Show, Infer and Tell: Contextual Inference for Creative Captioning

Ankit Khare
ankit.khare@mavs.uta.edu
Manfred Huber
huber@cse.uta.edu

The University of Texas at Arlington
Texas, USA

Abstract

Several attention based encoder-decoder architectures have been geared towards the task of image captioning. Yet, the collocations and contextual inference seen in captions written by humans is not observed in the output of these systems e.g., if we see a lot of different vehicles on the road, we infer "traffic" and say "a lot of traffic on the road". Further, "hallucination" of commonly seen concepts for fitting the language model is commonly observed in a lot of existing systems. For example, "a group of soldiers cutting a cake with a sword" would be hallucinated as "a boy cutting a cake with a knife". In this work we construct two simultaneously learning channels, where first channel uses the mean-pooled image feature and learns to associate it with the most relevant words. The second channel, on the other hand, utilizes the spatial features belonging to salient image regions to learn to form meaningful collocations and perform contextual inference. This way, the final language model gets the opportunity to leverage the information from the two channels to learn to generate grammatically correct sentence structures which are more human-like and creative. Our novel "spatial image features to n-gram text features mapping" mechanism not only learns meaningful collocations but also verifies that the caption words correspond to the region(s) of the image, thereby avoiding "hallucination" by the model. We validate the effectiveness of our one pass system on the challenging MS-COCO image captioning benchmark, where our single-model achieves a new state-of-the art 126.3 CIDEr-D on the Karpathy split, and a competitive 124.1 CIDEr-D (c40) on the official server.

1 Introduction

Deep neural network based encoder-decoder architectures [12, 23, 29, 34] have been quite successful in producing better captioning results as compared to earlier template based methods [9, 19, 20] and represent the current state-of-the-art. This is mainly due to their ability to form complex representations using large datasets [22] and to learn long-range sequences [14] which together make it possible to summarize an image. Image captioning is challenging because so far it is not possible to make machines go through the same experience and context that allows humans to understand the underlying complex concepts. 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)

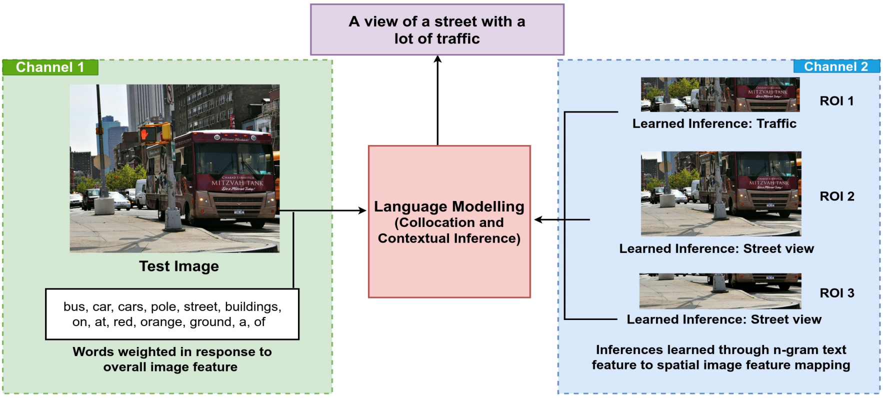
A view of a street with a lot of traffic
Channel 1
Test Image
bus, car, cars, pole, street, buildings, on, at, red, orange, ground, a, of
Words weighted in response to overall image feature
Language Modelling (Collocation and Contextual Inference)
Channel 2
ROI 1
Learned Inference: Traffic
ROI 2
Learned Inference: Street view
ROI 3
Learned Inference: Street view
Inferences learned through n-gram text feature to spatial image feature mapping

Figure 1: An illustrative example showing our approach for generating image captions.

intelligent utilization of their knowledge of language to verbally summarize those representations. The same has to be the case in neural network based automatic image captioning. The system must be capable of abstracting representations of objects and their attributes. Then it should be able to learn collocations in the language and produce contextual inferences.

Research shows that humans initially look at the overall image and then focus on the sequence of individual pieces to form collocations and infer contextual associations [4]. Finally, they come up with an abstract sentence which creatively describes the image. This phenomenon forms our intuition behind the ways to architect the interactions between the two modalities of vision and language. Intuitively, the first channel learns to associate text features and overall visual information as training progresses. This channel biases text generation towards a context most appropriate for the overall appearance of the image. In addition, we introduce a second channel that learns collocations and contextual inferences by "gaiting" (weighing) all the n-gram text features generated by the network until the current time-step to each individual spatial image feature (encoded feature corresponding to a spatial area of an image) corresponding to a region of interest (ROI) for all ROIs. The two simultaneously operating channels are concatenated to form an input to the language model. Thus, our language model learns to strengthen the contextual inferences using well associated multi-modal features to be able to produce creative and more human-like captions (Fig. 1). This happens in one pass without the overhead of running multiple passes involving multiple CNNs and RNNs. It is noteworthy that the two pass architectures [16, 38] incur this overhead and still under-perform lacking advantages of explicit n-gram text feature to spatial image feature mapping.

2 Related Work

Our intuition builds up from the observation of earlier attention models [10, 11, 33, 37] that learn to attend to images on the basis of a mechanism which can be spatial attention [33], text-based attention [37], or a combination of both [35]. Although these attention mechanisms have provided useful focus of attention, they are still unable to utilize the full extent of the richness of encoded image features. Moreover, many of them fit to frequently observed concepts in the language model and fail to generalize well when given a new scenario. Addressing this is very important since real world images are very diverse.

One of the early works is [33] that suffers from poor region proposals where objects at the boundaries are considered insufficiently. In our model, we resort to the use of the Region Proposal Network (RPN) to ground proposed regions in salient objects. Jia et al. [15] exploit the relation between images and their captions as the global semantic information to guide the language LSTM. The guidance is pre-specified, linear and fixed over time. Moreover, it lacks explicit spatial information from image regions. You et al. [35] and Gan et al. [11] incorporate semantic concepts where image features are vectors of confidences of attribute classifiers. This requires additional external resources and increased network overhead to train these semantic attributes. Despite this, they still are limited in terms of learning collocations. In contrast, [37] systematically incorporates time-dependent text-conditional attention, from 1-gram to n-gram. Still, it lacks region-based spatial attention and is prone to “hallucinating" previously seen concepts based on the language model.

Another area of related work is Text-Attention [24] which relies on ground truth captions to be used as a basis of selecting visual features. Here, a model would suffer from the error prone test time sampling where error would build up during captioning and propagate further while referencing the caption. For example, suppose that at test time the model hallucinates a person watching the TV whereas the test image originally has a kid playing in front of a TV, then the wrong captions would be referenced in the second pass and the network might propagate them further. Knowing When to Look [23] utilizes the impact of visually attending to an image only when a salient word is encountered. Words like ‘of’, ‘the’, and ‘with’ which are ignored in this model provide useful context for associating visual and verbal concepts. These prepositions form the basis to learn collocations and contextual inferences. Lastly, the Up-Down Captioner [2] focuses primarily on the generation of bottom-up features. However, there is no explicit mechanism to utilize the rich features from the encoder for obtaining sentence constructs reflecting human-like creativity.

3 Method

Addressing the aforementioned shortcomings, our architecture is shown in Fig. 2.

3.1 Encoder

To demonstrate the wide applicability of our approach and advantages of using a Region Proposal Network (RPN), we train two models, one encoded with Resnet-101 [13] features, and another using Faster-RCNN [28] in conjunction with Resnet-101 for encoding the image.

3.2 First Channel: Overall Context and Additive Biasing

Let V_s denote spatial features and $\bar{V}$ denote mean-pooled image feature or overall image feature:

$$\bar{V} = \frac{1}{k}\sum_{s=1}^{k} V_s \tag{1}$$

where $k = 100$. For h_i^t and x_i^t, subscripts denote LSTM layer number i.e., 1 for LSTM1 and 2 for LSTM2, and superscripts denote time-step. At each time-step, LSTM1 receives input, x_1^t containing mean-pooled image feature $\bar{V}$, the encoded representation of the last generated word, and LSTM2’s hidden state h_2^{t-1} from the last time-step:

$$x_1^t = [h_2^{t-1}, \bar{V}, W_{e_1}^t S^t] \tag{2}$$

$W_{e_1} \in \mathbb{R}^{e\times|\Sigma|}$ is the word embedding for our vocabulary Σ, e is the word embedding size, and S^t is a one-hot encoding of the last generated word. W_{e_1} is randomly initialized and

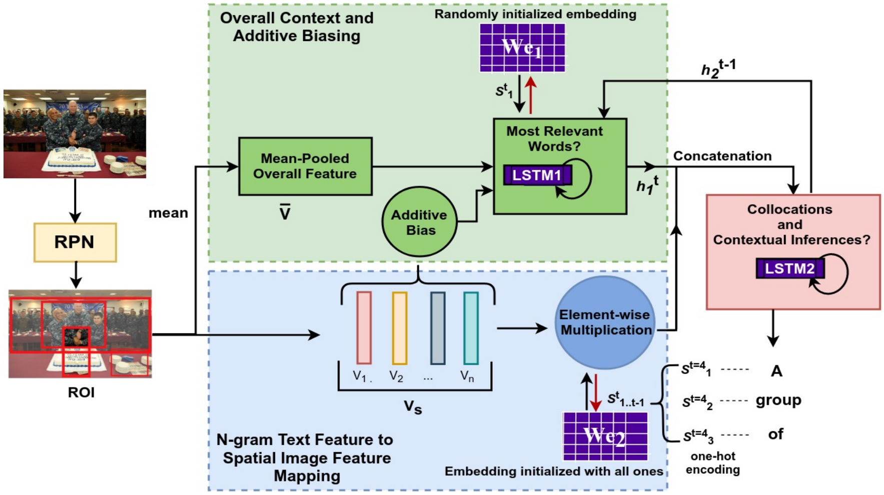
Overall Context and Additive Biasing
Randomly initialized embedding
W_{e_1}
s^t_1
h_2^{t-1}
Most Relevant Words?
LSTM1
Mean-Pooled Overall Feature
$\bar{V}$
Additive Bias
mean
RPN
ROI
Concatenation
h_1^t
Collocations and Contextual Inferences?
LSTM2
V_1
V_2
...
V_n
V_s
Element-wise Multiplication
$s^t_{1..t-1}$
W_{e_2}
$s^{t=4}_1$ ----- A
$s^{t=4}_2$ ----- group
$s^{t=4}_3$ ----- of
one-hot encoding
N-gram Text Feature to Spatial Image Feature Mapping
Embedding initialized with all ones

Figure 2: Architectural framework of the proposed system. Green and blue regions denote channel 1 and channel 2 respectively. Violet components constitute the learnable parts of the system. Red arrows signify that the two word embeddings learn semantics or text feature representations through backpropagation. All the feature representations of the words from W_{e_2} generated until the current time-step are utilized by channel 2 to associate them to spatial image features at every time-step.

learned from scratch. The hidden state h_1^t of LSTM1 combines with spatial features V_s:

$$f_s^t = w_f^T tanh(W_{vf}V_s + W_{hf}h_1^t) \tag{3}$$

$W_{vf} \in \mathbb{R}^{d\times v}$, $W_{hf} \in \mathbb{R}^{d\times m}$, 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 number of hidden units in each of the two LSTMs. We use batch normalization for the attention layer which we have observed leads to faster convergence. We form the first attention pathway as follows:

$$\alpha = softmax(f^t) \tag{4}$$

$$\hat{v}^t = \sum_{s=1}^{k} \alpha_s^t V_s \tag{5}$$

where at any given time-step t, α^t acts as a weight mask for spatial features V_s (convex combination). As a result of the additive operation (Eq. 3), the bias favors increasing the weight of correct spatial image regions which could have the objects strongly associating to the relevant representations within the first word embedding.

3.3 Second Channel: N-gram Text Feature to Spatial Image Feature Mapping

This channel introduces a novel "gaiting pathway" formed by coupling the previously generated word history, $S^t_{1..t-1}$, directly to the spatial image features, V_s using an element-wise multiplicative operation followed by non-linearity:

$$C_s^t = W_c^T tanh(V_s \odot W_{e_2} \sum_{i=1}^{t} \frac{S_{i-1}}{t}) \tag{6}$$

$$\beta^t = softmax(C^t) \tag{7}$$

where $W_c \in \mathbb{R}^{d \times 1}$. The second word embedding, W_{e_2}, is also trained from scratch but unlike the first one, it is initialized with all ones to verify if the network decides to keep the “ones” intact or if it structures the word vectors in an interpretable way which can resemble correlations among words used in similar contexts. Applying the weighted mask on V_s to form the input of LSTM2 yields:

$$\hat{v}_c^t = \sum_{s=1}^{k} \beta_s^t V_s \tag{8}$$

$$x_2^t = [\hat{v}^t, h_1^t, \hat{v}_c^t] \tag{9}$$

The text features from the second embedding are batch normalized. We observed that it leads to faster convergence and avoids saturation of neurons during training. The first channel provides more basic "semantics" with limited spatial information. In contrast, the second channel provides more sophisticated "semantics" and contains enough spatial information. It is thus directly concentrated on the image locations relevant to the language context. Due to the different uses in the two simultaneously operating channels, the two learnable word embeddings will represent different semantic information with one containing more general content, while the other is more specific. Together this results in a more complete representation of the language context of the caption. Additionally, the "n-gram text feature to spatial image feature" mapping allows the network to correlate all permutations and combinations of ROIs and relevant word representations for updating the weights in recurrent network as well as the semantic representations in the word embedding to be able to learn collocations and contextual inferences (Fig. 2). This leads to avoiding hallucinations and generating more human-like captions.

3.4 Learning

We train our model using cross entropy loss and further optimize it on CIDEr [32]. Suppose the maximum length of any caption is L. Our aim is to calculate the conditional probability P of the sequence of words $(y_1, y_2 ... y_L)$ over the vocabulary Σ. At any time-step t we take the hidden state of LSTM2, h_2^t, and calculate the softmax distribution to find the conditional probability as:

$$p(y^t | y^{1:t-1}) = softmax(W_p h_2^t + b_p) \tag{10}$$

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:

$$P = \prod_{t=1}^{L} p(y^t | y^{1:t-1}) \tag{11}$$

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

$$L(\theta) = -\sum_{t=1}^{L} log(P_\theta(y^{*^t} \mid y^{*^{1:t-1}})) \tag{12}$$

During CIDEr optimization, the metric is directly optimized to minimize the loss:

$$LF_R(\theta) = -E_{y^{1:L} \sim p_\theta}[r(y^{1:L})] \tag{13}$$

where the parameters of the network are given by θ, and r is the score function (CIDEr). From the method described in SCST [29], the gradient for LF is approximated as:

$$\nabla_\theta LF_R(\theta) \approx -(r(y_s^{1:L}) - r(\hat{y}^{1:L}))\nabla_\theta log p_\theta(y_s^{1:L}) \tag{14}$$

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.

Model	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Up-Down (resnet baseline) [2]	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
CIC-R101 (our resnet baseline)	**77.8**	**34.8**	**27.3**	**56.5**	**111.3**	**20.1**	**79.2**	**35.3**	**27.7**	**57.2**	**116.8**	**21.2**
Up-Down (bottom-up feats.) [2]	77.2	36.2	27.0	56.4	113.5	20.3	78.4	36.1	27.5	57.1	117.8	20.8
CIC-RCNN (our final model)	**78.0**	**36.5**	**27.7**	**57.3**	**116.7**	**20.7**	**81.1**	**39.3**	**28.8**	**58.9**	**126.3**	**22.0**
	Cross-Entropy Loss						CIDEr Optimization					

Table 1: Comparison between our models (in bold) and Up-Down Captioner on Karpathy split. Our models surpasses the corresponding up-down captioner by a significant margin.

4 Experiments

4.1 Dataset, Settings and Metrics

MS-COCO[22]: There are two standard splits of MS-COCO: the official online test split and the Karpathy split [17] for offline test. The first split has $82,783$, $40,504$ and $40,775$ train, val, and test images respectively, each of which has 5 human labeled captions. The second split has $113,287/5,000/5,000$ train/val/test images, each of which has 5 captions.

Settings: Our final model has two LSTMs [14] each with m=2,048 hidden units. The two word embeddings have a hidden size e of 1,024 units. The hidden unit size d of 1,000 is chosen for the attention layer. ADAM with amsgrad [27] optimizer is used. The 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. We trained on cross-entropy loss for 25 epochs and subsequently used SCST [29] for 15 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. During optimization, beam size was set to 5. We perform minimal text pre-processing by tokenizing on white space, 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.

Metrics: We used five standard automatic evaluation metrics: Bleu [26], METEOR [3], ROUGE-L [21], CIDEr-D [32], and SPICE [1]. In Tab. 1 and 2, B-N, M, R, C, and S denote BLEU-N, METEOR, ROUGE-L, CIDEr-D, and SPICE respectively.

4.2 Ablative Studies

4.2.1 Encoder Variants

Although our approach can be applied without using the Region Proposal Network (RPN), we train our final model using RPN [28] to ground visual feature-vectors in objects rather than simply using bilinear interpolation to resize the output to a fixed size spatial representation [33]. Objects form a natural basis [8, 30] for visual attention. Getting salient attention regions and visual features for each region, gives us a head start for the image captioning process from the encoder's end. For our Resnet baseline, referred as CIC-R101, the mean-pooled image features are obtained from the final convolutional layer of Resnet-101 pre-trained on ImageNet [7]. To obtain spatial image features we follow the approach used in SCST [29] and use bilinear interpolation to form fixed size spatial representations of 10×10. For our final model, referred as CIC-RCNN, we take the final output of the Faster RCNN [28] and perform non-maximum suppression. In our implementation we used an IoU threshold of 0.7 for region proposal non-maximum suppression, and 0.3 for object class non-maximum suppression. To select salient image regions, we simply selected the top k = 100 features in each image. Table 1 shows our resnet baseline and final model's performance as compared to bottom-up captioner [2] on MS-COCO dataset.

Model	Cross-Entropy Loss						CIDEr Optimization					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Mean-Pooled-CIC (ours)	77.5	36.2	27.2	56.9	114.8	20.4	78.8	36.8	27.9	57.6	119.2	21.1
CIC-RCNN (ours)	78.0	36.5	27.7	57.3	116.7	20.7	81.1	39.3	28.8	58.9	126.3	22.0

Table 2: Performance of our models on Karpathy split illustrating quantitative improvements when using spatial image features in the second channel of our architecture.

4.2.2 Variation in Channel-2 of Decoder

In order to evaluate the importance of our "n-gram text feature to spatial feature" mapping within the second channel, we replace the spatial image features with the mean-pooled feature-vector to "gait" the n-gram text features (sentence history). Hence, Eq. 6 and the input to LSTM2 (Eq. 9) changes as follows:

$$C^t = W_c^T tanh(\bar{V} \odot W_{e_2} \sum_{i=1}^{t} \frac{S_{i-1}}{t}) \tag{15}$$

$$x_2^t = [\hat{v}^t, h_1^t, C^t] \tag{16}$$

We refer to this fully trained experimental model as Mean-Pooled-CIC. Intuitively, since the mean-pooled feature has limited spatial information, the possibility of "hallucination" of objects that are not present in an image by the model should increase. Further, the level of detail in captions should decrease. The quantitative and qualitative results shown in Tab. 2 and Fig. 3, respectively confirm our intuition. The scores in CIDEr and SPICE which correlates the most to ground truth captions written by humans decrease significantly in Mean-Pooled-CIC. In Fig. 3 (a), Mean-Pooled CIC and up-down captioner hallucinate "a leash" and the position of the dog as "standing" (Red colored fonts in Fig. 3). By contrast, our final model, CIC-RCNN is capable of understanding the interaction between objects and generalizing them creatively in language despite the fact that in most of the images in the training set that contain leashed dogs, they are either sitting or standing on the floor. Thus, we obtain a more descriptive, creative and human-like caption describing the correct positioning of the dog with respect to a person where the model doesn't confuse a charging cable with a leash. Similarly, in part (b) we see a higher level of detail (Blue fonts show high quality captions from CIC-RCNN and Brown fonts represent less details within captions) and creativity shown by our final model CIC-RCNN where cutting the cake by a group of soldiers is again an out-of-context scenario but CIC-RCNN nevertheless describes it quite well. Likewise, in

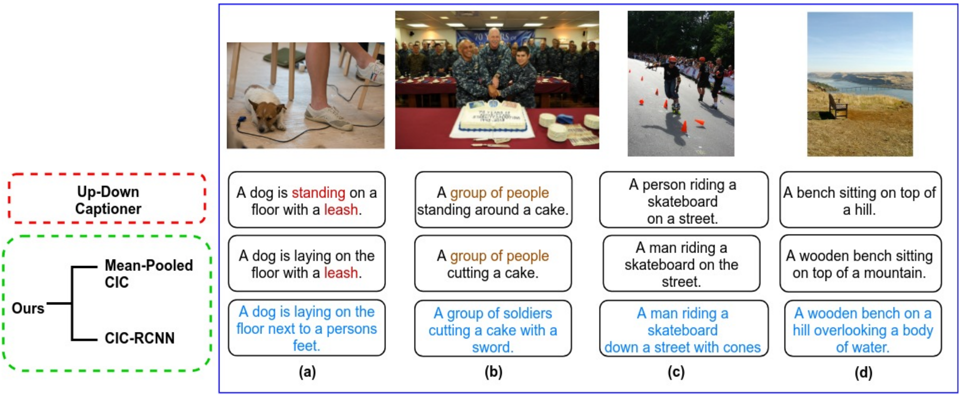
Up-Down Captioner
Ours
Mean-Pooled CIC
CIC-RCNN
A dog is standing on a floor with a leash.
A dog is laying on the floor with a leash.
A dog is laying on the floor next to a persons feet.
(a)
A group of people standing around a cake.
A group of people cutting a cake.
A group of soldiers cutting a cake with a sword.
(b)
A person riding a skateboard on a street.
A man riding a skateboard on the street.
A man riding a skateboard down a street with cones
(c)
A bench sitting on top of a hill.
A wooden bench sitting on top of a mountain.
A wooden bench on a hill overlooking a body of water.
(d)

Figure 3: Qualitative examples illustrating the importance of our "n-gram text feature to spatial image feature" mapping for creative and more human-like captioning.

part (c) and (d) the captions are more descriptive and creative. In part (d) the position of the wooden bench and its association with the body of water are inferred from the interactions and spatial positioning of the salient objects: body of water, terrain, and bench.

4.2.3 Novel Semantic Constellations

In the example (COCO_val2014_000000005820.jpg) in Fig. 3 (d), the caption generated by CIC-RCNN generalizes the image quite well and produces novel semantic constellations. There are 1,843 instances in 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. 3 (d). Our model learns from these examples, how to associate objects and object's attributes to form collocations and contextual inferences.

In another example in Fig. 3 (c), 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 [22] 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 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 constellations.

4.2.4 Embedding Analysis

Dense word embeddings can be successful in capturing semantic relations among words. Presence of a meaningful semantic structure in their respective vector spaces is highly probable [31]. 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 associated with image features to learn collocations in language. In our exploration, we refer to the category theory [25] and construct a KNN adjacency matrix containing the pair-wise similarities among representations within the word embedding. Let A_m be a $\Sigma \times \Sigma$ matrix (Σ is our vocabulary size). 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 ground truth captions are reflected in the nearest words. The first embedding, W_{e_1}, learns to associate nouns with their prepositions and other related verbs/nouns, while the second, W_{e_2}, learns collocations commonly used for the object under consideration (Table 3).

Word	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
cat	a	of	on	with	the	sitting	at	cats
	black	domestic	laying	looking	sleeping	kitchen	standing	hides
car	a	at	in	sitting	pulling	down	are	that
	parked	traffic	road	traveling	trunk	window	luggage	seat
frisbee	a	next	to	on	flying	playing	with	while
	dog	playing	catching	grass	beach	throws	park	holding
bananas	of	UNK	to	a	on	near	for	sitting
	hanging	eating	yellow	surrounded	table	plate	full	fruit

Table 3: Eight nearest words (N_1-N_8) in the two embeddings of the CIC-RCNN model for randomly selected words. For each word, row 1 and 2 represent W_{e_1} and W_{e_2}, respectively.

We observed that the commonly seen words had dense representations and relatively large deviations from the initial value (embedding W_{e_2} was initialized with all ones) whereas

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr-D	SPICE
SCST:Att2all [29]	-	-	-	34.2	26.7	55.7	114.0	-
Up-down [2]	79.8	-	-	36.3	27.7	56.9	120.1	21.4
MLAIC [36]	80.7	63.9	49.0	36.9	27.7	57.5	119.1	-
STACK-CAP [12]	78.6	62.5	47.9	36.1	27.4	56.9	120.4	20.9
CIC-RCNN	**81.1**	**65.6**	**51.1**	**39.3**	**28.8**	**58.9**	**126.3**	**22.0**

Table 4: Performance of our final model, CIC-RCNN, on MS-COCO Karpathy split. Bold figures represent the highest scores. Our single model (without ensemble) outperform other state-of-the-art single models by a significant margin in all metrics.

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† [34]	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 [12]	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† [2]	80.2	**95.2**	**64.0**	**88.8**	49.1	**79.4**	36.9	**68.5**	27.6	36.7	57.1	72.4	117.9	120.5
MLAIC [36]	**80.3**	94.4	63.4	87.3	48.3	77.4	36.0	66.1	27.4	36.1	57.0	71.8	113.9	116.4
CIC-RCNN	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: For each column, bold figure represent the highest score. † is used to denote the use of ensemble of several differently initialized models. Our single model is able to outperform previous state-of-the-art results by a significant margin on the MS-COCO test server.

words found rarely within captions stayed closer to their initialization. In CIC-RCNN, compared to Mean-Pooled-CIC, we observed an increased structuring of the representation space, indicated by increasing deviations from the initialization values, particularly for collocations in the English language. To further examine the embedding, we tried different embedding sizes (512, 1,000, and 1,024) for the two embeddings. We found that a ratio of 1:1 between the LSTM size and the embedding size worked best in terms of the quality of the representations inside the embeddings.

4.3 Comparison with State-of-The-Arts

Comparing Methods: Though there are various captioning models developed in recent years, for fair comparison, we only compared CIC-RCNN with some encoder-decoder methods trained by the RL-based reward, due to their superior performances. Specifically, we compared our methods with SCST [29], StackCap [12], Up-Down [2], LSTMA3 [34], and MLAIC [36]. Among these methods, SCST and Up-Down are two baselines where the more advanced self-critic reward and visual features are used. Compared with SCST, StackCap proposes a more complex RL-based reward for learning captions with more details.

Results on Karpathy split: From Table 4, we can see that our single model achieves a new state-of-the-art score among all the compared methods in all metrics. CIC-RCNN gains on up-down captioner by an absolute 5.9 on CIDEr-D.

Leaderboard Results: Our model outperforms Up-Down Captioner, SCST-Att2all and other leading state-of-the-art ensemble models using only a single model on MS-COCO Leaderboard [5] in terms of METEOR, ROUGE-L, and CIDEr-D while being competitive (almost always first or second) in Bleu scores (Tab. 5). Note that Bleu, initially proposed for machine translation, is based on explicit word matching and fails to spot semantic similarity when common words are scarce. It is affected by word vocabularies and synonyms [18]. It is thus a less effective metric for caption evaluation. Compared to the other non-ensemble method, StackCap, our method still performs better by utilizing the advantages of obtaining spatial image features using RPN and mapping them with n-gram text features, even when our RL-reward is not as sophisticated as theirs.

5 Conclusion

We propose a single pass encoder-decoder framework for image captioning comprised of two simultaneously learned channels: (i) overall context and additive biasing and (ii) n-gram text feature to spatial image feature mapping. Our model learns to form collocations and contextual inference to produce more human-like captions in a single pass without using multiple combinations of CNNs and RNNs. It outperforms existing state-of-the-art models with a significant margin without incurring an overhead of using any external resource in terms of supervision and training. We experimentally validate the advantages of our approach using extensive ablative studies. We observed that our model avoids hallucinations even in out-of-context scenarios and the captions generated are creative and more human-like as compared to existing state-of-the-art models (Fig. 3).

Our future work will focus on analyzing our models on learning-based metric [6] and extending our approach to other tasks lying at the intersection of vision and language like Visual Question Answering (VQA).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Satanjeev Banerjee and Alon 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*, pages 65–72, 2005.

[4] Timothy J. Buschman and Earl K. Miller. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–1862, 2007. ISSN 0036-8075. doi: 10.1126/science.1138071. URL http://science.sciencemag.org/content/315/5820/1860.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Robert Egly, Jon Driver, and Robert D Rafal. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123(2):161, 1994.

[9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[10] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

[11] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.

[12] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. *arXiv preprint arXiv:1709.03376*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415, 2015.

[16] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[18] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*, 2016.

[19] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[20] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard 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)*, volume 6, page 2, 2017.

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Text-guided attention model for image captioning. In *AAAI*, pages 4233–4239, 2017.

[25] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[27] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[29] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[30] Brian J Scholl. Objects and attention: The state of the art. *Cognition*, 80(1-2):1–46, 2001.

[31] Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[33] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[34] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*, pages 22–29, 2017.

[35] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[36] Wei Zhao, Benyou Wang, Jianbo Ye, Min Yang, Zhou Zhao, Ruotian Luo, and Yu Qiao. A multi-task learning approach for image captioning. In *IJCAI*, pages 1205–1211, 2018.

[37] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 305–313. ACM, 2017.

[38] Zhihao Zhu, Zhan Xue, and Zejian Yuan. Think and tell: Preview network for image captioning. In *British Machine Vision Conference*, pages 3–6, 2018.