequencies of words it contains, i.e., for the i^{th} word type,

$$d_{S1,i} = \frac{count(i)}{|S1|} \tag{1}$$

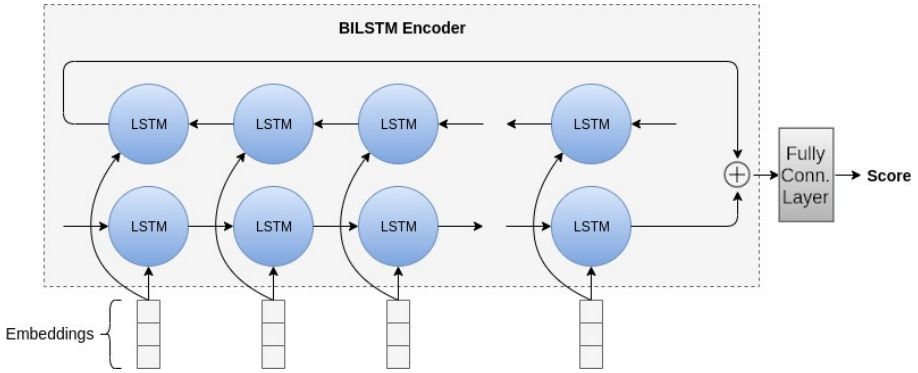

Fig. 1. An overview of language discriminator, the discriminator provides a score in the range of 0-1 for each input sentence.

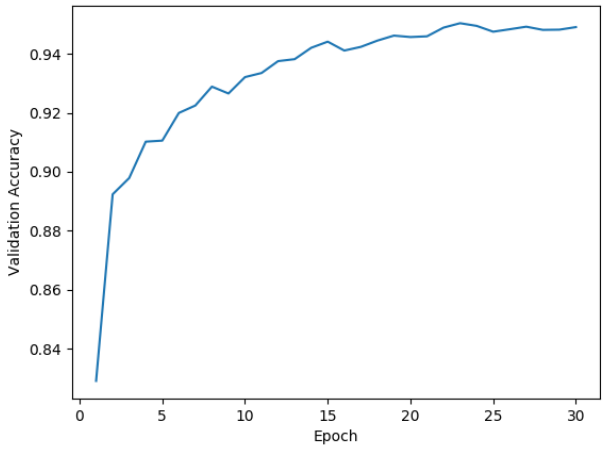

Fig. 2. Performance of language discriminator on validation set.

where $|S1|$ is the total word count of sentence S1, and $d_{S2,i}$ is defined similarly.

Now let the i^{th} word be represented by $W_i \in \mathbb{R}^{\text{m}}$, i.e., an m-length embedding, allowing us to define distances between the i^{th} and j^{th} words, denoted by $\Delta(i,j)$. $\mathbf{V}$ is the vocabulary size. We have followed Kusner et al. [4] and used the Euclidean distance $\Delta(i,j) = \|\mathbf{V}_i - \mathbf{V}_j\|_2$.

The WMD between sentence, $S1$ and sentence, $S2$ is then the solution to the linear program:

$$WMD(S1,S2) = \min_{\mathbf{T}\geq 0} \sum_{i=1}^{\mathbf{V}} \sum_{j=1}^{\mathbf{V}} \mathbf{T}_{i,j}\Delta(i,j) \tag{2}$$

subject to:

$$\forall i, \sum_{j=1}^{\mathbf{V}} \mathbf{T}_{i,j} = d_{S1,i} \tag{3}$$

$$\forall j, \sum_{i=1}^{\mathbf{V}} \mathbf{T}_{i,j} = d_{S2,j} \tag{4}$$

$\mathbf{T} \in \mathbb{R}^{\text{V x V}}$ is a non negative matrix. $\mathbf{T}_{i,j}$ denotes the number of i^{th} words in $S1$ which are associated with j^{th} words in $S2$. The constraints ensure that the flow of a given word cannot exceed its weight. Specifically, WMD ensures that the entire outgoing flow from word i equals $d_{S1,i}$, i.e., $\sum_{j=1}^{\mathbf{V}} \mathbf{T}_{i,j} = d_{S1,i}$. Additionally, the amount of incoming flow to word j must match $\sum_{i=1}^{\mathbf{V}} \mathbf{T}_{i,j} = d_{S2,j}$.

The above optimization problem is a special case of Earth Mover's Distance (EMD) [14], specialized solvers are available to solve this problem [8], [11].

Dissimilarity Score Let FC_i be the i^{th} sentence of final caption at time instance t. Then, dissimilarity score of sentence s with respect to final caption FC is calculated as:

$$\text{D}S(s,FC) = \frac{\sum_{i=1}^{t} WMD(s,FC_i)}{t}, t > 0 \tag{5}$$

Length Penalty We have also given a length penalty to the sentences with short length to prevent them from getting selected for the final caption. Let $|s|$ be the length of a candidate sentence, s. The median length of the candidate sentences CS is represented as $ML(CS)$, MML denotes minimum median length (*this is set to prevent short truncated sentences from getting selected*). Then, the length penalty is calculated as:

$$\text{L}P(s,CS) = \min\left(1, \frac{|s|}{\max(MML, ML(CS))}\right) \tag{6}$$

We have used minimum median length (MML) as 3 while calculating length penalty. This is done to prevent selection of sentences with less than 3 words when the median length of candidate sentences CS is smaller than 3.

The Fig. 3. shows length penalty corresponding to different median sentence lengths.

C. Final Caption Generation

The first sentence for the final caption is the sentence with maximum language score from the first sentences of the candidate paragraph captions. For subsequent sentences, similarity of candidate sentences with the final caption at that

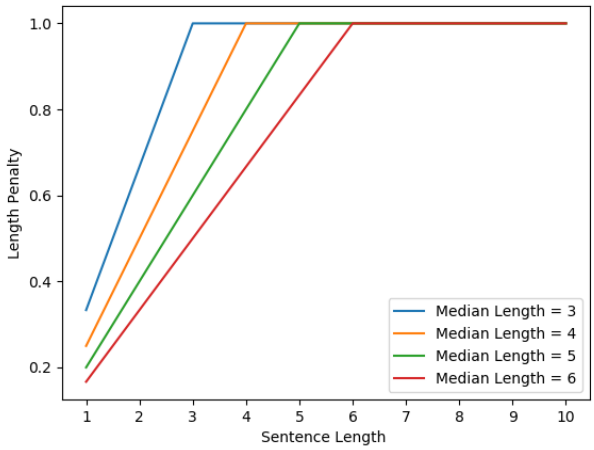

Fig. 3. Graph showing length penalty given for different values of median sentence lengths.

instance of time is also considered, and based on that, best sentence is selected.

First sentence for the final caption is selected as:

$$S_{ent} = \underset{s}{\operatorname{argmax}} LS(s) \tag{7}$$

Other sentences are selected as:

$$S_{ent} = \underset{s}{\operatorname{argmax}}(LS(s) + DS(s, FC)) * LP(s, CS) \tag{8}$$

In equations 7 and 8, s denotes a sentence, $LS(s)$ is language score for a sentence. $DS(s, FC)$ is dissimilarity score between sentence s and final caption FC at that point of time. $LP(s, CS)$ is the length penalty calculated as discussed in previous section.

We have tried different values ranging from 0.5 to 4.0 (with an interval of 0.25) as minimum score which a candidate sentence must have for selection in final caption. The threshold of 2.25 for minimum score *(candidate sentences with scores lower than this bound are not added to the final caption)*, for a sentence provides best results. Fig. 7. shows graph of METEOR score obtained for different minimum score thresholds.

Fig. 4. shows our approach to caption generation. Fig. 5. shows candidate captions and final generated caption after the application of the above approach.

IV. Results

To get diverse candidate captions, we have generated captions by sampling words according to their probabilities; due to this, the final captions have semantically similar meanings with ground truth sentences but differ in wordings. So, to effectively measure the performance of this model, we have used the METEOR score. In contrast to other evaluation metrics, such as BLEU [10], CIDEr [15] and ROUGE [6], METEOR [2] performs stemming and synonymy matching. The use of probability based word sampling generates captions which describe the image information in different wordings, which cannot be evaluated by metrics without synonymy matching. Thus, we have used METEOR for the evaluation purpose of this work.

Table III shows a comparison of our model with some prior works; our model performs better than previous works. Fig. 6. shows a qualitative comparison of our model with the up-down SCST model from Anderson et al. [1]. It can be noted from the figure that with this approach we have removed the repetition. Also, the output of our model contains sentences with complex structures and uses more unique words.

We have also calculated the number of unique words used for generating the captions to compare the language diversity of the models. It can be seen from Table IV that captions generated using our approach utilize approximately four times more unique words in comparison to the up-down SCST model [1] with an increase of only 1/3 in caption length.

We have also shown that use of language score and dissimilarity score with length penalty provides better results compared to other combinations of selection parameters. Table V shows results obtained by using different sets of parameters for selection of sentences for final caption.

TABLE V
METEOR scores obtained for different combinations of selection parameters.

Parameter Combination	METEOR Score
Language Score	14.65
Dissimilarity Score	14.51
Language Score & Dissimilarity Score	17.81
Language Score, Dissimilarity Score & Length Penalty	**19.01**

Fig. 7. shows graph of results obtained for different minimum score threshold.

A. Statistical Significance Test

We have also conducted statistical t-test [17] by obtaining METEOR scores for 5 different runs. METEOR scores obtained by up-down SCST and our approach are shown in Table VI.

A p-value [17] smaller than 0.05 shows that results are statistically significant. We have obtained a p-value of $1.5531633797359418e-15$ for t-test showing that improvements obtained by our proposed approach are statistically significant.

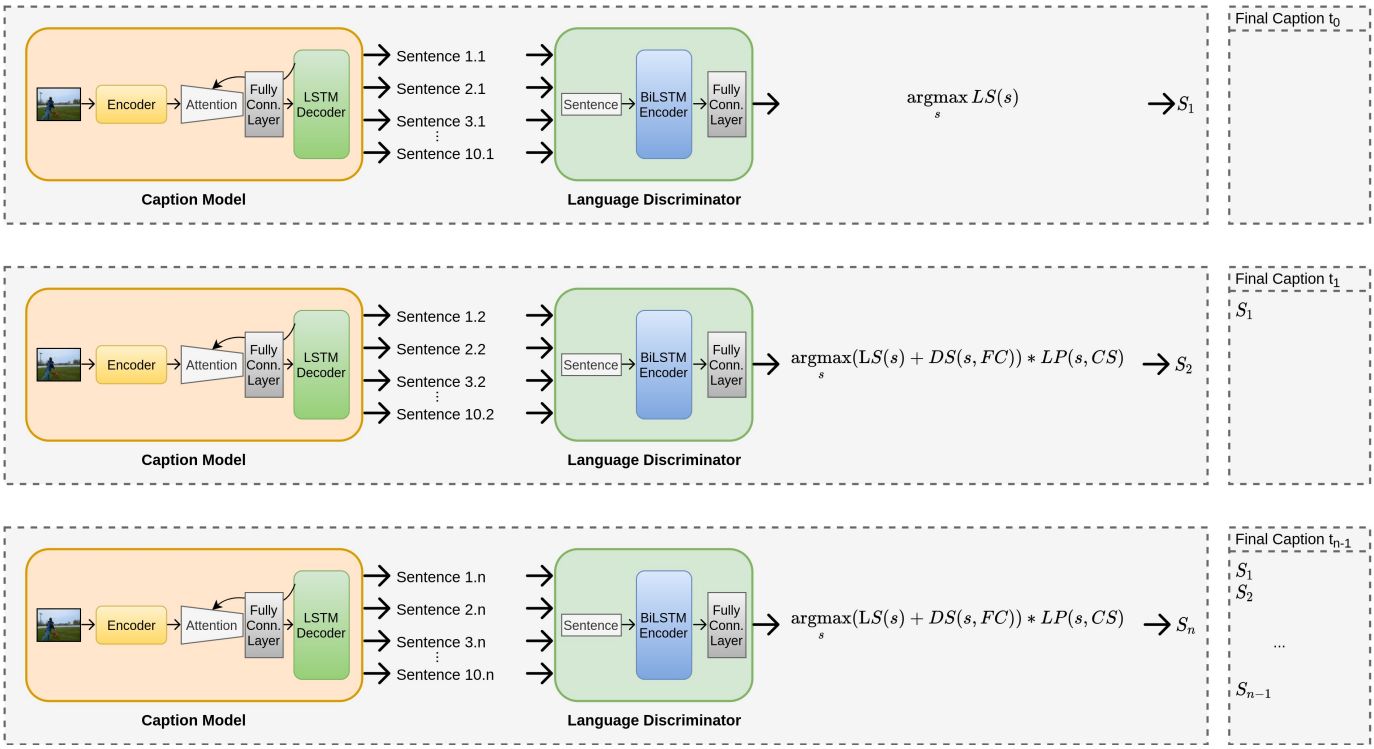

Fig. 4. Figure shows our approach of generating paragraph captions. Sentence X.Y symbolizes Y^{th} sentence of X^{th} paragraph. The final captioning box shows final caption at an instance of time. $LS(s)$ is language score for sentence s, $DS(s, FC)$ is dissimilarity score of sentence s with respect to final caption FC and $LP(s, CS)$ is length penalty for sentence s with respect to candidate sentences CS.

TABLE III
A COMPARISON OF RESULTS OF OUR PROPOSED APPROACH WITH THE RESULTS OF PREVIOUSLY PUBLISHED WORKS.

Model Name	**METEOR Score**
Krause et al. [3] (Template)	14.31
Krause et al. [3] (Flat w/o object detector)	12.82
Krause et al. [3] (Flat)	13.54
Krause et al. [3] (Hierarchical)	15.95
Liang et al. [5] (w/o discriminator)	16.57
Liang et al. [5] (w/ discriminator)	17.12
Anderson et al. [1] (XE training, Up-Down)	13.66
Anderson et al. [1] (SCST, Up-Down)	13.63
Melas-Kyriazi et al. [9] (XE training, w/ rep. penalty)	15.17
Melas-Kyriazi et al. [9] (SCST, w/ rep. penalty)	17.86
Ours (w/ language score, dissimilarity score & length penalty)	**19.01**

TABLE IV
A COMPARISON OF UNIQUE WORDS USED FOR CAPTION GENERATION.

Model Name	**Words Per Para**	**Total Unique Words**	**Unique Words Per Para**
Anderson et al. (SCST, Up-Down)	91.01	541	15.19
Ours (w/ language score, dissimilarity score & length penalty)	123.63	2478	57.51

Candidate 1: A woman wearing the frisbee is running frisbee . The frisbee is green . The man is at the frisbee . The dog is standing on the grass . The small wearing a gray polo . The dog has long ears . The dog is looking in the side . The frisbee is blue . The dog is standing on a field . There are many tree on the left . There are several trees surrounding the field . There are a lot of trees in the background .
Candidate 2: Two people are playing frisbee in a field . The grass is in the air . The dog is brown . The woman is looking up . The girl . is dog on the ground . The frisbee is white . The woman is wearing a green shirt . The man is wearing a blue collar . The woman is standing in the grass . The sky is gray and overcast . There are trees in the background . There is a building near the distance . The woman is holding a frisbee . There are signs behind the light .
Candidate 3: A woman is kneeling in a field . The smiling is yellow . The dog is frisbee . The woman is wearing a blue shirt . The woman is holding a Frisbee . The boy is bare by the dog . The dog is skin . The frisbee is silver and blue . The woman is wearing a jacket . The woman is short and brown . The grass is green . The sky is bright with the clouds . . The dog is long and white .
Candidate 4: A woman is running on a field . The woman is standing than the woman towards the frisbee . The dog is gray blue The woman is wearing is white and blue color . The dog is small on the grass . The woman is standing in a grassy field . The field is surrounded by trees . The man is wearing a brown shirt . The water is made of stone . A short fence is standing near the dog .
Candidate 5: there is a woman and a girl in the grass . the person is brown , and . The frisbee is black in color . the woman is running on a field . The person is wearing green shirt . The woman is holding a red Frisbee . the person is standing by the white and yellow one . The woman is running on a fence . The sky is blue .
Candidate 6: There are two women sitting on a field . The girl is bending a frisbee . There is a dog on the lake . The dog is brown . The woman is wearing a black . The woman is playing the many person . There is is in the field is The grass is green . There are trees beside the trees . The trees are yellow . There is a green tree near the field .
Candidate 7: The woman is standing on the yellow field . trees dog is black on the Frisbee . The is a woman standing behind her dog . The lady is wearing a . The woman 's tail is resting on the dog . The dog is green . The trees are tall . Behind the mountain is a bright green .
Candidate 8: Two woman is standing on a grassy field . The woman is wearing a blue . The girl is wearing a green jacket . The woman is wearing a black shirt . The woman is standing on a grassy field . The woman is blue and black . is a gray fence in the grass . There is a large group of trees and the woman . Green bushes are standing near the fence . There is a crowd by the trees . There is a fence on the court .
Candidate 9: The woman is running in the grass . The woman is looking up . A dog is running . The girl is wearing a blue scarf on The woman is standing in a field . The dog is white in color . The girl is sitting in grass field . The field is green and green . There area is surrounded by trees . A tree is standing behind the woman . The sky is blue .
Candidate 10: A woman and a dog are standing outside on a grassy field . The her park is very small . The woman is wearing a blue shirt . The woman is white up are black . The woman is holding a frisbee . The field is short and green . The dog is looking to a building . The grass is green and green . There is a fence behind the woman . A green tree is on the side of the fence . A couple is standing on the second of the lake .

Final Caption: A woman and a dog are standing outside on a grassy field . The girl is bending a frisbee . The frisbee is black in color . The dog is standing on the grass . The person is wearing green shirt . The field is surrounded by trees . The woman is wearing a jacket . The woman is short and brown . There are many tree on the left . The sky is gray and overcast . There is a fence on the court . There is a building near the distance .

Ground Truth: A woman is throwing a frisbee in the air . The frisbee is round and green . It is flying in the air. The woman's dog is on the grass jumping . The dog is large and brown . The woman is wearing blue jeans and a blue jacket . The sky is cloudy and gray . The woman and dog are on green grass . It is surrounded by a short white wall .

Fig. 5. 10 candidate captions generated using up-down SCST from Anderson et al. with probability based word sampling and final generated caption using our approach. Sentences in blue color are selected for final caption.

Up-Down SCST: A man is standing on a tennis court . He is wearing a white cap , blue shirt and blue shorts . ~~The man is holding a racket in his hand . The man is holding a racket in his hand .~~ The man is wearing a white shirt . ~~The man is holding a racket in his hand .~~ The court is green and white . ~~The court is white and white . The court is white and white . The court is white and white .~~

Ours: The man is taken on a court . He is wearing a black cap on his head and a navy blue shirt . There is a red and white tennis racket on the court . There are people on the wall behind the person . The tennis boy is holding a tennis tennis ball . The side of the tennis court is white and gray . Part of a crowd of white people are sitting around the player . Two boys are sitting next the field watching the game .

Ground Truth: A man is standing on a tennis court. He is wearing a white baseball cap, blue and orange jersey, blue and white wristband and blue shorts. The man is holding a black, red and white racket in his hand. The court has white painted lines on it. Part of another man can be seen sitting on the court inside of a black and white box. He is wearing a green shirt and blue jeans.

Up-Down SCST: A man is standing on a tennis court . He is wearing a white shirt and white shorts . ~~The man is holding a racket in his hand . The man is holding a racket in his hand . The man is holding a racket in his hand . The man is holding a racket in his hand .~~ The court is green and white . The man is wearing a white shirt . ~~The man is holding a racket in his hand .~~ The court is made of metal .

Ours: A man is wearing a white shirt . He is wearing dark shorts , white shorts , white socks and a blue headband . He is playing tennis and is holding a racket in his hands . A large white fence can be seen in the background of the man . The court has also black lines on it . A tall chain link fence is on the window . There is a black fence next to the person . The player on the racket is white and black . The railing is made of brown with metal .

Ground Truth: A tennis player is outdoors playing tennis in the daytime. He is swinging a tennis racket that is colored red with a red W on it. The player is also wearing a blue sweatband on his head, a white t shirt, a watch, and blue shorts. There is a green shrub in the background with a few people behind it.

Up-Down SCST: A box of donuts is sitting on a table . ~~The box is white . The box is white . The box is white . The box is white . The box is white . There is a white cup on the table . There is a white cup on the counter . There is a white cup on the counter . There is a white cup on the counter . There is a white cup on the counter . There is a white cup on the counter .~~

Ours: There are two donuts on the table . The box is brown and there is a white and yellow bin in the container . The donuts are sitting under an orange bag . Some of the donuts are brown , some are on a block . Part of a cabinet is sitting behind the counter . There is a clear plastic mug . There 's a bottle of the drink near the container . The table has an oven . The paper is red and white . There are various dessert on the counter near the tray .

Ground Truth: A black and silver microwave sits on a white counter top in this photo. A black and silver coffee maker is sitting by the microwave. Two white coffee mugs and a container of sugar are also on the counter top. A stack of multi colored cups have tipped over behind the coffee mugs. A box of four doughnuts are sitting on the counter top. One of the doughnuts is glazed, two have white icing and the last one has chocolate icing on it.

Fig. 6. Comparison of paragraph captions generated by proposed approach (Ours) with paragraph captions generated by the up-down SCST model from Anderson et al. [1]. The strike-through text shows repeated sentences in the caption. Sentences with complex structure and language diversity are shown with green font color. Underlined text shows semantically wrong information.

TABLE VI
METEOR SCORES OBTAINED FOR CAPTIONS GENERATED USING UP-DOWN SCST AND CAPTIONS GENERATED USING PROPOSED APPROACH (OURS) FOR 5 DIFFERENT RUNS

Run	SCST METEOR	Ours METEOR
1	13.74	19.01
2	13.69	18.96
3	13.76	18.86
4	13.78	18.97
5	13.78	18.92

V. Error Analysis

We have done a thorough error analysis on the outputs provided by the proposed approach. The observations are as follows:

To increase the diversity of generated captions, we have sampled words for candidate captions according to their probabilities. Due to this there are some sentences in the generated captions, which are grammatically correct but do not correctly capture features of objects in the input image (underlined text in Fig. 6.). This issue can be resolved by using a relatedness discriminator, for comparing relationships between images and sentence content.

VI. Conclusion

The current work aims to increase diversity and decrease redundancy in the paragraph generation task from an image. In this work, we have shown that paragraph captions can be made more diverse by using language discriminator and dissimilarity score. By using our proposed methods, we improved the

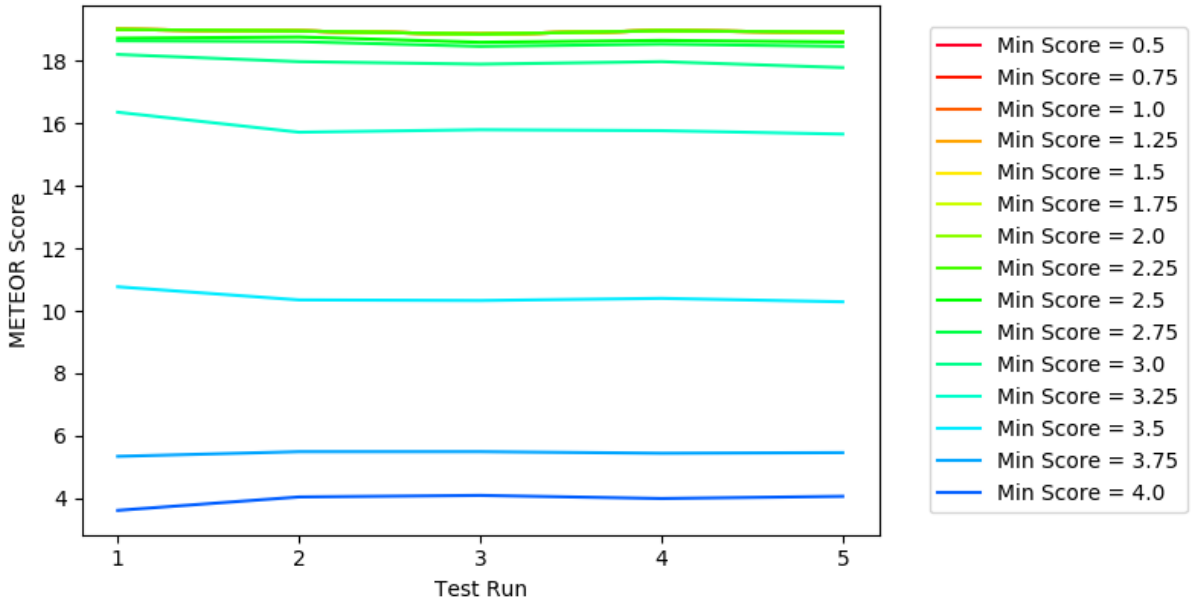

Fig. 7. METEOR score obtained for different values of minimum score threshold for 5 runs. Some plots are overridden as METEOR scores obtained for threshold value of 0.5 to 2.25 are same.

METEOR score of the state-of-the-art up-down SCST model [1] by 5.38 for the Visual Genome dataset. For increasing the diversity, we sample words with their corresponding probabilities, which make some sentences semantically incorrect. In the future, we will work on solving this challenge using relatedness discriminator, for comparing relationships between image and sentence content.

Acknowledgement: Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017. [Online]. Available: http://arxiv.org/abs/1707.07998

[2] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *ACL*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72.

[3] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," *CoRR*, vol. abs/1611.06607, 2016. [Online]. Available: http://arxiv.org/abs/1611.06607

[4] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *ICML, 2015, Lille, France*, 2015, pp. 957–966. [Online]. Available: http://proceedings.mlr.press/v37/kusnerb15.html

[5] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition GAN for visual paragraph generation," *CoRR*, vol. abs/1703.07022, 2017. [Online]. Available: http://arxiv.org/abs/1703.07022

[6] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[7] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[8] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 840–853, 2007.

[9] L. Melas-Kyriazi, A. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *EMNLP*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 757–761.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *ACL*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[11] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *ICCV*, 2009, pp. 460–467.

[12] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR, 2016, San Juan, Puerto Rico*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.06732

[13] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," *CoRR*, vol. abs/1612.00563, 2016. [Online]. Available: http://arxiv.org/abs/1612.00563

[14] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *ICCV*, 1998, pp. 59–66.

[15] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, June 2015.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CoRR*, vol. abs/1411.4555, 2014. [Online]. Available: http://arxiv.org/abs/1411.4555

[17] J. D. Winter, "Using the student ' s t-test with extremely small sample sizes," 2013.