ULTRA-CONTEXT: MAXIMIZING THE CONTEXT FOR BETTER IMAGE CAPTION GENERATION

by

ANKIT KHARE

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2019

To my selfless friends Devi Prasad Tripathy and Stacey Aria Young who helped me stay on track and assisted me in many ways towards the completion of my thesis. And to my parents who gave me financial stability so that I never had to worry about anything but my thesis.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervising professor Dr. Manfred Huber for constantly advising and encouraging me. His invaluable advice, deep expertise in various computer science sub-domains and huge breadth in computer science as a whole has proved to be extremely helpful during the entire course of my graduate studies. I wish to thank my committee members Dr. Vassilis and Dr. Gao for their interest in my research and for taking time to be in my committee.

I would also like to extend my appreciation to Dr. Huber for providing financial support for my studies and giving me a stress free environment for continuing my research on many topics I have been interested in. I am especially grateful to Devi Prasad Tripathy for his interest in my research and for the helpful discussions and invaluable comments. I wish also to thank my friend Stacey Young for taking the time to critically evaluate this manuscript and my research papers.

I am grateful to all the professors who taught me during the years I spent as a graduate student. I would like to thank Dr. Kamangar whose computer vision class inspired me to study neural networks using online courses.

Finally, I would like to express my deep gratitude to my parents who have encouraged me and also sponsored my graduate studies.

Feb 25, 2019

ABSTRACT

ULTRA-CONTEXT: MAXIMIZING THE CONTEXT FOR BETTER IMAGE CAPTION GENERATION

ANKIT KHARE, M.S. Computer Science

The University of Texas at Arlington, 2019

Supervising Professor: Dr. Manfred Huber

Several combinations of visual and semantic attention have been geared towards developing better image captioning architectures.

In this work we introduce a novel combination of word-level semantic context with image feature-level visual context, which provides a more holistic overall context for image caption generation. This approach does not require training any explicit network structure, using any external resource for training semantic attributes, or supervision during any training step.

The proposed architecture addresses the significance of learning to find context at three levels to achieve a better trade-off as well as a balance between the two lines of attentiveness (word- level and image feature-level).

The structure of the visual information is very different from the structure of the captions to be generated. Encoded visual information is unlikely to contain the maximum level of structural information needed for correctly generating the textual description in the subsequent decoding phase. Attention mechanisms aim at streamlining the two modalities of language and vision but often fail to find a balance between them. Our novel approach to establish this balance where the encoder-decoder pipeline learns to pay balanced attention to the two modalities leads to the captions not drifting towards the language model irrespective of the visual content of the image or towards the image objects regardless of the saliency observed in the generated sentence history.

We demonstrate how the encoder's convolutional feature space attended in a top-down fashion and in parallel conditioned over the entire n-gram word space, can provide maximum context for sophisticated language generation.

Effective architectural variations to produce hybrid attention mechanisms streamline a model towards better utilization of rich image features while generating final captions. The impact of this mechanism is demonstrated through extensive analysis using the MS-COCO dataset.

The proposed system outperforms state-of-the-art results, illustrating how this context-based architectural design opens up new ways of addressing context and the overall task of image captioning.

TABLE OF CONTENTS

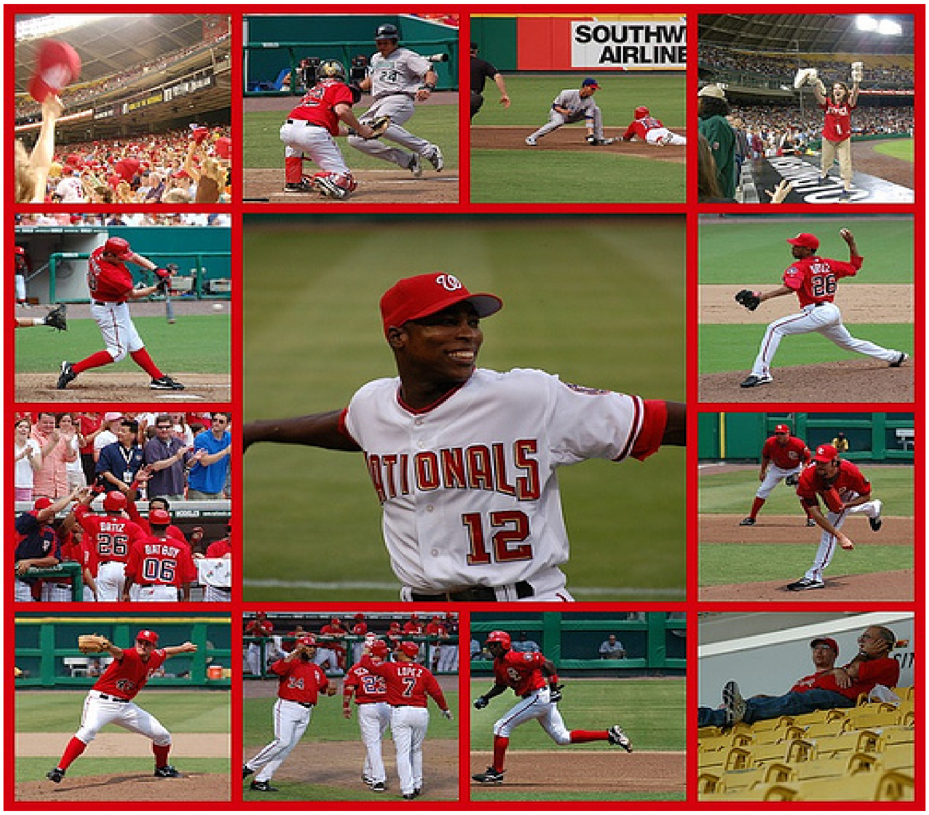
LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Image captioning [1] requires deep scene understanding and subsequent generation of grammatically correct sentences, reflecting abstraction of the understood concepts. It is challenging not only due to the very nature of the process itself but also because it is so far not possible to make machines go through the same experience and context that allows humans to understand the underlying complex concepts which might sometimes not even be visually obvious. Deep neural network based encoder-decoder architectures [2, 3, 4, 5, 6] have been quite successful in producing good captioning results and represent the current state-of-the-art. This is mainly due to their ability to form complex representations using large datasets [7, 8, 9] and to learn long-range sequences [10] which together make it possible to summarize an image. Broadly, during the process of describing an image, humans extract information from an image because of: (i) their ability to form meaningful representations of what they see, and (ii) intelligent utilization of their knowledge of language to verbally describe its essence. The second part which comes from language skills is equally important here. Even in the learning mechanism in *Homo Sapiens* [11], knowledge of language and its association with learned visual representations is very significant to allow it to express the meaning of observed or virtual scenarios. The same has to be the case in image captioning architectures where training the model to understand semantic language concepts and making it learn how to associate them with visual features is of high importance. The inspiration for our approach to produce quality captions comes from the same consideration. Coupling of visual representations with word vectors is an extremely critical part of the entire process and the context provided to the network to aid in this process has a definitive influence on theoverall automatic image captioning task. Ground truths provided in the training dataset reflect the context in which annotators caption the image. Neural networks learn to understand that context and their inference is evaluated on the basis of the extent of their success in the process. Therefore, it is obvious that the better the network gets at understanding this context, the better it can describe an image (Fig. 1.1).



Soft Attention [6]**:** A baseball player throwing a ball on a field.
CaptionBot [12]**:** I think it's a group of baseball players standing on top of a grass covered field.
Ground Truth [8]**:** A poster has many photos of players for the Nationals baseball team.
Ultra-Context (ours): A series of photos showing a baseball game.

Figure 1.1: Better context leads to better associations between visual and semantic elements, and hence fewer "hallucinations" of commonly seen concepts

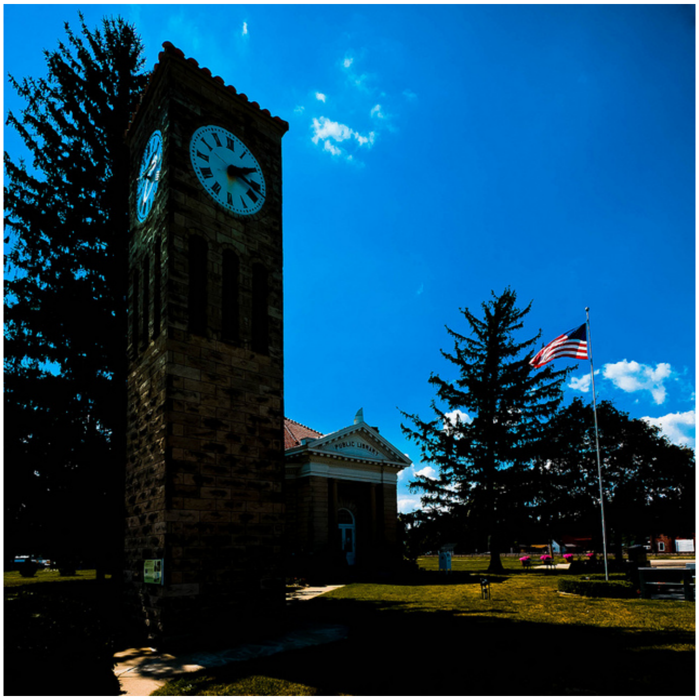
CHAPTER 2

Related Work

Although our approach aimed at maximizing the context can be applied while using any CNN encoder [13, 14, 15], we primarily use bottom-up features [16] in the current work (our baseline model is similar to the Up-Down Captioner [16] in terms of the architecture and is an open source implementation [17]). The reason behind adopting them comes from the observation that objects form a natural basis [18, 19] for visual attention. These encoded features provide salient attention regions (grounded in objects) and a visual representation of each region with additional attribute features to aid the image captioning process, thereby providing the first level of attention context. The basis of selection of region boundaries is well established, hence it becomes imperative that the further process would have a strong initiation context. Conventionally, attention models [2, 6, 20] learn to refine encoded features from different image regions to the most salient ones on the basis of a mechanism which can be spatial attention [6], text-based semantic attention [20], or a combination of both [21]. Although these attention mechanisms provide useful direction towards caption generation, they are still unable to utilize the full extent of the richness of image features received from object recognition architectures [13, 14]. Moreover, many of them fit to frequently observed concepts and fail to generalize well when given a new scenario. It is of even higher importance to address this, since real world images are very diverse.

Realizing the aforementioned deficits and the unprecedented progress observed in object recognition, we propose the **CTX_initemb** and **Ultra_CTX** models, where task specific attentiveness (or 'top-down' attention) of encoded visual features is further combined with n-gram word-specific attentiveness [22] to form a more complete contextual input for the language model to learn further. These models attain context from various angles ((i) initiation context from bottom-up features, (ii) task specific context and (iii) semantic context from coupling of visual and textual features) without incurring the overhead of infusing attributes created from external resources, training additional network structure to fit in existing methods, or externally supervising the training process. In only one pass with a single model (without an ensemble), these context-based models jointly learn where to focus their attention on the basis of the holistic context available to them. They learn to channelize their internal representations, leveraging the visual evidences in the test-time image to form meaningful captions in terms of grammar, saliency, and generalization. As a result, the best model (Ultra_CTX) achieves a score of 28.3 in METEOR [23], 124.2 in CIDEr [24], 58.1 in ROUGE-L [25] and 21.7 in SPICE [26] on the widely used MS-COCO Karpathy split [27] when trained with cross entropy loss and a subsequent Self Critical Sequence Training (SCST) [5] reward function. Fig. 2.1 shows an example illustrating the improved captions that the models generate when compared to the baseline architecture.

The background for our approach comes from a diverse line of works. One of those early works is [6] which suffers from poor region proposals where objects at the boundaries of the grid are not considered sufficiently. Without a strong basis [18] of dividing the regions, critical initiation-point context is lost. They also lack context from semantic word-level attention and might fail to generalize in a lot of scenarios where words form an important basis of saliency. Another shortcoming in such spatial models is that they generally resort to weighted pooling on the feature map, thus leading to spatial information being lost inevitably. However, weighted pooling leads



Baseline: A clock tower with a clock on the top.
CTX_initemb: A clock tower in the middle of a park.
Ultra_CTX: A very tall clock tower with a clock on each of its sides.

Figure 2.1: Evolution from baseline to our first variation, CTX_initemb, and to the full context model, Ultra_CTX

to better generalization to some extent. Thus, more context is required to compensate for the negative aspect of weighted pooling.

Another line of work [21] incorporates semantic concepts, where image features are vectors of confidences of attribute classifiers. Jia et al. [28] exploit the relation between images and their captions as the global semantic information to guide the language LSTM. They require external resources to train these semantic attributes and still lack in maximizing context. Jia et al. introduce guidance for the LSTM but their approach uses pre-specified guidance that is linear and fixed over time. In contrast, [20] systematically incorporates time-dependent text-conditional attention, from 1-gram to n-gram. Theirs was the first model which illustrated the importance of n-gram text-conditional attention. Still, their model lacks region-based initiation context and is therefore prone to “hallucinating” previously seen concepts based on guidance from the partially generated caption. Better context mechanisms would have enabled it to balance between saliency in regions and words. Such a balanced mechanism is necessary to learn new concepts from the features at hand (the image’s visual features and word vectors) rather than fitting itself to seen ones. Additionally, it aids in proper utilization of rich visual features and their attributes while summarizing the image.

SCA-CNN [2] takes advantage of the CNN’s natural ability to obtain channel-wise and spatial attention. They consider channel-wise attention to be similar to semantic attention. However, the context of their architecture still limits it to task specific attention and facilitates only limited associations of visual and verbal concepts.

Another area of related work is Text-Attention [29] which relies on the ground truth captions to be used as a basis of selecting visual features. Following this approach, a model would definitely suffer from the error prone test time sampling where error would build up during captioning and propagate further during re-reference of the caption.

Knowing When to Look [4] utilizes the impact of visually attending to an image only when a salient word is encountered. Again, visual features conditioned on n-gram text features [20] enhances a model with strong textual saliency. Besides, words like ‘of’, ‘the’, and ‘with’ provide useful context for associating visual and verbal concepts. Many times they are necessary to be attended to for more context. Real world images are diverse and no model should rely exclusively on language or visual elements for context to predict the next word. In other words, a balanced learning approach, coupling language and visual elements strongly at different levels within an architecture, is needed to attend to diverse concepts and to generalize well.

Lastly, the Up-Down Captioner [16], focuses primarily on the generation of bottom-up features. However, the overall context achieved from addressing multiple lines of attentiveness could certainly provide better captions.

CHAPTER 3

Method

Addressing the aforementioned shortcomings, a basis of coupling visual and language features is developed by experimenting with: (i) the behavior of word embeddings when added to visual features at different levels of training, and (ii) the correspondence of attention weights, produced by the coupled features, with the output of the caption model.

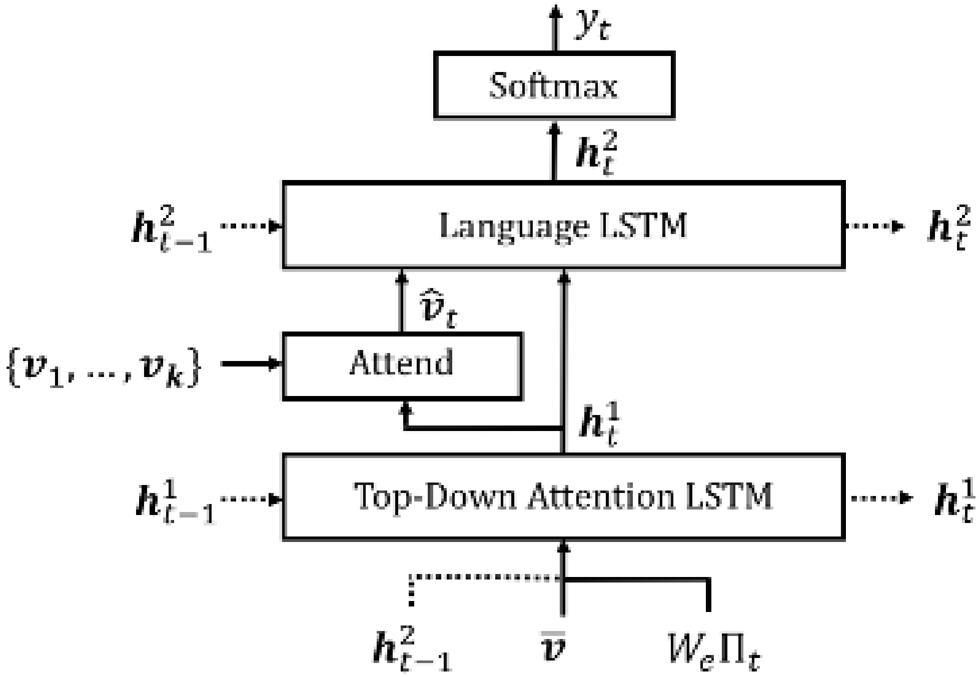


Figure 3.1: Architecture of baseline model taken from Up-Down captioner [16]

In practice, the first context for attention is achieved using bottom-up features [16] relying on RCNN [30] which provides salient region proposals and Resnet-101 [13] that provides visual representations corresponding to each proposed region. We use parallel coupling via (i) visual features coupled with the partially generated caption (which is the encoded representation of the last generated word) and (ii) semantically oriented coupling that distributes and weighs feature representation(s) (mean-pooled image features/spatial image features) over n-gram word vectors (all word vectors corresponding to previously generated words during caption generation). Lastly, we concatenate the coupled outputs to feed to the language LSTM which will have maximum context to produce captions systematically. The architectural considerations we have made for our baseline model are inspired by LRCN [3].

For each image, the number of proposed regions vary up to a maximum of 100. Let V_s denote spatial features, and define mean-pooled spatial features as:

$$\bar{V} = \frac{1}{k} \sum_{i=1}^{k} V_i \tag{3.1}$$

where $1 \leq k \leq 100$. Broadly, the model consists of two LSTM layers: LSTM1 and LSTM2 [10]. The factored case from [3], where the first LSTM layer is shielded from image features and is given the responsibility of representing only the partial caption independent of the visual input, is not used. Rather, our innovation is in the ways to couple the image features with word representations.

3.0.1 Baseline Model

We describe the architecture of our models starting from baseline to our final Ultra_CTX model. Figures 3.2, 3.3, and 3.4 show the architecture of our three models. In all three figures, violet boxes represent learnable components of the model, red

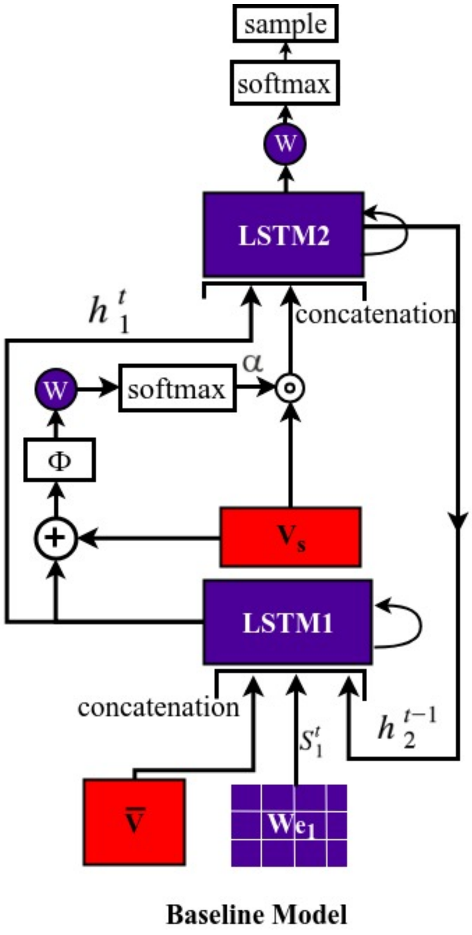


Figure 3.2: Architecture of baseline model redrawn for easy comparison with our improved architectures

boxes show features from the encoder, $\circledcirc$ represents a convex combination, and $\odot$ represents element-wise multiplication.

To introduce the novel couplings that enhance the context, we develop a mathematical basis to it. At each time-step, LSTM1 receives input x_1^t containing mean-pooled image features ($\bar{V}$ from Eq. 3.1), an encoded representation W_{e_1} of the last generated word, and LSTM2's hidden state, h_2^{t-1}, from the last time-step. For h_i^t and x_i^t subscripts denote LSTM layer number and superscripts denote time-step.

$$x_1^t = [h_2^{t-1}, \bar{V}, W_{e_1}^t S^t] \tag{3.2}$$

$W_{e_1} \in \mathbb{R}^{e \times |\Sigma|}$ is the word embedding for our vocabulary Σ, e is the word embedding size and S^t is the one-hot encoding of the word generated by LSTM2 in the previous time-step $t-1$. Word embedding W_{e_1} is learned from scratch and randomly initialized. The hidden state h_1^t of LSTM1 fuses with spatial features V_s as follows:

$$f_s^t = w_f^T tanh(W_{vf} V_s + W_{hf} h_1^t) \tag{3.3}$$

$W_{vf} \in \mathbb{R}^{d \times v}$, $W_{hf} \in \mathbb{R}^{d \times m}$, and $W_f \in \mathbb{R}^d$ are linear layers where d is the number of hidden units in the attention layer and m is the n umber of hidden units in the two LSTMs. These are learned through back-propagation. We use batch normalization for the W_{vf} layer which we have observed leads to faster convergence. Softmax distribution taken at every time-step creates an attention weight for each feature V_s:

$$\alpha = softmax(f^t) \tag{3.4}$$

$$\hat{v}^t = \sum_{s=1}^{k} \alpha_s^t V_s \tag{3.5}$$

This is how optimized attention models [16, 3] frequently define their attention layer, where at any given time-step t, α^t acts as a weight mask for spatial features V_s, and LSTM2 is fed with the current state of LSTM1 concatenated with attended features as follows:

$$x_2^t = [\hat{v}^t, h_1^t] \tag{3.6}$$

Suppose the maximum length of any sequence of words is L. Our aim is to calculate the conditional probability distribution P of the entire sequence of words $(y_1, y_2 ... y_L)$ over the vocabulary. At any time-step t we take the hidden state of LSTM2, h_2^t, apply a linear layer, and calculate the softmax distribution to find the conditional probability as:

$$p(y^t | y^{1:t-1}) = softmax(W_p h_2^t + b_p) \tag{3.7}$$

where $W_p \in \mathbb{R}^{|\Sigma| \times m}$ and $b_p \in \mathbb{R}^{|\Sigma|}$. The joint probability distribution can be computed using the chain rule as:

$$P = \prod_{t=1}^{L} p(y^t | y^{1:t-1}) \tag{3.8}$$

Given a sequence of words $y^{*1:L}$ as ground truth we compute cross-entropy loss as:

$$L(\theta) = -\sum_{t=1}^{L} log(P_\theta(y^{*^t} \mid y^{*^{1:t-1}})) \tag{3.9}$$

where $y^{*^{1:L}}$ is the ground truth sequence and θ is the parameter set of the model.

3.0.2 CTX_latemb Model

This is the first variation of the baseline which uses a 2-step training process where the model is pre-trained as the baseline model before an additional component based on the sentence history is added to increase semantic context.

At any time-step t we have $t-1$ previously generated word(s) (where t starts from 1 and at the first time-step there is no generated word but we just provide a 'START' token). Inputs and outputs for LSTM1 remain the same. For LSTM2, the modification to the input (Eq. 3.6) is as follows:

$$M^t = tanh(\bar{V} \odot W_{e_2} \sum_{i=1}^{t} \frac{S_{i-1}}{t}) \tag{3.10}$$

$$x_2^t = [(\hat{v}^t + M^t), h_1^t] \tag{3.11}$$

where $\odot$ stands for element-wise multiplication, $\bar{V}$ is from Eq. 3.1, $\hat{v}$ from Eq. 3.5, W_{e_2} is a second embedding initialized with all ones, S_i is the one-hot encoding of the i^{th} word, and M^t is the result of coupling mean-pooled spatial features with respect to the sequence of $t-1$ words generated so far.

Intuitively, training further with a new time-dependent mask of word representations should either enhance the context for LSTM2 by allowing focus of attention

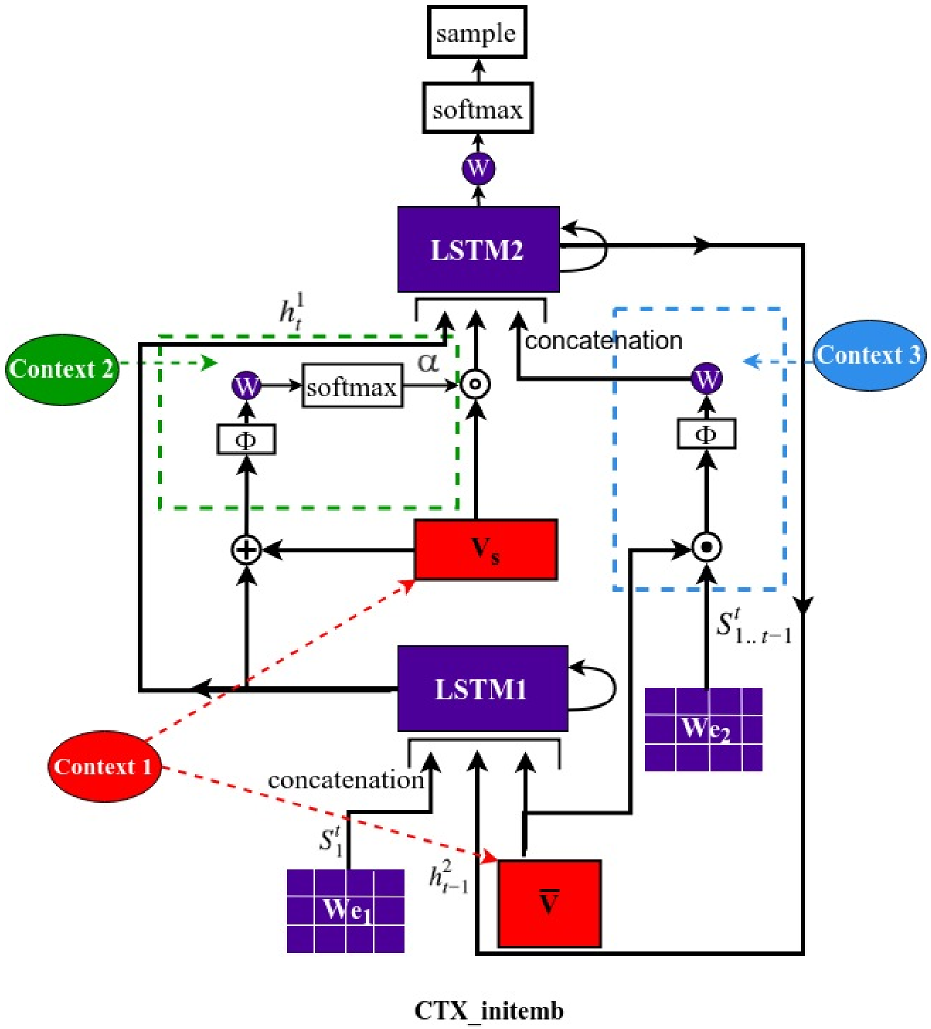


Figure 3.3: Architecture of CTX_initemb model

to be based on the entire history rather than only the most recent word, or lead to some (potentially temporary) performance degradation due to the increased complexity in the architecture. In practice, we found that the model degrades for some time due to the addition of a new untrained embedding and then starts to improve again till a point where it surpasses its initial scores (detailed comparative performance results are shown in Table 5.2). This demonstrates that spatial features should be coupled with word vectors at this specific point since most of the reference LSTM2 has for word representations has to propagate through LSTM1 and is lost or at least strongly "faded" by this point. Furthermore, providing a convex combination of all word representations generated so far with region features contributes significantly to increased semantic context.

3.0.3 CTX_initemb Model

This variant of the baseline utilizes the word sequence to spatial feature coupling, M^t from the very beginning of the training process. In contrast to CTX_latemb, however, it uses it explicitly as an additional input to LSTM2 rather than first combining it with $\hat{v}$:

$$x_2^t = [(\hat{v}^t, M^t, h_1^t] \tag{3.12}$$

This change stems from the observation that the addition operation in Eq. 3.11 between $\hat{v}$ and M from Eq. 3.10, leads to a loss of information. The advantage of this operation in the previous model was that it kept the input dimensions of LSTM2 constant while adding M_t, which is essential for the 2-step training. Here, this need is eliminated since no pre-training using the baseline model is used and we thus concatenate the three entities (Eq. 3.12) to better preserve the contextual information. Results shows that this leads to further improved performance (Table. 5.2).

3.0.4 Ultra_CTX Model

Our final model, Ultra_CTX, streamlines the architecture further to achieve maximum context geared towards better utilization of visual features.

An even stronger coupling is formed by coupling the previously generated word history, $S^t_{1..t-1}$, directly to the spatial features, V_s (rather than the mean-pooled fea-

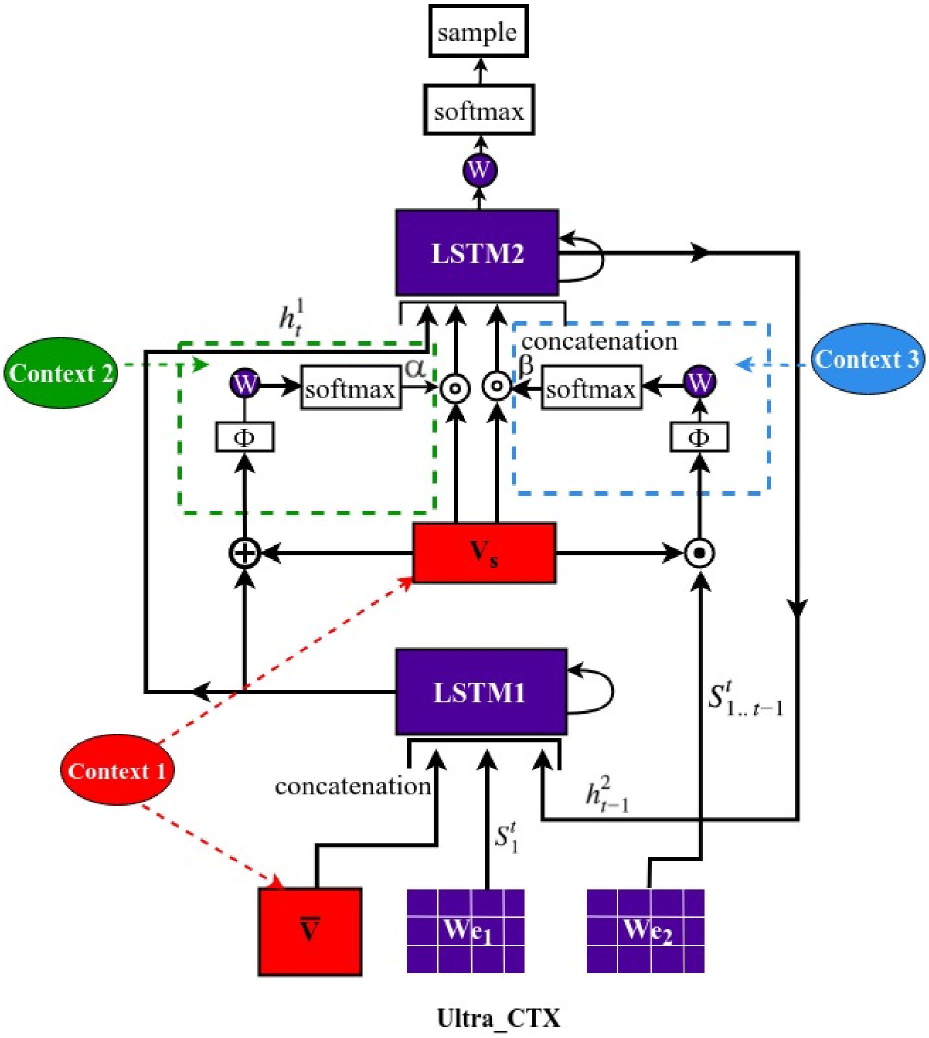


Figure 3.4: Architecture of Ultra_CTX model

tures, $\bar{V}$) through a second weighted mask, β (Eq. 3.4 represents the first weighted mask):

$$C_s^t = W_c^T tanh(V_s \odot W_{e_2} \sum_{i=1}^{t} \frac{S_{i-1}}{t}) \tag{3.13}$$

$$\beta^t = softmax(C^t) \tag{3.14}$$

where $W_c \in \mathbb{R}^{d \times 1}$ and β is the weighted mask on spatial features V_s conditioned on word vectors (encoded representation of words from time-step 1 to $t-1$). Similar to the previous variants, this model also uses a trainable embedding, W_{e_2}, initialized with all ones. Applying the weighted mask on V_s to form the input of LSTM2 yields:

$$\hat{v}_c^t = \sum_{s=1}^{k} \beta_s^t V_s \tag{3.15}$$

$$x_2^t = [\hat{v}^t, h_1^t, \hat{v}_c^t] \tag{3.16}$$

Intuitively, each salient region has a set of features which should be coupled with word representations to receive more context. In other words, the model should learn to attend to the visual features, balancing between the three different levels of context (initiation context from bottom-up features, task specific context, and semantic context).

CHAPTER 4

Analysis

4.0.1 Qualitative Analysis

To illustrate the findings and evaluate the two carefully crafted couplings qualitatively, we look specifically: (i) at the learned embeddings, W_{e_1} and W_{e_2}; (ii) at our final model's ability to attend to correct regions by analyzing the weighting maps, α and β (Eq. 3.4 and Eq. 3.14) produced by coupling the features; and (iii) at the semantic construct of captions.

4.0.1.1 Embedding Analysis

Dense word embeddings can be successful in capturing semantic relations among words. However, in many cases, their semantic structure is heterogeneously distributed across the embedding dimensions which makes it hard to interpret them [31, 32]. We aim to bring light to the semantic concepts implicitly represented by various dimensions of a word embedding in order to establish an intuitive insight into how they could be meaningfully coupled with image visual features to provide more context. In our exploration, we refer to the category theory [33] and construct a KNN adjacency matrix, A_m, such that A_m is a $\Sigma \times \Sigma$ matrix (Σ is our vocabulary size) containing the pair-wise similarities among representations within the word embedding. The euclidean distance between two points in the matrix A_m, representing distances between words, is used to calculate the closest distances to each word in the vocabulary. We find that the associations seen in the generated captions are reflected in the nearest words. An interesting observation is that both embeddings learned different associations. The first embedding, W_{e_1}, seems to learn to associate words with their prepositions and relatively few related verbs/nouns, while the second, W_{e_2}, seems to learn to associate mostly nouns, adjectives, and verbs (Table 4.1).

Word	N_1	N_2	N_3	N_4	N_5
cat	a	of	on	with	in
	cats	several	hat	sitting	black
car	a	with	in	to	building
	cars	power	traffic	parked	bikers
frisbee	a	next	to	on	flying
	umbrella	park	fire	next	hydrant
bananas	of	UNK	to	a	stairway
	pizza	eating	yellow	surrounded	table

Table 4.1: Five nearest words (N_1-N_5) in the two embeddings of the Ultra_CTX model for randomly selected words. For each word, row 1 and 2 represent W_{e_1} and W_{e_2}, respectively

The CTX_latemb model, even with the introduction of the second embedding at a later stage of training (Sec. 3.0.2) manages to form meaningful associations within the new embedding. We found that even in this case the commonly seen words had dense representations and relatively large deviations from the initial value whereas words found rarely within captions stayed closer to their initialization. As the models moved progressively to a more tightly integrated second embedding, advancing through CTX_initemb to Ultra_CTX, we observed an increased structuring of the representation space, indicated by increasing deviations from the initialization values. To further examine the embedding, we tried different embedding sizes (512, 1000, and 1024) for the two embeddings. We found that a ratio of 1:1 between the LSTM size and the embedding size generally works best in terms of the model's convergence and the quality of the representations inside the embedding.

Besides improvement in word relations, a slight enhancement in grammar was also observed. We have not performed any extensive analysis to examine this and it is thus currently purely anecdotal based on qualitatively observing captions.

4.0.1.2 Attention Weight Analysis

Analyzing focus of attention mechanisms in deep neural network architectures is a complex task, however, it also provides crucial information for the understanding of the operation of the architecture. To perform this analysis we analyzed the Ultra_CTX model with respect to the α and β attention weight masks leveraging box coordinates [16] corresponding to the weight masks. The resulting sum of α and β gave the final weight of the region where the model focuses. Here β has a supplementary effect by shifting the α weights and its value alone is not analyzable.

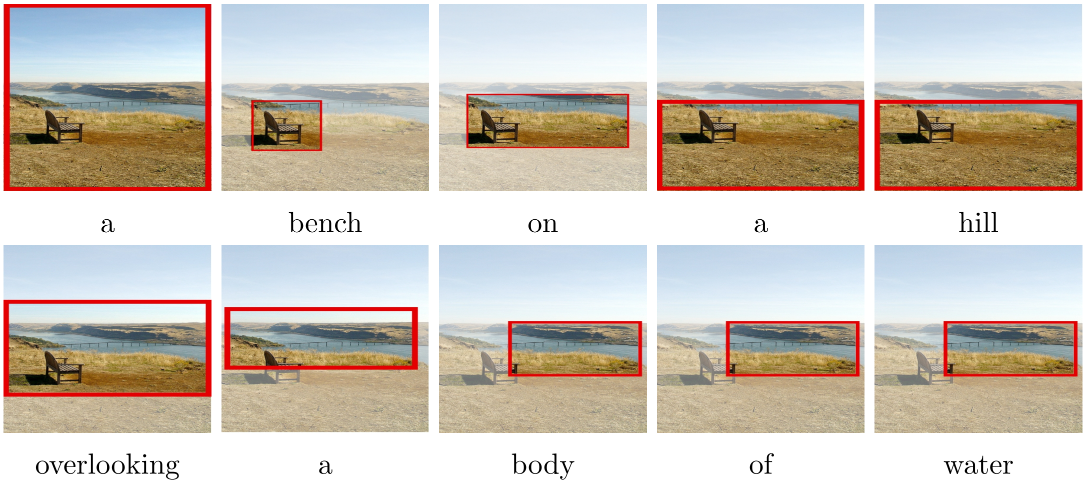


Figure 4.1: Example analysis of the Ultra_CTX model during caption generation. Red boxes show the dominant region proposals at the corresponding time step as identified from the α and β weights, illustrating the spatial focus of the network during the generation of the word indicated below the corresponding image frame

Using this analysis method, we found that the network focused on relevant regions that closely resemble the closer context of the word being generated at each time-step. Fig. 4.1 shows an example image and generated word sequence with the corresponding attention region highlighted. It is critical that a model learns to ground the generated caption in objects and other salient image features it considers. Otherwise it might be fitting to the data without learning to generalize to new unseen images. Our model is successful at finding meaningful associations between words, objects, object's attributes and surroundings, and at weighing them properly while describing images.

4.0.1.3 Novel Semantic Constructs

In the example (COCO_val2014_000000005820.jpg) in Fig. 4.1, the caption generated by our Ultra_CTX model trained using cross-entropy loss generalizes the image quite well and produces novel semantic constellations. There are 1,843 instances in the 2014 Train/Val annotations from MS-COCO where the phrase, "a body of water" is used, out of which only two cases are similar to the one shown in Fig. 4.1. Our model learns from these examples how to associate words, objects, and object's attributes and surroundings. Furthermore, it is also capable of generalizing the deciphered associations quite well (also evident from the ability to attend to the correct regions). Fig. 2.1 shows another example where the architecture produces a semantic relationship of the same form and context not present in the dataset.

In another example in Fig. 4.2 our model produces the caption, "a man riding a skateboard down a street with cones". The image (COCO_test2014_000000194910) is taken from the MS-COCO [8] 2014 test images. In this case, there is not a single image in the entire training-set where a caption combines the skateboarding, street, and cone context. There is only a single instance where the phrase, "street with

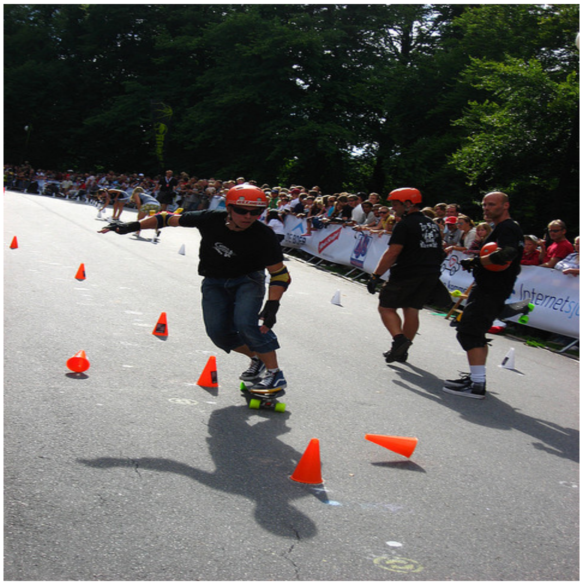


Figure 4.2: Novel semantic construct: “A man riding a skateboard down a street with cones”

cones” is used in the entire dataset. This highlights our architecture’s strength to be able to learn from a variety of instances and produce novel semantic constructs. We observed many such instances before coming to the inference.

4.0.2 Implementation Details

Our CTX_initemb model has two LSTMs each with the number of hidden units, m, set to 1,000 units. The two word embeddings have a hidden size e of 1,000 units. Hidden size d of 512 units is used for the attention layer. We perform minimal text pre-processing by tokenizing on white spaces, converting every word into lower case, and filtering out words that occur less than 5 times. Finally, a vocabulary of 9,487 words is formed. Captions are trimmed to a maximum of 16 words for computational efficiency.

ADAM with amsgrad [34] optimizer is used. Batch size was chosen to be 100 with an initial learning rate of 0.0005 which is lowered at a rate of 0.8 after every

3 epochs starting from epoch 10. Scheduled sampling [35] is used while training with cross-entropy loss, starting from the beginning of the training process. We start with 0 and increase the sampling probability by 0.05 every epoch until it reaches a maximum of 0.25. We trained on cross-entropy loss for 33 epochs and subsequently used the SCST approach for 20 additional epochs. While training with SCST, our learning rate was set to 0.00005 with a decay rate of 0.5 after every 3 epochs starting from epoch 43. Training the model fully takes one whole day on a NVIDIA Tesla P100 GPU. During optimization, beam size was set to 5.

For our Ultra_CTX model the only changes are: LSTM size of 2,048 units (for both LSTMs), input encoding size of 1,024 units (for the two word embeddings), and hidden attention size of 1,000 units. Due to the large size of the network, we used 2 NVIDIA Tesla P100 GPUs in parallel to train the model. We noticed that the model converges relatively fast and reaches its plateau in 25 epochs. Similarly, during CIDEr optimization it converges in just 15 additional epochs. We found that the architecture of the Ultra_CTX model was robust to the increase of the size of the LSTMs from 1,000 to 2,048 without over-fitting the data, which is due to the architectural strength due to carefully crafted couplings.

4.0.3 Automatic Evaluation Metrics

While analyzing the captions generated by our different models during local evaluation (`https://github.com/tylin/coco-caption`), the shortcomings of evaluation metrics becomes noticeable. During CIDEr optimization, the metric is directly optimized with an objective of minimizing the loss:

$$LF_R(\theta) = -E_{y^{1:L} \sim p_\theta}[r(y^{1:L})] \tag{4.1}$$

where the parameters of the network are given by θ, and r is the score function (CIDEr). From the method described in SCST [5], the gradient for this loss function LF can be approximated as:

$$\nabla_\theta LF_R(\theta) \approx -(r(y_s^{1:L}) - r(\hat{y}^{1:L}))\nabla_\theta logp_\theta(y_s^{1:L}) \tag{4.2}$$

where $y_s^{1:L}$ is a sampled sequence of words and $r(\hat{y}^{1:L})$ is a greedily decoded score from the current model.

While producing captions, the margin of mistakes made by a model trained with Reinforcement Learning (RL) is significantly reduced. However, the model gets trained to predict captions that score high (on CIDEr) without any regards to the actual content of the image. In other words, SCST relies on the assumption that metrics are efficient at mimicking human behavior of evaluation while it can be safely said that they are inadequate to an extent [36]. Apart from that, RL-based approaches are language oriented and thus fail to ground words in visual composition of an image.

Figure 4.3 shows the caption generated by 3 models: 1. Ultra_CTX model at CIDEr 114.2, 2. Baseline model at CIDEr 110.2, and 3. Resnet101 baseline model at CIDEr 107.6, all trained on cross-entropy loss. The image is taken from MS-COCO val2014 dataset (COCO_val2014_000000104392.jpg) and has one of its ground truth caption as, "a modern kitchen with an oven, stove, and a fridge" [8]. The Caption generated by our model covers three items correctly while conventional attention models describe only two, out of which one may even be wrong (there is no sink in the image). Even then, the caption from the Ultra_CTX model scores the lowest in every metric. There are many similar images in the dataset where kitchen appliances/furniture is shown, or a bathroom is shown with fixtures. Our model captions tend to include more than two items in many cases and gets penalized similar to the case shown in Figure 4.3. This results in overall low scores (especially on CIDEr and SPICE) even when the caption quality is actually high.

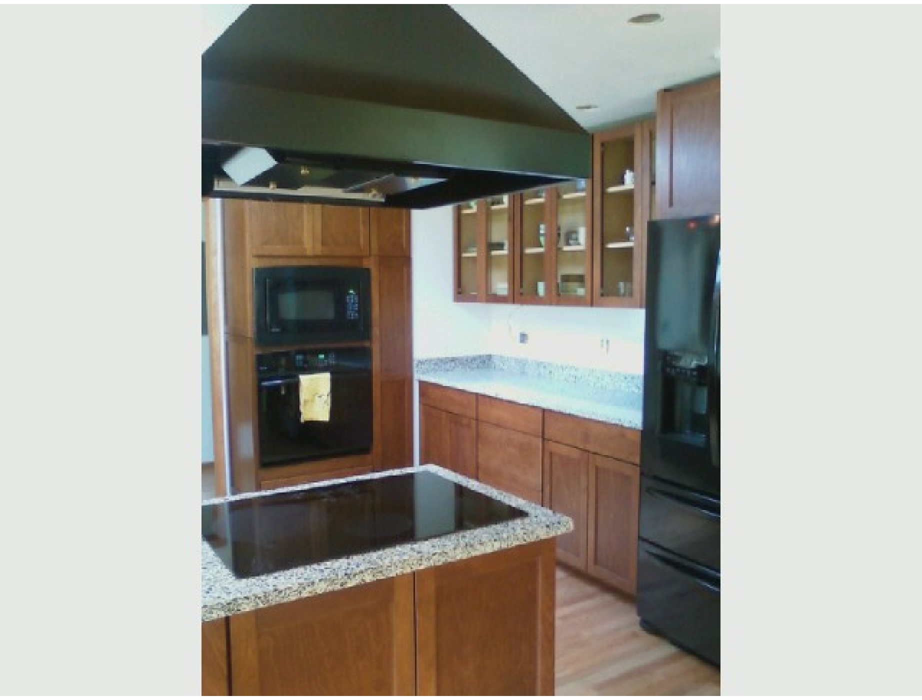


Figure 4.3: Analysis of captioning evaluations by automatic evaluation metrics. C stands for CIDEr, M for METEOR, R for ROUGE-L and S for SPICE

Ultra_CTX: a kitchen with a stove a microwave and a counter
Scores: C: 68 M: 22 R: 60 S: 17
Baseline RL: a kitchen with a stove and a table
Scores: C: 106 M: 25 R: 65 S: 19
Resnet101_baseline: a kitchen with a stove and a sink
Scores: C: 96 M: 25 R: 65 S: 19

Similar deficiencies have been pointed out by others:

"CIDEr is designed as a specialized metric for image captioning evaluation, however, it works in a purely linguistic manner, and only extends existing metrics with tf-idf weighting over n-grams. This sometimes causes unimportant details of a sentence to be weighted more, resulting in a relatively ineffective caption evalua-

tion" [37] and *"First, we observe that all the scores decrease when some words are replaced with their synonyms. The change is especially significant for SPICE and CIDEr"* [37].

We observe the same phenomenon where our model produces a lot of synonyms and sometimes slightly different semantic constructs (Sec. 4.0.1.3) while preparing the captions (due to the novel coupling between spatial features and words representations) and is penalized by the metric in many such cases. Our sole purpose of specifying the limitations and drawbacks of automatic captioning metrics (even after achieving state-of-the-art results on the same) is to suggest that there is a dire need to attend to the way we evaluate the caption quality and also to emphasize that our future work will be towards the same.

CHAPTER 5

Results

5.0.0.1 Resnet-101 Features

To demonstrate the wide applicability of our approach we use Resnet-101 features and verify the performance of our model on the Karpathy split [27]. Similar to the method we followed for bottom-up features, the mean-pooled image features are substituted with the ones from the final convolutional layer of Resnet-101 pre-trained on ImageNet [7]. To obtain spatial features we follow the approach used in SCST [5] and use bilinear interpolation to form fixed size spatial representations of 10×10 which is equal to the maximum number of spatial regions we could use with bottom-up features. The results are shown in Table 5.1.

Model	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Up-Down_R101 [16]	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
Ultra_CAP101	**77.8**	**34.8**	**27.3**	**56.5**	**111.3**	**20.1**	**79.2**	**35.3**	**27.7**	**57.2**	**116.8**	**21.2**
	Cross-Entropy Loss						CIDEr Optimization					

Table 5.1: Comparison of results between our final architecture labeled as Ultra_CAP101, and Up-Down Captioner's Resnet baseline labeled as Up-Down_R101, when Resnet-101 features are used. B-1 stands for BLEU 1, B-4 for BLEU 4, M for METEOR, C for CIDER, R for ROUGE-L and S for SPICE

5.0.0.2 Bottom-Up Features

In Tables 5.2 and 5.3 we compare the performance of our four models on the Karpathy split [27] which distributes validation and training images from the MS-

COCO training-set into a split of 5,000 images for validation, 5,000 for testing, and 113,287 for training. Table 5.2 shows the performance after cross-entropy training while Table 5.3 shows the performance after further CIDER optimization.

Model	B1	B4	M	R	C	S
Baseline	76.5	34.3	26.8	56.1	110.4	19.9
CTX_latemb	76.7	34.5	26.9	56.3	111.2	20.0
CTX_initemb	76.9	35.3	27.0	56.5	112.0	20.1
Ultra_CTX	**77.7**	**35.5**	**27.1**	**56.9**	**114.2**	**20.5**

Table 5.2: Results of training our models on the Karpathy split using the cross-entropy loss function. B1 stands for Bleu-1, B4 for Bleu-4, M for METEOR, R for ROUGE-L, and S for SPICE

Model	B1	B4	M	R	C	S
Baseline	78.4	36.1	27.5	57.1	117.8	20.8
Late_emb	78.8	36.5	27.9	57.6	119.2	21.1
Init_emb	79.2	37.3	28.1	58.0	122.7	21.5
Ultra_CAP	**81.1**	**39.3**	**28.8**	**58.9**	**126.3**	**22.0**

Table 5.3: Results on Karpathy split after CIDER optimization. B1 stands for Bleu-1, B4 for Bleu-4, M for METEOR, R for ROUGE-L, and S for SPICE

Here the CTX_latemb model is mainly used as a means to signify the importance of providing context through tightly coupled features since it allows to directly see the improvement due to the added second, increased context training phase when compared to the baseline model that was used for pre-training in the first phase. The improvement observed is suggestive of the fact that dense embeddings can learn semantically meaningful representations when properly coupled with spatial features. In general, however, it is not recommended to couple the features with the additional embedding when it is already trained, as illustrated by the better results achieved using the CTX_initemb model which does not use any pre-training. In Tables 5.1, 5.2, and 5.3, abbreviation B1 stands for Bleu-1 [41], B4 for Bleu-4 [41], M for METEOR [23], R for ROUGE-L [25], and S for SPICE [26]. In all of our experiments, we do not form ensembles of our models but instead use only single network systems to achieve state-of-the-art results. Table 5.4 shows a comparison between our final model and a range of state-of-the-art models on the Karpathy split. Our final model here outperforms all other models.

Model	Bleu1	Bleu2	Bleu3	Bleu4	METEOR	ROUGE-L	CIDER-D	SPICE
Soft Att [6]	71.8	50.4	35.7	25.0	23.0	-	-	-
TextATTResnet [29]	74.9	58.1	43.7	32.6	25.7	-	102.4	-
Jia-glstm [28]	67.0	49.1	35.8	26.4	22.74	81.25		
E2Eglstm [20]	71.6	54.5	40.5	30.1	24.7	-	97.0	-
Adaptive [4]	74.2	58.0	43.9	33.2	26.6	-	108.5	-
SCST: Att2in [5]	-	-	-	34.8	26.9	56.3	115.2	-
Semantic-ATTFCN [21]	70.9	53.7	40.2	30.4	24.3	-	-	-
Up-down [16]	79.8	-	-	36.3	27.7	56.9	120.1	21.4
MLAIC [38]	80.7	63.9	49.0	36.9	27.7	57.5	119.1	-
STACK-CAP [39]	78.6	62.5	47.9	36.1	27.4	56.9	120.4	20.9
Ultra_CAP	**81.1**	**65.6**	**51.1**	**39.3**	**28.8**	**58.9**	**126.3**	**22.0**

Table 5.4: Performance of our final model on MSCOCO Karpathy split. Bold figures represent the scores of our models. † is used to denote the use of ensembles of several differently initialized models. Our single models (without ensemble) outperform other state-of-the-art models by a significant margin

5.0.1 Microsoft-COCO Leaderboard

CTX_initemb and Ultra_CTX models are tested on the MS-COCO competition leaderboard (Table 5.5).

Both of our models outperform Up-Down Captioner and SCST-Att2all ensemble models in terms of METEOR, ROUGE-L, and CIDEr scores using only a single model. We used MS-COCO 2014 training and validation set (123,287 images) to train our

Model	Bleu-1		Bleu-2		Bleu-3		Bleu-4		METEOR		ROUGE-L		CIDER-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
LSTM-A3† [40]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Stack-Cap† [39]	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3
Up-Down† [16]	**80.2**	**95.2**	**64.1**	**88.8**	49.1	**79.4**	36.9	**68.5**	27.6	36.7	57.1	72.4	117.9	120.5
SCST-Att2all† [5]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
CTX_initemb	79.5	94.3	63.7	87.9	49.1	78.6	37.2	67.8	28.0	37.0	57.7	**72.6**	119.8	121.3
Ultra_CTX	79.8	94.4	64.0	88.0	**49.3**	78.6	**37.3**	67.6	**28.1**	**37.0**	**57.8**	72.5	**121.8**	**124.1**

Table 5.5: MS-COCO test server results. Bold figures represent the highest scores within the table. † is used to denote the use of ensembles of several differently initialized models

two models for the MS-COCO test server [1]. The Ultra_CTX model achieves a margin of ≈ 1% on ROUGE-L, ≈ 3% on CIDEr-D and ≈ 2% on METEOR over the state-of-the-art Up-Down captioner (an ensemble of 4 models) [16] with just a single model. Note that Bleu [41] initially proposed for machine translation, is based on explicit word matching (n-gram matching). It fails to spot semantic similarity when common words are scarce and is affected by word vocabularies [37].

5.0.2 Qualitative Results

We utilize the open source implementation [17] of the Up-Down Captioner to compare and contrast the improvements in our final Ultra_CTX model. We demonstrate unique features of our model in terms of the quality of captions it produces. More specifically, we present cases where the model describes: (i) more than two items present in an image (Fig. 5.1), (ii) the presence of black and white photos (Fig. 5.2), (iii) 'out of context' images (Fig. 5.3), (iv) the images with captions comprised of novel semantic constructs generated due to fewer "hallucinations" and more visual understanding (Fig. 5.4, 5.5, and 5.6), and (v) the images with captions showing better utilization of rich visual features (Fig. 5.7). Next, cases where the model fails to generalize well (Fig. 5.8), are shown. Furthermore, Fig. 5.9 highlights our model's ability to shift its context focus to the word being generated and Fig. 5.10 includes instances relating to the skateboard context where our model successfully understands the context and associates words, objects, object's attributes and object's surroundings. Finally, we present additional examples (Fig. 5.11) demonstrating the quality of captions produced by our model:

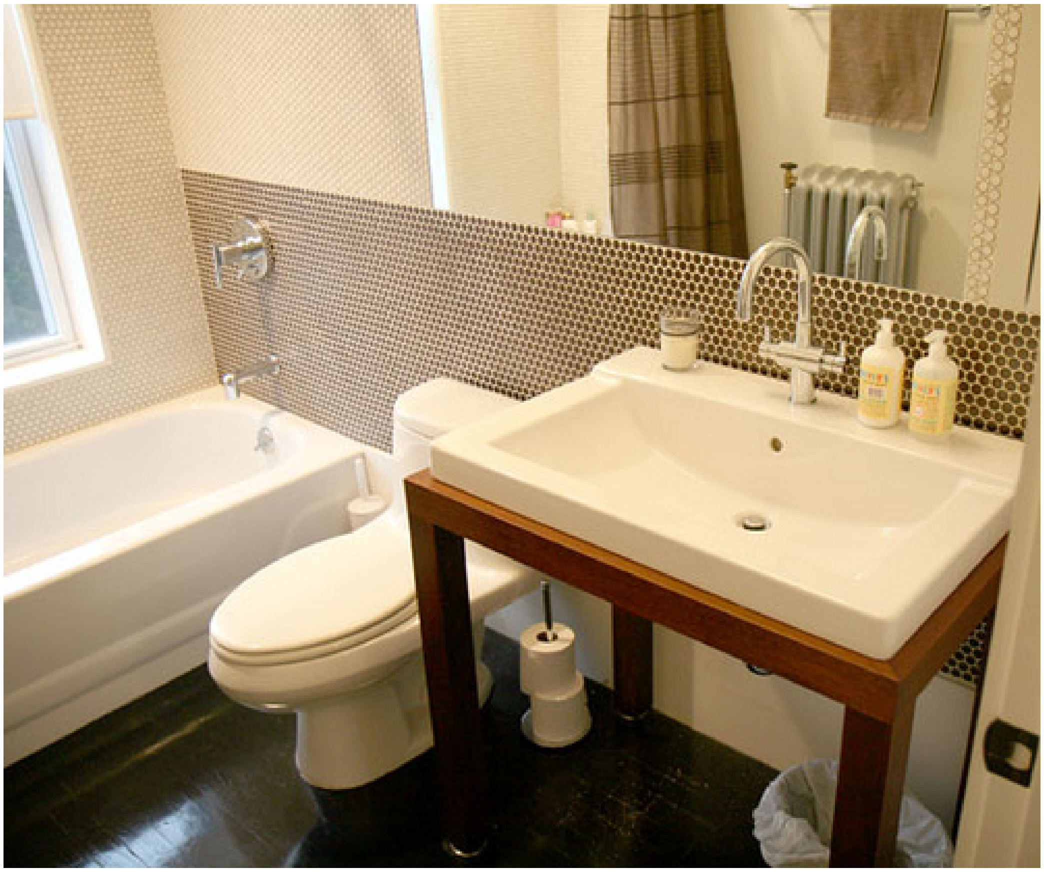


Figure 5.1: Detection of more than two items in images

Up-Down: A bathroom with a toilet and a sink.
Ultra_CTX: A bathroom with a sink a toilet and a bathtub.
GT: A bathroom with a bathtub next to a white toilet and a sink.

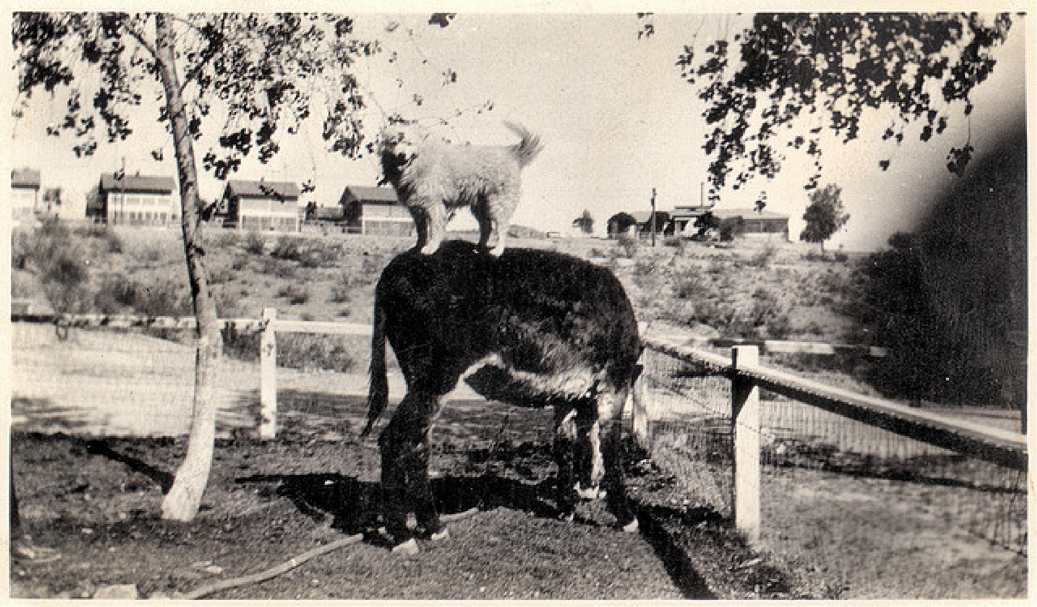


Figure 5.2: Detection of black and white photo frames

Up-Down: A cow and a sheep are standing in a field.
Ultra_CTX: Black and white photo of a cow and sheep standing on a field.
GT: Black and white photograph of a dog standing on a cow's back.

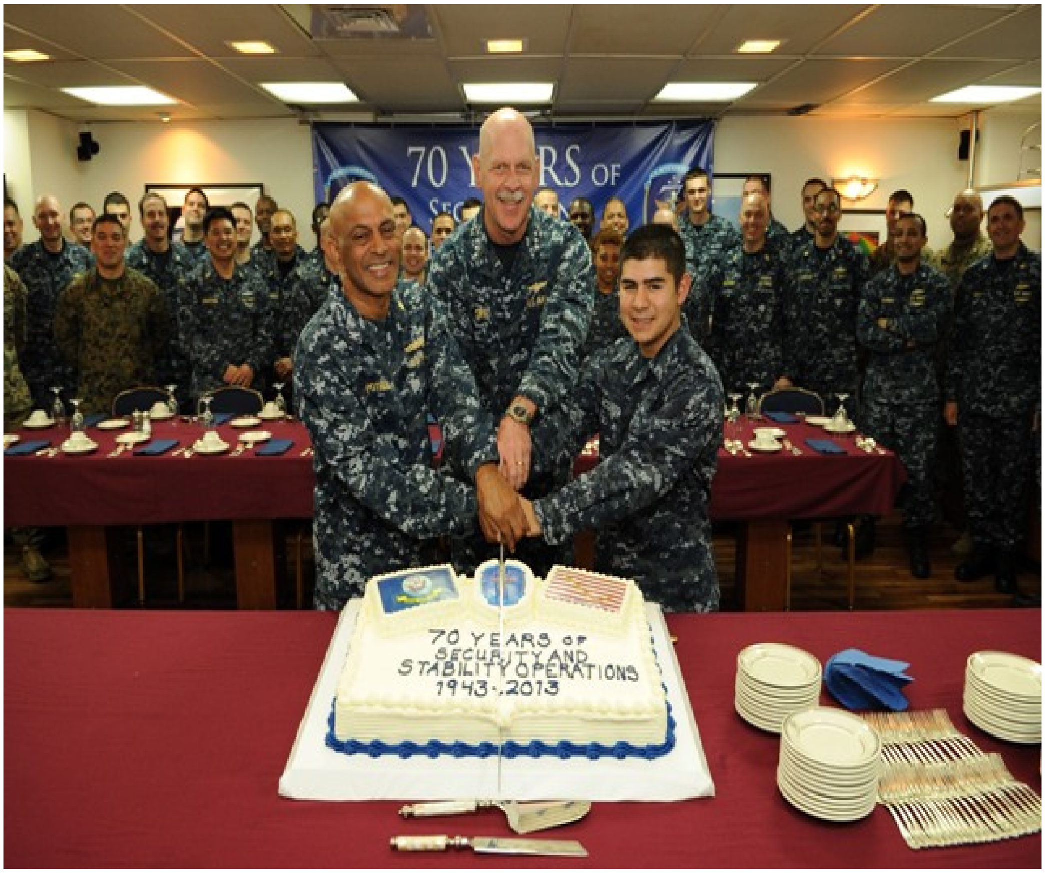


Figure 5.3: An example of the model describing a less common concept is shown. Generally, soft attention mechanisms tend to "hallucinate" a knife for cutting the cake but our model is relatively more closely grounded in what it sees

Up-Down: A group of people standing around a cake.
Ultra_CTX: A group of soldiers cutting a cake with a sword.
GT: A group of solders cutting up a sheet cake.

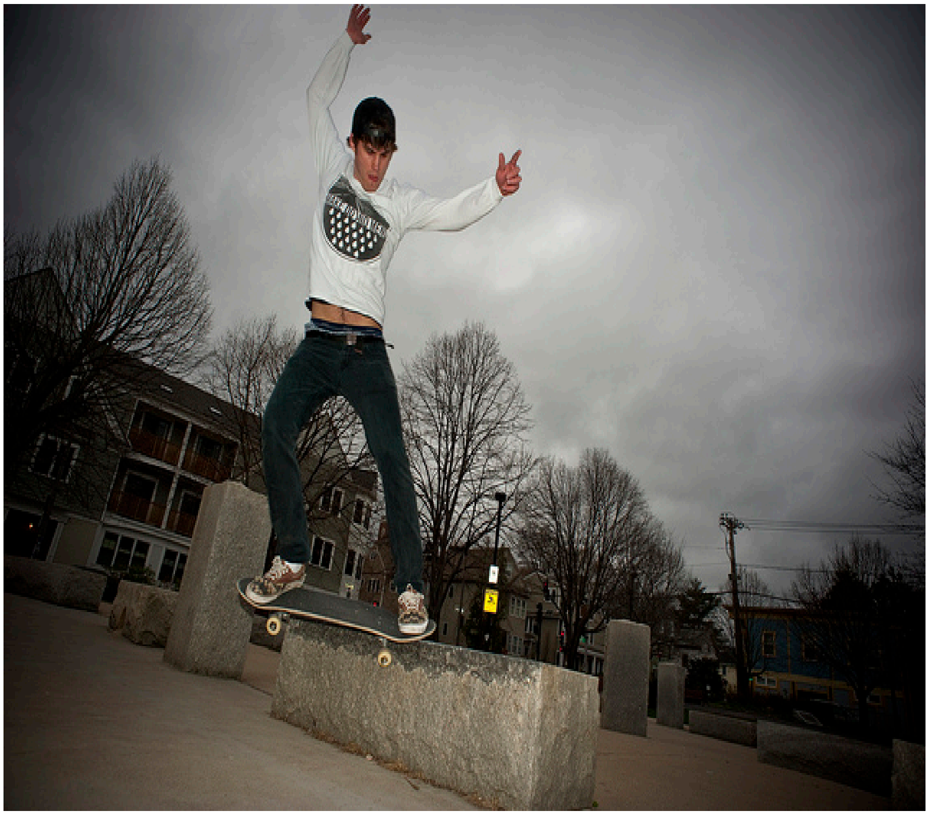


Figure 5.4: There are many instances in the dataset where the model sees a cement block and the skateboarders doing tricks but mostly in different contexts. Still, it learns to associate them well.

Up-Down: A man doing a trick on a skateboard.
Ultra_CTX: A man riding a skateboard on top of a cement block.
GT: A skateboarder riding on a park bench, on a cloudy day.

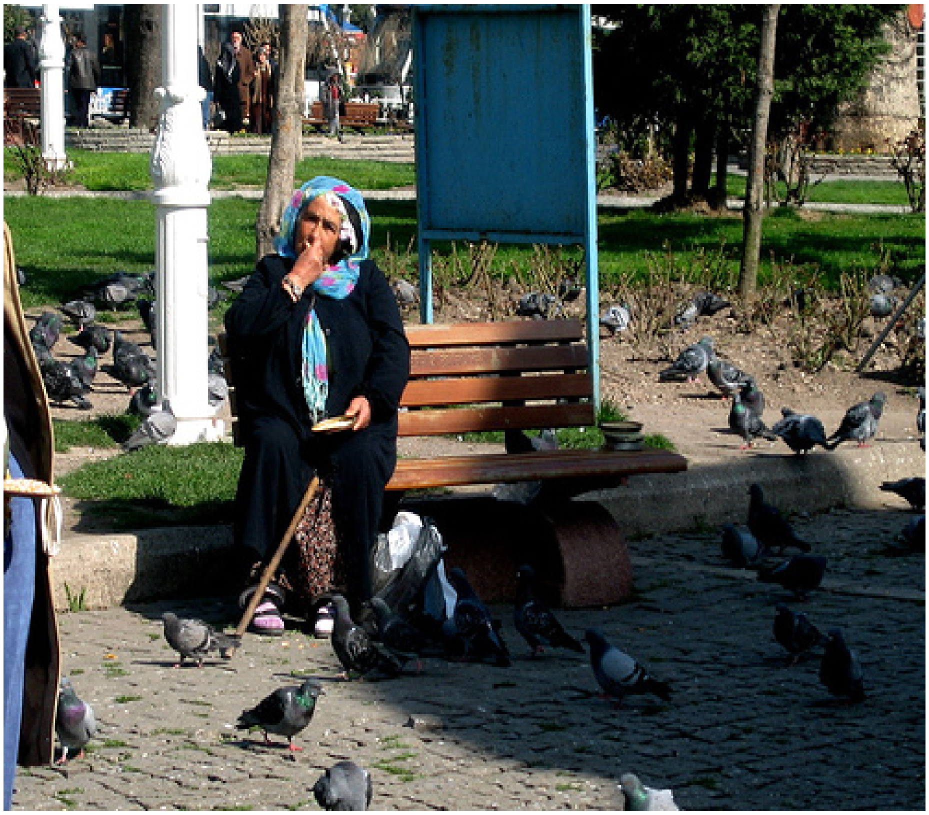


Figure 5.5: The model describes the position of the women relatively more precisely in relation to the presence of pigeons within the image

Up-Down: A woman sitting on a bench next to pigeons.
Ultra_CTX: A woman sitting on a bench surrounded by a flock of pigeons.
GT: The woman sits on a wooden bench near a large flock of pigeons.

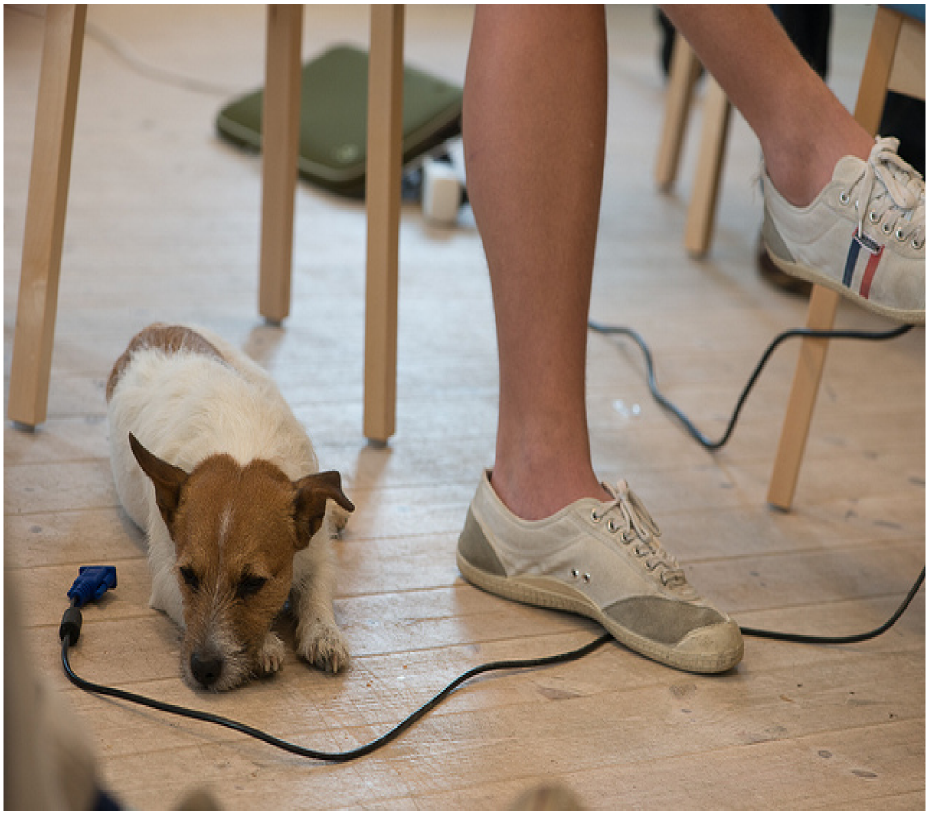
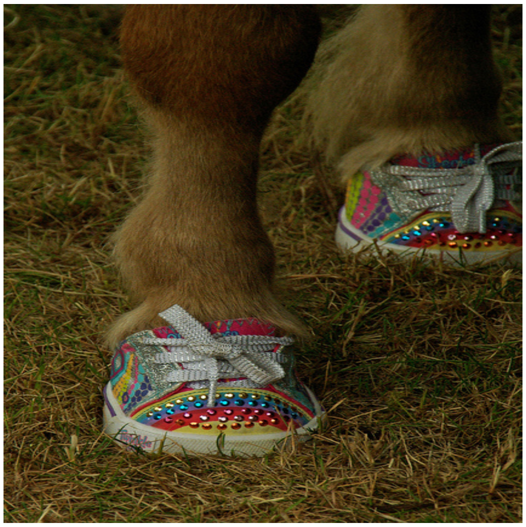


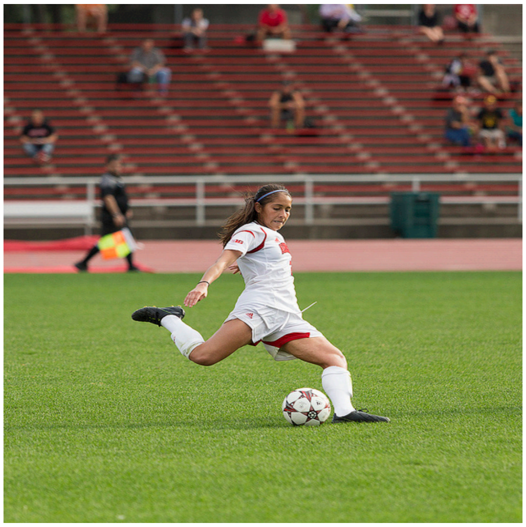
Figure 5.6: Another example where the model does not "hallucinate" commonly seen concepts and forms novel sentence in terms of semantics

Up-Down: A dog is standing on a floor with a leash.
Ultra_CTX: A dog is laying on the floor next to a persons feet.
GT: A little dog laying under the table at someones feet.



A failure case where the model makes an intelligent guess that is better than the Up-Down model

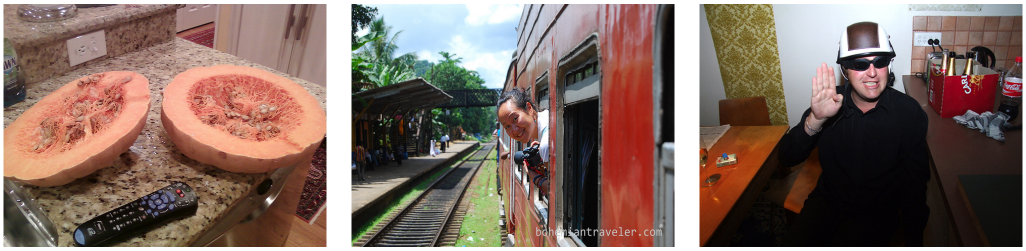
Up-Down: A dog wearing a tie and a sweater.
Ultra_CTX: A dog wearing a pair of shoes.
GT: Animal with hoofs made to wear colorful pair of shoes.



The model describes the color of the jersey along with the girl's current action

Up-Down: A woman kicking a soccer ball on a field.
Ultra_CTX: A girl in a white uniform kicking a soccer ball.
GT: A young lady kicking a soccer ball on a field.

Figure 5.7: Two examples are shown where the model utilizes the image visual features to a better extent

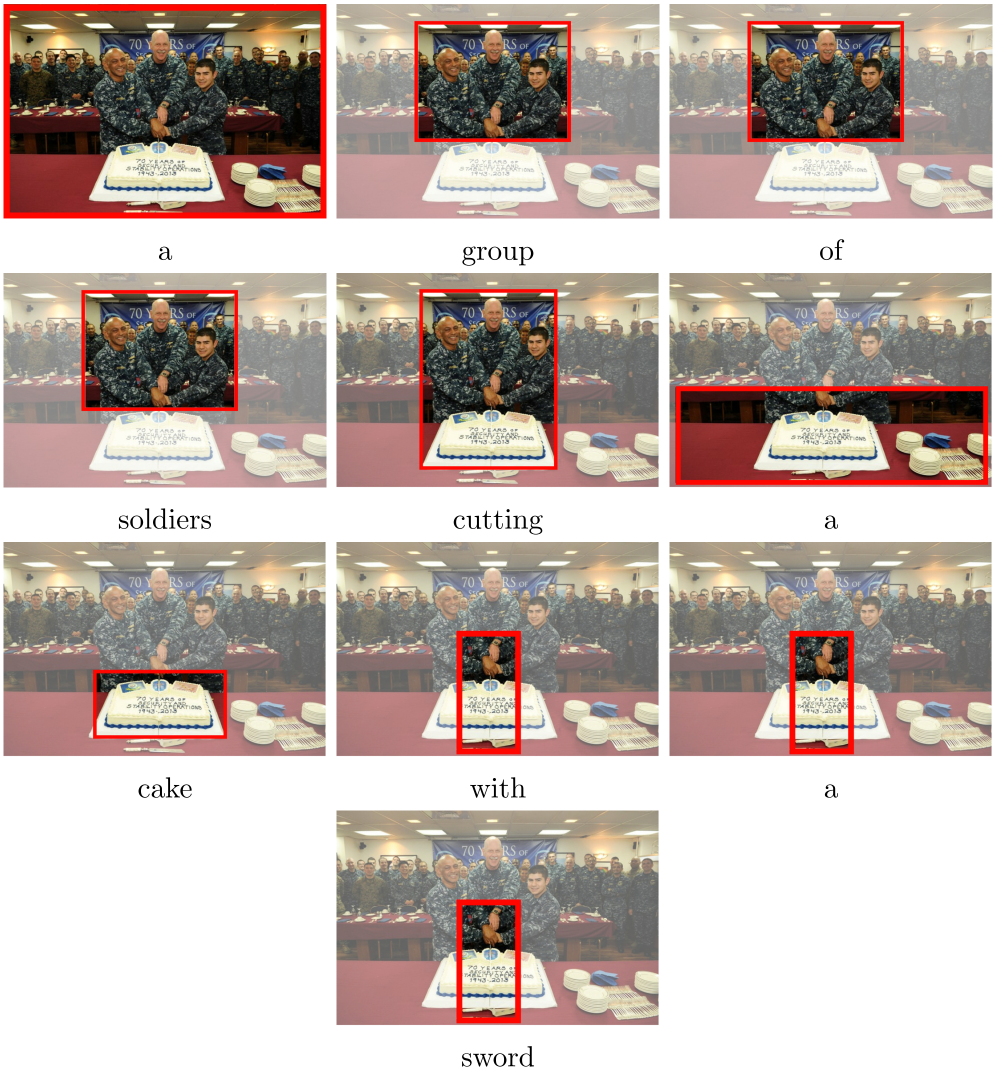


Up-Down: Two uncooked pizzas sitting on a counter top.
Ultra_CTX: A pizza sitting on top of a cutting board.
GT: The halved melon is on the counter next to the remote.

Up-Down: A woman is looking at her phone in front of a train.
Ultra_CTX: A woman is standing in front of a train.
GT: A woman looking out the window of a train.

Up-Down: A woman wearing a helmet and holding a cup.
Ultra_CTX: A woman wearing a helmet and glasses holding a knife.
GT: A man dressed in a helmet and goggles indoors has a goofy smile on and his hand raised.

Figure 5.8: Examples of failure cases



a

group

of

soldiers

cutting

a

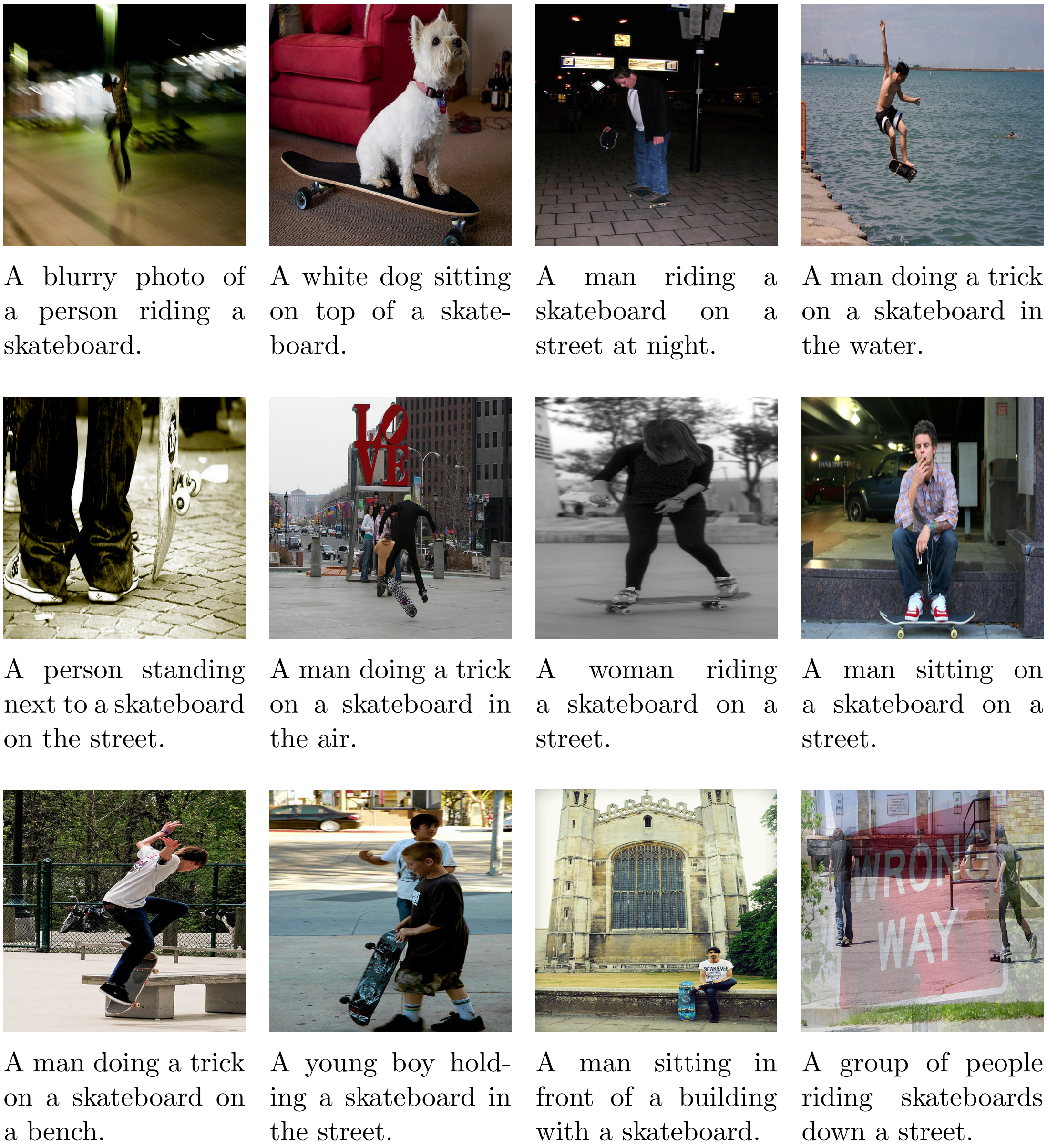
cake

with

a

sword

Figure 5.9: Example analysis of the Ultra_CTX model during caption generation. Red boxes show the dominant region proposals at the corresponding time step as identified from the sum of α and β weights, illustrating the spatial focus of the network during the generation of the word indicated below the corresponding image frame



A blurry photo of a person riding a skateboard.

A white dog sitting on top of a skateboard.

A man riding a skateboard on a street at night.

A man doing a trick on a skateboard in the water.

A person standing next to a skateboard on the street.

A man doing a trick on a skateboard in the air.

A woman riding a skateboard on a street.

A man sitting on a skateboard on a street.

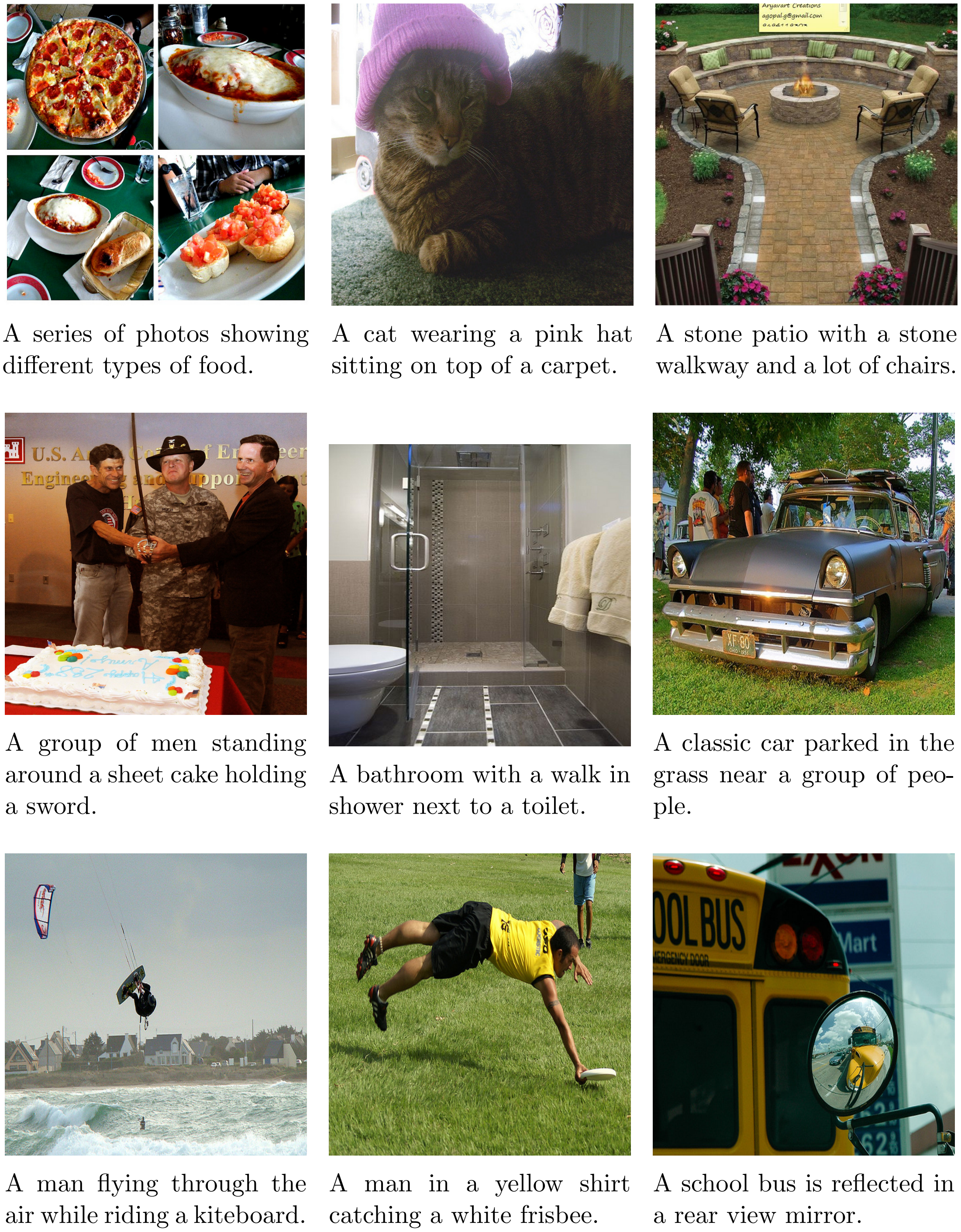
A man doing a trick on a skateboard on a bench.

A young boy holding a skateboard in the street.

A man sitting in front of a building with a skateboard.

A group of people riding skateboards down a street.

Figure 5.10: A series of images highlighting our model's ability to associate words, objects, object's attributes and object's surroundings. We specifically chose the skateboard related context with diverse surroundings and a variety of different interactions between objects. All images are taken from the MS-COCO 2014 test-set and thus, they are never seen by the model during training. Moreover, there is no ground truth caption available for these images since they are used for testing for the MS-COCO Leaderboard competition



A series of photos showing different types of food.

A cat wearing a pink hat sitting on top of a carpet.

A stone patio with a stone walkway and a lot of chairs.

A group of men standing around a sheet cake holding a sword.

A bathroom with a walk in shower next to a toilet.

A classic car parked in the grass near a group of people.

A man flying through the air while riding a kiteboard.

A man in a yellow shirt catching a white frisbee.

A school bus is reflected in a rear view mirror.

Figure 5.11: Additional examples showing the model's ability to utilize the rich image features from the encoder, showing fewer "hallucinations" while captioning

CHAPTER 6

Conclusion

6.1 Conclusion

Description of images by humans usually reflect a contextual basis. Realizing the significance of this basis and its nuances, we introduced a novel context-based mechanism that strongly couples the visual and language features on three contextual grounds: (i) region-based image feature-level context, (ii) task-specific context, and (iii) semantic word-level context. A thorough evaluation of the couplings demonstrates qualitative advantages of our approach in the form of novel semantic constructs and effective utilization of the encoder's features. We maximize focus of attention (and thus caption generation performance) by integrating complementary mechanism that couple visual features (from the input image) and textual features (from the already generated caption components). The inspiration behind our approach is explained and the intuition behind our final model is progressively laid using three increments to baseline. Our architecture: (i) maintains a balanced attention between the two structurally different modalities of vision and language, and (ii) enhances information retention in the learned components of the architecture from both modalities leading to increased visual and textual feature utilization. Our final model outperformed state-of-the-art models with a wide margin using an end-to-end jointly trainable architecture without incurring an overhead of using any external resource in terms of supervision and training. We experimentally established our model's ability to: (i) correctly focus on salient image region while generating words, (ii) form meaningful semantic associations within learned embeddings, and (iii) generate novel semantic constructs. Our work could have wide implications in creating attentive encoder-decoder architectural pipelines in any task lying at the intersection of vision and language.

6.2 Future Work

Our future work will focus on analyzing our captioning architectures on learning-based metrics [42].

REFERENCES

[1] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollr, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv:1504.00325*, 2015.

[2] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2017, pp. 6298–6306.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[4] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 6, 2017, p. 2.

[5] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision.* Springer, 2014, pp. 740–755.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] P. Bloom, *How children learn the meanings of words.* MIT press, 2002.

[12] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, “Rich image captioning in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 49–56.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2017, pp. 2261–2269.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[16] 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 visual question an-

swering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] R. Luo, "Unofficial pytorch implementation for self-critical sequence training for image captioning," https://github.com/ruotianluo/self-critical.pytorch, 2017.

[18] R. Egly, J. Driver, and R. D. Rafal, "Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects." *Journal of Experimental Psychology: General*, vol. 123, no. 2, p. 161, 1994.

[19] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, 2001.

[20] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 2017, pp. 305–313.

[21] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[22] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, 2007. [Online]. Available: http://science.sciencemag.org/content/315/5820/1860

[23] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[24] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[25] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[26] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision.* Springer, 2016, pp. 382–398.

[27] A. Karpathy and L. 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.

[28] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.

[29] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning." in *AAAI*, 2017, pp. 4233–4239.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[31] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2014, pp. 302–308.

[32] L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur, "Semantic structure and interpretability of word embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[33] G. Murphy, *The big book of concepts.* MIT press, 2004.

[34] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ

[35] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[36] D. Elliott and F. Keller, “Comparing automatic evaluation measures for image description,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2014, pp. 452–457.

[37] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning,” *arXiv preprint arXiv:1612.07600*, 2016.

[38] W. Zhao, B. Wang, J. Ye, M. Yang, Z. Zhao, R. Luo, and Y. Qiao, “A multi-task learning approach for image captioning.” in *IJCAI*, 2018, pp. 1205–1211.

[39] J. Gu, J. Cai, G. Wang, and T. Chen, “Stack-captioning: Coarse-to-fine learning for image captioning,” *arXiv preprint arXiv:1709.03376*, 2017.

[40] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 22–29.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “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.

[42] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, “Learning to evaluate image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5804–5812.

BIOGRAPHICAL STATEMENT

Ankit Khare was born in Bhopal, M.P., India in 1991. He received his B.S. degree from Lovely Professional University, India, in 2014, his M.S. degree from The University of Texas at Arlington in 2019 all in Computer Science. From 2014 to 2016, he was with Nagarro Software Pvt. Ltd., India as a software developer. His current research interest is in the area of Computer Vision and Natural Language Processing using Deep Neural Networks.