MRRC: Multiple Role Representation Crossover Interpretation for Image Captioning With R-CNN Feature Distribution Composition (FDC)

Chiranjib Sur

Computer & Information Science & Engineering Department, University of Florida.

Email: chiranjibsur@gmail.com

arXiv:2002.06436v1 [cs.LG] 15 Feb 2020

Abstract—While image captioning through machines requires structured learning and basis for interpretation, improvement requires multiple context understanding and processing in a meaningful way. This research will provide a novel concept for context combination and will impact many applications to deal visual features as an equivalence of descriptions of objects, activities and events. There are three components of our architecture: Feature Distribution Composition (FDC) Layer Attention, Multiple Role Representation Crossover (MRRC) Attention Layer and the Language Decoder. FDC Layer Attention helps in generating the weighted attention from RCNN features, MRRC Attention Layer acts as intermediate representation processing and helps in generating the next word attention, while Language Decoder helps in estimation of the likelihood for the next probable word in the sentence. We demonstrated effectiveness of FDC, MRRC, regional object feature attention and reinforcement learning for effective learning to generate better captions from images. The performance of our model enhanced previous performances by 35.3% and created a new standard and theory for representation generation based on logic, better interpretability and contexts.

Index Terms—language modeling, representation learning, tensor product representation, image description, sequence generation, image understanding, automated textual feature extraction

I. INTRODUCTION

BIG data is struck with volume and variety of data and is creating huge challenges to merge them for analysis and inference. While large part of data is in images and videos and helping in better expression, entertainment and communication, it is difficult for machines to understand, merge, recommend and track the sentiments and trends of the specific and generalized happenings. To overcome this, captioning research must explore its fullest potential, for generalized and domain adaptation. Image captioning architectures [1] demonstrated diverse ways of effective sentence generation from visual features through various ways of representation generation and compositions from image features like Vgg [51], ResNet ([89], [54], [73]), Inception [47] etc, mostly relying on object and attribute detectors to describe images ([51], [53], [87]) and later focused on attention based model ([49], [47], [60], [54], [38], [45], [15]) and semantic factorization [73]. Recent works with top-down objects from image regions ([41], [6]) has used hierarchical models. This architecture [41] has some serious drawbacks and we have introduced some solutions that are more robust and capable of introducing truly interpret-able representation instead of just relying on the network to learn them independently. Also, our architecture has much lower number of weights and can be trained in less than 20 epochs, while their architecture requires 50 epoch of rigorous training and then fine-tuning. Another drawback is related to global overview, derived out of average of all regional components. An overlap of a happy person and a sad one will produce only sad context or a negative way of representation because of the structural nature of the situation, which machine fails to interpret. Averaging suppresses many informative objects and overlapping average never works. The compositional characteristics of image-attention LSTM and intermediate transfer layer has limitations and creates a heuristic weighted selection (as **a** in [41]) for image regions $\{v_1, \ldots, v_n\}$. Here, apart from a 2048 dimension LSTM, the intermediate transfer layer weights $\mathbf{W}_b \in \mathbb{R}^{b \times d}$, $\mathbf{W}_h \in \mathbb{R}^{b \times d}$, $\mathbf{W}_a \in \mathbb{R}^{b \times d}$ are more than required. Hence, we curtailed the image-attention LSTM and $\mathbf{W}_b$ and introduced $\mathbf{W}_h \in \mathbb{R}^{b \times d}$ and $\mathbf{W}_a \in \mathbb{R}^{b \times d}$ with much lower weights and this can help in propagation of information much faster. More details are provided in Section IV. Our approach outperformed top-down approach in [6] which used only this image caption data for training, while our results were very close to top-down approach in [41], which used three data (image captioning, VQA and Visual Genome Question Answering) to train their system for representation and use the features for different applications. Hence, direct comparison of our result with [41] will be unfair since we only use MSCOCO image caption dataset [73]. We introduced a weighted image composition scheme to produce Feature Distribution Composition (FDC) through selection of the distribution of objects with the real global feature from the original image feature. This helps in better and quick selection of the weights for each corresponding regional object features. Also the intermediate transfer layer creates a fat structure, creating fan out situation for the information and further propagation of information to the image-attention LSTM layer get heavily diminished. Also, with lower gradient propagation, learning is slow and it is difficult to work with low value for initialization. Also, in [41], the intermediate self-attention feedback are the hidden layer outputs ($\mathbf{h}_1$ and $\mathbf{h}_2$) and most of the time propagates the initial generated errors, while the correct word embedding need to be propagate through the network for better learning. Overall, the model contains too many information fusion and it is difficult to control such mixing in a very heavy model with relatively lower resources

(data and computation). We replaced these heavy weights with a simple and light weighted model and paid attention to composition of features from image features, global objects representation from the original image and use of semantic composition for construction of sentences. We showed that our composition of feature representation is a much advanced way of composition of representation that the system can distinguish and decode as sentences, while our proposed R-CNN feature composition architecture provides similar kind of performance, when experimented with MSCOCO dataset.

The novelty of our work is multiple-fold. First, we advocate that crossover of representation can enhance image captioning, then we introduced a novel R-CNN features composition architecture for composition of structural representation in the form of Feature Distribution Composition (FDC) and lastly combined crossover representation, feature distribution composition and semantic composition for a better network. Lastly, we used SCST reinforcement learning scheme [45] for better performance enhancement and as a moderating way of incorporating the cumulative topological effects of sequence of the sentence and generate changes in the network weights.

The rest of the document is arranged with the details of theory of tensor product and its influence on our architecture in Section III, the description of the architecture and description of the different feature composition techniques in Section IV, the intricacies of our experiments, results and analysis in Section V, revisit of the existing works in literature in Section II, concluding remarks with future prospects in Section VI.

II. Literature Review

There are quite a number of works being done on image captioning like [5] CNN features and hash-tags from users as input, [6] with sentence 'template', [7] sentiment-conveying image descriptions. [8] reported the comparison of context-aware LSTM captioner and co-attentive discriminator for image captioning. [9] used question features and image features, [11] parsing tree StructCap, [12] sequence-to-sequence framework, and [13] dual temporal modal, Image-Text Surgery in [14], [15] attribute-driven attention, [16] generative recurrent neural network, [17] MLAIC for better representation. Also there is [18] text-guided attention, [19] reference based LSTM, [21] adversarial neural network, high-dimensional attentions [22], [23] coarse-to-fine skeleton sentence, [24] specific styles, [25] structural relevance and structural diversity, multimodal attention [26], [27] popular brands caption [28] diversified captions, [29] stylish caption, [30] sub-categorical styles, [31] personalized captions, [32] studied actor-critic reinforcement learning, [33] scene-specific attention contexts, [34] policy network for captions, [35] reinforcement learning based training, [36] distinguish between similar kind for diversity, [37] improved with correctness of attention in image, [39] adaptivity for attention, [40] used combination of computer vision and machine translation, [42] used adaptive re-weight loss function, [43] personalized captioning, [46] high level semantic concept, [47] used visual features and machine translation attention combinations, [56] different caption styles, [57] shifting attention, [58] characteristics of text based representations, [62] variational autoencoder representation, [63] dependency trees embedding, [64] character-level language modeling, [65] fixed dimension representation, [66] 3-dimensional convolutional networks, [67] human judgments, out-of-domain data handling, [70] semantic attention, [73] Semantic Compositional Network (SCN), [74] localize and segment objects, [76] extra semantic attention, [78] content planning and recognition algorithms, [80] new tree based approach to composing expressive image descriptions, [81] transposed weight sharing scheme, [82] different emotions and sentiments, and [85] where nouns, verbs, scenes and prepositions used for structuring sentence.

III. Image Weighted Tensor Crossover & Representations Theory

The concept is motivated by multiple extraction of different aspects of the images and combining them, in absence of enough trigger neurons for many unattended part of the image features. Uniformly distribution of triggering of every sector of the image is also not feasible. Hence, we pretend that this overall combinations is an approximation event for neural composition network function. Mathematically, we can say for image $\mathbf{I}$, we define an approximate context function $\Phi(.)$ through the combination of individual neuron triggered transformation of the image regions defined as $\sigma(\mathbf{W}_1\mathbf{I} + \mathbf{b}_1)$, $\sigma(\mathbf{W}_2\mathbf{I} + \mathbf{b}_2)$ and so on. The structural component of features in images are flattened through [89].

A. Approximation

Approximation has always lead to ease of computation, but here we are focused on approximation for scalability. For better approximation and compositional characteristics of representations, our architecture learns to predict two near-optimum representation tensors that can be combined to generate the maximum likelihood representation. The purpose is to complement each other, but in case of presence in both representations, non-linearity can neutralize such effects. These near-optimum tensors are learned from data and are generated as a function of visual features $\mathbf{v}$ for image $\mathbf{I}$. For example, we consider the time-independent representation $\mathbf{R}$ in Equation 1,

$$\mathbf{R} = f_1(\mathbf{v})f_2(\mathbf{v}) \tag{1}$$

with functions $f_1(.)$ and $f_2(.)$, we can also include time and rewrite representation $\mathbf{R}_t$ as,

$$\mathbf{R}_t = \sum_{i=0}^{i=t} c_i p_i^T \tag{2}$$

where we interpret functions $f_1(.)$ and $f_2(.)$ as c_i and p_i and are the context and positional significance vector for the image and is used to generate partial global representation $c_i p_i^T$ for each word in the sentence (or elements in a topological sequence), where the word embedding vector is defined as c_i and the positional or semantic interpretation is provided as p_i. We retrieve back c_i from decoder, since we have defined likelihood and we know $\mathbf{R}_t$ as the structure of sentence with positional significance $\{p_1, p_2, \ldots, \}$ for

each word $\{c_1, c_2, \ldots, \}$. Imposed orthogonality generates the detection theory for decoding as $p_i^T p_j = 0 \mid \forall\ i \neq j$ and $p_i^T p_j = 1 \mid \forall\ i = j$, if we know c_i. The approximate structure of the function creates retrieval as $c_1 p_1^T p_1, c_2 p_2^T p_2, c_3 p_3^T p_3$ and so on from each $\mathbf{R}_t$. Equivalently, we can write, estimated c_i as $\widehat{c}_i$,

$$\widehat{c}_i = \mathbf{R}_t p_i = c_i p_i^T p_i \qquad (3)$$

where $\mathbf{R}_t$ is Multiple Role Representation Crossover (MRRC). To apply this concept for language generation, we have two situations. One is where we know and defined one set of vectors and we can retrieve the other, with the help of $\mathbf{R}_t$ from contexts (here images). Another is when we derive both the vectors and $\mathbf{R}_t$ from from contexts (here images) and retrieve the other vectors. Experiments show that the latter is slightly better and this is due to approximations.

For image caption generation problem, it is difficult to retrieve whole $\mathbf{R}_t$ from $\mathbf{I}$ and hence we can define an approximate scheme where we compose partial $\mathbf{R}_t$ as $f(\mathbf{v})$ and add it to $\mathbf{R}_{t-1}$ for partial sequence context and establishing the topological dependency and thus retrieve context word c_i simultaneously from the image as attention. This architecture can be represented mathematically as the following conditional probability equation.

$$\mathcal{P}(c_i \mid f(\mathbf{v}), c_1, c_2, \ldots) = \mathcal{P}(c_i \mid f(\mathbf{v})) + \mathcal{P}(c_i \mid c_1, c_2, \ldots)$$
$$= \mathcal{P}(c_i \mid f(\mathbf{v})) + \mathcal{P}(c_i \mid \mathbf{R}_{t-1}) \qquad (4)$$

We have $\mathcal{P}(c_i \mid f(\mathbf{v}), c_1, c_2, \ldots)$ as the likelihood representation, attended as $\mathbf{R}_t$. The overall mathematical interpretation can be written as the following equation.

$$\mathbf{R}_t = c_1 p_1^T + c_2 p_2^T + \ldots + f(\mathbf{v})$$
$$= \sum_{i=0}^{i=t} c_i p_i^T + f(\mathbf{v}) \qquad (5)$$

where $f(\mathbf{v})$ can be regarded as structural correction by image attention. The ideal situation would have been when $c_t p_t^T = f(\mathbf{v})$ or $c_t p_t^T = f(\mathbf{v} \mid \mathbf{h}_{t-1})$ and is denoted as the following equation.

$$\mathbf{R}_t = c_1 p_1^T + c_1 p_1^T + \ldots + c_t p_t^T$$
$$= \sum_{i=1}^{i=t} c_i p_i^T \qquad (6)$$

where we know the sentence $\{w_1, w_2, \ldots, w_n\}$ (for image $\mathbf{I}$) and we can easily retrieve $\mathbf{R}_t$ by adding $c_i p_i^T$ to $\mathbf{R}_{t-1}$ from image and retrieving back c_i as $c_i p_i^T p_i$ (as $p_i^T p_i = 1$) through a sequence decoding technique. So apparently, we are getting the next word c_i from image $\mathbf{I}$ but through $\mathbf{R}_{t-1}$, but after each step we only know w_i at time t and before time t based on the decoded sequence, instead of all w_i before and after time t. This strategy will solve many problem related to generation of a complex and combined representation for the sentence and helps in scaling up the context diversity and generation capability. Since this is a non-linear transformation that is made to behave like reverse estimation, we will also predict the positional or semantic interpretation representation, an equivalent to topological information for a sequence. If time

factor is considered, the initial projection of the tensors goes through the following series of updation,

- $\mathbf{R}_0 \leftarrow Null$
- $\mathbf{R}_1 \leftarrow \mathbf{R}_0 + f_1(\mathbf{v})$
- $New\_Word \leftarrow w_1 = \mathbf{R}_1 * p_1$
- $\mathbf{R}_2 \leftarrow \mathbf{R}_1 + w_1 * p_1 + f_1(\mathbf{v})$
- $\mathbf{R}_{t+1} \leftarrow \mathbf{R}_t + w_t * p_t + f_1(\mathbf{v})$

where p_i is defined or trained from the data and $\{w_1, w_2, \ldots, w_n\}$ is sentence for image $\mathbf{I}$. This idea is converted into a form of structural neural function, where, we train neural network to define the functions.

B. Divide and Conquer Strategy In Deep Learning

The fundamental difference between "divide and conquer" strategy and neural ensemble strategy is that the former is focused on combination generation instead of utilization individually as concatenated series. Here, our approach is to divide the situation and later combine the inference, instead of expecting details in single function. While approximation is bounded through mathematical structures, the details of the functional approximation is operated in a "divide and conquer" algorithm, where, instead of relying on some weights to generalize well for all the images, we diversify the work to different weights for better generalization. This is a perfect example of "divide and conquer" strategy in deep neural network, mainly concentrating on task of decoding and related to sequence generation and processing. Sequence processing is computation intensive and contextual segregation is inevitable for proper generation of topologically significant sequences. While, most approaches are dependent on weighted transformation, the models lack generalization for sequence generalization and hence, we introduce the idea of Multiple Role Representation Crossover for better learning of usefulness from images. Mathematically, we can define this "divide and conquer" strategy in deep neural network as the following.

$$\mathbf{y} = \Phi(\mathbf{x}) = \Phi(\Psi_1(\mathbf{x}), \Psi_2(\mathbf{x}), \ldots, \Psi_n(\mathbf{x})) \qquad (7)$$

where we have defined $\Phi(.)$ as the transformation function of pre-likelihood and $\Psi_i(.)$ as the ith intermediate extraction of contexts. Instead of characterizing each function $\Psi_i(.)$ for specific purposes, we define $\Psi_i(.)$ as component function that take their required role and is trained to that for different situations.

C. Local Recurrent Strategy In Deep Learning

Local Recurrent Strategy is defined as identification of the most relevant and useful context (local context) in the search space instead of processing the whole context (global context). In this work, we have devised this approach for our strategy and is more like the swarm intelligence strategy, where the swarm always tries to follow some local best for improvement instead of global ones. Several local contexts will lead to the sentence, instead of learning a global context for the sentence. In this work, for sentence generation, the purpose is identification of some of the best local structural context for proper generation of the words for the sentence and these are

derived from lower level image information. Mathematically, we can consider the following equation as one kind of Local Recurrent Strategy.

$$\mathbf{y}_t = \Phi(\mathbf{I}) = \Phi(\mathbf{A}_t \mathbf{I}) \qquad (8)$$

where we have defined $\Phi(.)$ as the transformation function to pre-likelihood, the combination of local contexts $\mathbf{A}_t \mathbf{I}$ as context for the sentence at time t through the adjustment matrix $\mathbf{A}_t$. This can be regarded as Local Recurrent Strategy for the recurrent decoding, where the topology is established through variables at the language decoder level. Local Recurrent Strategy helps in diversification of contexts through "context shifts" strategies and is generated from the image features directly. Lastly, Local Recurrent Strategy helps in estimation of contexts that are not biased that is not learned through bias in the language decoder LSTM, where the learning of the language decoder plays important role.

IV. ATTENTION COMPOSITION ARCHITECTURES

The main motivation of introducing this architecture is the lack of previous architectures in defining a model that can capture different aspects of the situation. Attention is not generalized. But with series of local contexts can approach generalization and is a much powerful tool. Our architecture is composed of three components: Feature Distribution Composition (FDC) Attention, Multiple Role Representation Crossover (MRRC) Attention and Semantic Composition based Language decoder. FDC helps in feature composition from images, MRRC helps in selection through prediction of the next words through generation of semantics for the decoder LSTM layer. Different concepts of attentions are introduced and fused in the architecture that can generate the perfect combination for image captions. Figure 2 provided a diagrammatic overview of the architecture and subsequently, we provided more details of the individual components.

A. Feature Distribution Composition Attention

Feature Distribution Composition Attention is characterized for detention of the semantic composition from the regional objects features and the semantic composition expresses more with more regularization of the lower level extracted features. Previous works on this kind of architecture utilized lower level CNN based feature of images for caption generations. CNN feature provides a limited and selective overview of the images and the composition cannot be varied for the network to be leveraged upon. But, RCNN features consist of series of features in individual forms that can heuristically be selected to compose a constructive component. Individually, the RCNN features are much more sparse, but the weighted component composition can be much more meaningful and constructive. We define an effective and efficient way of weighted selection scheme and call it Feature Distribution Composition (FDC) Attention Layer. Compared to the bottom up approach [41], it is much more effective and lower in the number of weights and this scheme helped in out-perform previous works like [6]. This architecture will perform much better than [41], if compared on the same platform, as [41] is trained with data

from multiple sources. Mathematically, Feature Distribution Composition (FDC) Attention, denoted as $\widehat{v}_t$, can be derived by the following set of equations.

$$\overline{v} = \frac{1}{k} \sum_{i=1}^{i=k} v_i \qquad (9)$$

where $\{v_1, v_2, \ldots, v_k\}$ are the regional CNN features and $v_i \in \mathbb{R}^{2048} \; \forall \; i \in \{1, 2, \ldots, k\}$. The initial parameters of the LSTM for the language decoder is important as the hidden states are related to establishment of the topological relationship and time and we initialized them as the followings, considering that the FDC is related to the top overview of the image.

$$\mathbf{h}_0, \; \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \qquad (10)$$

where we have defined $\mathbf{a}_t \in \mathbb{R}^{b \times d}$, $\mathbf{a}_t \in \mathbb{R}^{b \times d}$ for transformation of the features. The next task of this model is development of the Intermediate Transfer Layer

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \qquad (11)$$

where $\mathbf{a}_t \in \mathbb{R}^{k \times 1}$, $\mathbf{W}_a \in \mathbb{R}^{m_x \times k}$, $\mathbf{W}_h \in \mathbb{R}^{d \times m_x}$, k is the number of considered maximum regional features in images.

$$\alpha_t = \text{softmax}(\mathbf{a}_t) \qquad (12)$$

where $\sum_{i=1}^{i=k} \alpha_{i,t} = 1$ and $\alpha_t \in \mathbb{R}^{k \times 1}$ and the Feature Distribution Composition Attention $\widehat{v}_t$ is provided as a summation of different regions to compose the attention.

$$\widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \qquad (13)$$

where we have $\widehat{v}_t \in \mathbb{R}^{b \times 2048}$ where b is the batch size and d is the hidden layer dimension. Figure 1 provided a pictorial overview of the Feature Distribution Composition (FDC) Attention architecture and the attention can be utilized for different applications. FDC helps in deciding the features at a very high level of the whole image and mainly concentrates on determination of the combination of the lower level features instead of pre-likelihood or distribution. Since, we experimented different architectures and proposed different fusion models, we will discuss the utility from each prospect.

B. Multiple Role Representation Crossover Attention

While, we utilized different features from the image for fusion in the network, like the global view for initialization and selection of weighted combination, we also introduce Multiple Role Representation Crossover (MRRC) Attention, where the aim is to provide attention, which is derived as a crossover of different aspect of the network. Residual and skip-through like architecture has been successful in many applications for gathering different residuals of the network. Some of our model do the same, where instead of utilization of the image features, it gathers the transformed features from some node of the network. Since, the depth of our MRRC architecture is very shallow, compared to ResNet, our residual collection can be any gate of the architecture, utilizing different aspects
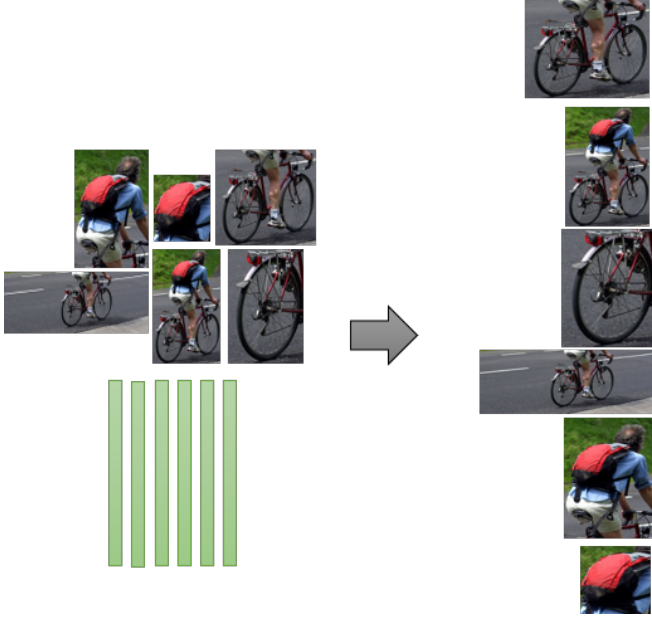
Fig. 1. Feature Distribution Composition (FDC) Attention Layer.

from images for generation and crossover the prospects to generate a new one with variations and developing the ability of the model to respond to those variations. Multiple Role Representation Crossover (MRRC) Attention Generation layer comprises of Image Weighted Tensors and are crossover-ed to generate the required attention dependent on time and a shifting weight scenario. The shifting of weights generate ample variations for the language decoder to be able to derive the sentences and generate diverse sentences. We can also call this attention as Image Weighted Tensor Crossover (IWTC) Attention when we use the FDC features as the prospects in place of $f(\mathbf{I})$ in Equation 14. Mathematically, Multiple Role Representation Crossover (MRRC) Attention Generation layer can be denoted by the following set of equations.

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}}\mathbf{h}_{t-1} + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(f(\mathbf{I}) \ \sigma(\mathbf{W}_{s_{21}}\mathbf{h}_{t-1} + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$(14)$$

where we have $\mathbf{x} \in \mathbb{R}^{b \times d}$. Figure 3 has given an overview of the Multiple Role Representation Crossover Attention architecture with more details of the involvement of the image characteristics and compositions. Multiple Role Representation Crossover Attention works on contexts related to the composition of the RCNN and thus can be considered as a dual prospect of FDC. In many situations, the weights of the neural network are biased to certain patterns and MRRC neutralizes the situation through extra weights and variances. This phenomenon is prevalent in statistics and is considered as interaction, where the variations are captured into the mathematical model through interaction term. However, due to the limitations in the number of considered variables in statistics, they are literally useless for intelligent systems.

Hence, in this work, we have devised such kind of situation, where MRRC is derived from FDC and $\mathbf{h}_{t-1}$. $\mathbf{h}_{t-1}$ is the next state derivation from the RCNN features and must not confused as the hidden state of LSTM. In this work, we are talking about MRRC architecture, which does not comply with LSTM, as we never propagate the hidden state (like in LSTM). Instead, we deal with transformed RCNN features. Hence, this architecture is far different from the traditional recurrent neural network architectures. It is to be noted that MRRC also deals with the previous contexts and thus can be regarded as a much enhanced criteria for context generation and the topologically synchronized attention.

C. Language Decoder

Language Decoder operates alongside the attentions and previous contexts. Unlike LSTM, this language decoder operates on the influence of attentions and previous context and does not explicitly depend on hidden states for sequential relevance. The equations for Language Decoder is provided as the followings.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_t + \mathbf{W}_{qi}\mathbf{q}_t + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \quad (15)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_t + \mathbf{W}_{qf}\mathbf{q}_t + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \quad (16)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_t + \mathbf{W}_{qo}\mathbf{q}_t + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \quad (17)$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_t + \mathbf{W}_{qg}\mathbf{q}_t + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \quad (18)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (19)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (20)$$

$$\mathbf{x}_t = \max \arg \mathrm{softmax}(\mathbf{W}_{hx}\mathbf{h}_t) \quad (21)$$

where the attentions $\mathbf{q}_t$ and $\mathbf{T}_t$ are derived from previous blocks and the $\mathbf{p}_t$ is the previous transformed context embedding. While, most of the recurrent units are dependent on attentions ($\mathbf{h}_{t-1}$, $\mathbf{A}_t$) and pretext ($\mathbf{W}_e \mathbf{x}_t$), our language model depends on combinations, which are generated out of the local context from the feature space and what to select as local context is learned in the feature space. Next, we will discuss the variations of architectures of these concepts and how they are unique in generating the combinations of contexts.

D. FDC-MRRC

Feature Distribution Composition - Multiple Role Representation Crossover (FDC-MRRC) is the basic architecture, where the main contributions are related to feature distribution generation and fusion of image weighted crossover attention. While FDC is about attention based on local enhancement, IWT provides a compositional construction that correlates with the local enhancement. This idea is motivated from the fact that both global and local prospect is important for diversified composition and establish local identity of feature space from the visual features. Figure 2 provided a diagrammatic overview of the FDC-MRRC architecture and their individual components. Mathematically, we can define FDC-MRRC with the following set of equations. Starting with regional image features $\{v_1, \ldots, v_n\}$ for image $\mathbf{I}$, $\overline{v} \in \mathbb{R}^{2048}$ is the average
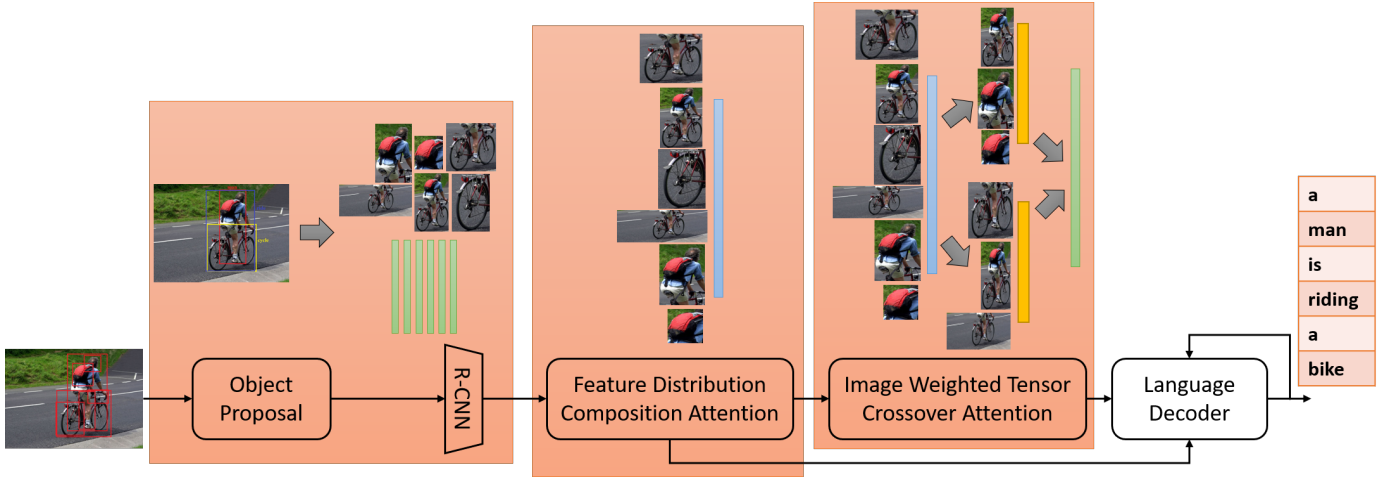
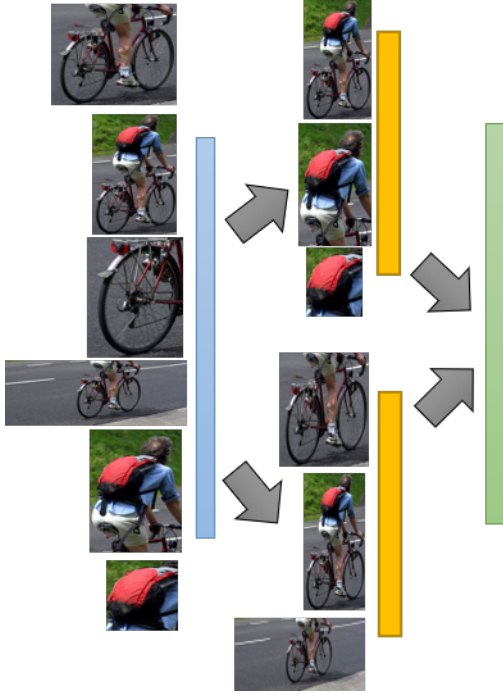Fig. 2. Overall FDC-MRRC Architecture With FDC, IWTC and RCNN Feature Fusion.

Fig. 3. Image Weighted Tensor Crossover (IWTC) Generation Layer. Here, the regional image segments are decomosed to generate newer ones with better comprehension and information.

of all the possibilities over the feature space that the image represent.

$$\overline{v} = \frac{1}{k} \sum_{i=1}^{i=k} v_i \qquad (22)$$

The initial parameters for the language decoder are initialized as the following for time $t = 0$.

$$\mathbf{h}_0, \ \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \qquad (23)$$

where the weights are defined as $\mathbf{W}_{h_0} \in \mathbb{R}^{2048 \times d}$, $\mathbf{W}_{h_0} \in \mathbb{R}^{2048 \times d}$ and these parameters effectively helps in determination of the FDC for the language decoder. The intermediate FDC component is defined as $\widehat{v}_t$, through the help of the

hidden components, which constitutes the status of previous composition and thus establishes the topological inference for the next instructions.

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \qquad (24)$$

where $\mathbf{a}_t \in \mathbb{R}^{k \times 1}$, $\mathbf{W}_a \in \mathbb{R}^{m_1 \times k}$, $\mathbf{W}_h \in \mathbb{R}^{d \times m_1}$, k is the number of considered regional features in images, m_1 is an intermediate dimension. Next, $\mathbf{a}_t$ is reduced to a softmax layer and the new α_t is used for FDC composition.

$$\alpha_t = \text{softmax}(\mathbf{a}_t) \qquad (25)$$

with $\alpha_t = \{\alpha_{1,t}, \ldots, \alpha_{k,t}\} \in \mathbb{R}^{k \times 1}$ and FDC composition is,

$$\widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \qquad (26)$$

where we have $\widehat{v}_t \in \mathbb{R}^{b \times 2048}$ where b is the batch size and d is the hidden layer dimension. We define new components $\mathbf{q}_t$ and $\mathbf{p}_t$ with $\mathbf{W}_e \in \mathbb{R}^{V \times e}$ as the embedding vector for languages and is derived from Stanford GloVe and V is the vocabulary size and e is the dimension of the embedding.

$$\mathbf{q}_t = \widehat{v}_t \qquad (27)$$

$$\mathbf{p}_t = \mathbf{W}_e \mathbf{x}_{t-1} \qquad (28)$$

Next, we define the MRRC feature selection layer as $\mathbf{T}_t$ as the following,

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} f(\mathbf{q}_t) + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}} f(\mathbf{q}_t) + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$= \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} \mathbf{h}_{t-1} + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}} \mathbf{h}_{t-1} + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$(29)$$

where we have defined $\otimes$ as an algebraic operation (like tensor product or dot product or similar) for the matrices to generate a feature vector using the different composition states of the regional image features in $\mathbf{h}_{t-1}$, which is more than hidden states of the language decoder. Here, we considered $\otimes = \odot$, element-wise multiplication, as we try to rectify and complement one context with the other context from the same image sources. Finally, we have,

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_t + \mathbf{W}_{qi}\mathbf{q}_t + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \tag{30}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_t + \mathbf{W}_{qf}\mathbf{q}_t + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \tag{31}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_t + \mathbf{W}_{qo}\mathbf{q}_t + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \tag{32}$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_t + \mathbf{W}_{qg}\mathbf{q}_t + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \tag{33}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{34}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{35}$$

where we have $\mathbf{h}_t$ as the pre-likelihood estimation of the topological sequence and is decoded as $\mathbf{x}_t$ as in Equation 21.

Mathematically, Feature Distribution Composition - Multiple Role Representation Crossover (FDC-MRRC), denoted as $f_{FDC_1}(.)$, can be described as the followings probability distribution estimation.

$$f_{FDC_1}(\mathbf{I}) = \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi(v_1 \dots, v_K), \ \mathbf{W}_{L_1})$$

$$\prod_i \Pr(\mathbf{T}_t \mid f_1(v_1 \dots, v_K), \ f_2(v_1 \dots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right), \mathbf{W}_{L_1}\right))$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \dots, v_K), \ f_2(v_1 \dots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k \mathrm{Q}_M(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right)\right))$$

$$\prod_t \mathrm{Q}_F(\mathbf{T}_t \mid f_1(v_1 \dots, v_K), \ f_2(v_1 \dots, v_K)) \tag{36}$$

using the weights of the LSTM in the architecture is denoted as $\mathbf{W}_{L_1}$, w_i as words of sentences, v_i as regional image features, $a_i v_m$ as intermediate learnt parameters, $\mathrm{Q}_M(.)$ and $\mathrm{Q}_F(.)$ are the Image Caption generator and feature generator function respectively. $\mathrm{Q}_F(.)$ derives $\mathbf{T}_i$ from $\mathbf{v}$ of $\mathbf{I}$.

E. Semi-Factorized-FDC-MRRC

Semi-Factorized Feature Distribution Composition - Multiple Role Representation Crossover (Semi-Factorized-FDC-MRRC) creates a factorized version of the $\mathbf{q}_t$ before being decoded for the topological interpretation. Figure 4 has provided a pictorial overview of the Semi-Factorized-FDC-MRRC architecture. From the prospect of feature composition, this is much more robust as these are dealing with factorization of the feature space to create diversity. Also, the FDC features are derived from regional local importance of the images based on

RPN (Region Prediction Network), whose works is to predict the regions of importance based on previously accounted probable regions. While, factorization for neural space is not equivalent to matrix factorization based on heuristics. This is because of the fact that such factorization do not scale well for applications and difficult to find the optimal space of factors. Also, for decoding tasks, such factorized model may not be helpful and do not fit well. Hence, a weighted factorization is introduced which helps in determination of partial usefulness of the feature vector that can be utilized for the model. Mathematically, we can define Semi-Factorized-FDC-MRRC as the following equations, where most of the notations are described for FDC-MRRC architecture.

$$\overline{v} = \frac{1}{k}\sum_{i=1}^{i=k} v_i \tag{37}$$

The initial parameters are initialized as,

$$\mathbf{h}_0, \ \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \tag{38}$$

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \tag{39}$$

$$\alpha_t = \mathrm{softmax}(\mathbf{a}_t) \tag{40}$$

$$\widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \tag{41}$$

The assembling and selector layer is defined for IWT as,

$$\mathbf{q}_t = \widehat{v}_t \tag{42}$$

$$\mathbf{p}_t = \mathbf{W}_e \mathbf{x}_{t-1} \tag{43}$$

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}}f(\mathbf{q}_t) + \mathbf{W}_{w_1}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}}f(\mathbf{q}_t) + \mathbf{W}_{w_2}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$= \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}}\mathbf{h}_{t-1} + \mathbf{W}_{w_1}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}}\mathbf{h}_{t-1} + \mathbf{W}_{w_2}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3) \tag{44}$$

The next equation introduces the factorization portion for the FDC $\mathbf{q}_t$ and is feed as $\mathbf{q}_{t,n}$ in place of each $\mathbf{q}_t$ for $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ and $\mathbf{g}_t$ of the language decoder. Equation 49 is called factorization step as the shared weights are decomposed into relevant slices of structured components with the help of a pre-trained distribution of the objects in the images. The main reason of factorization is to keep track of the composition of objects in images. Normally, FDC composition from RCNN objects feature representation will be heuristic and depend on the RPN prediction of Faster-RCNN, hence, factorization can help in establishing the structural properties between RCNN features and distribution of objects much faster. Unlike, previous works that operate on weights, directly related to the image features, our architecture is dependent on FDC composition and the factorized weights is expected to know the FDC composition than the whole image.
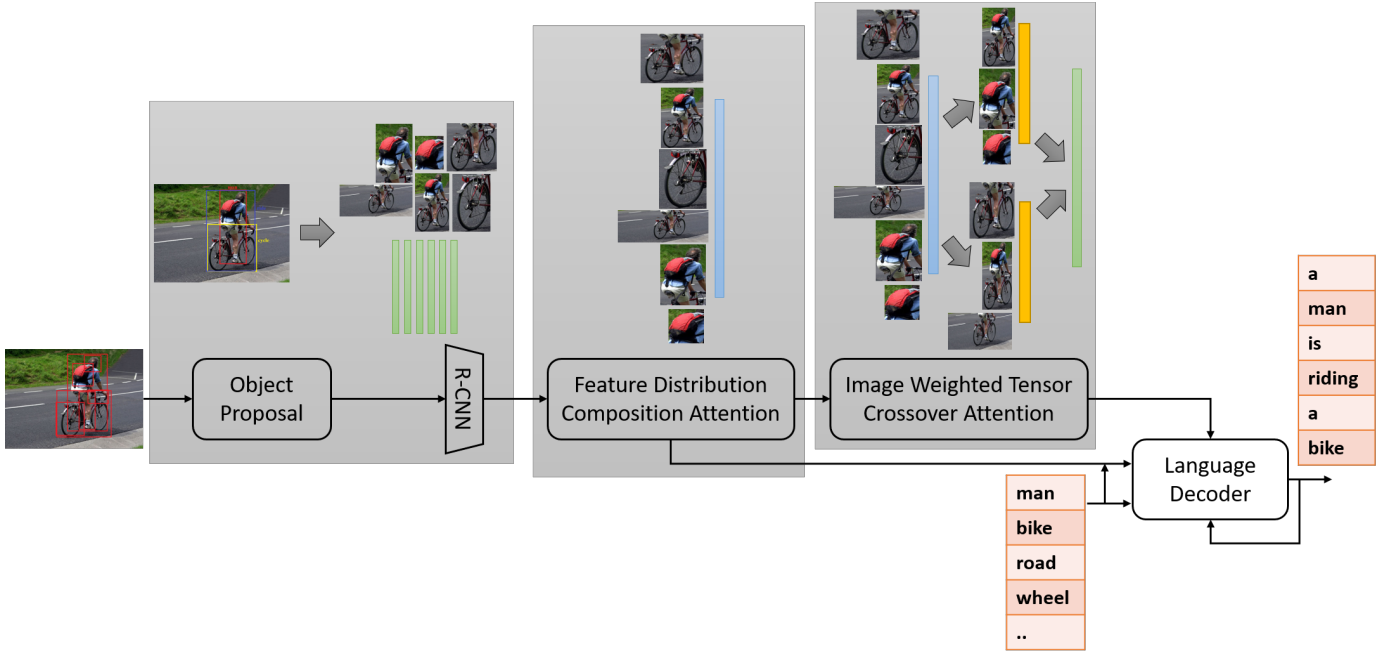
Fig. 4. Overall Semi-Factorized-FDC-MRRC Architecture With FDC, IWTC, Semantic and RCNN Feature Fusion.

Mathematically, Semi-Factorized Feature Distribution Composition - Multiple Role Representation Crossover (Semi-Factorized-FDC-MRRC), denoted as $f_{FDC_2}(.)$, can be described as the followings probability distribution estimation.

$$f_{FDC_2}(\mathbf{I}) = \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi(v_1\ldots,v_K), \ S, \ \mathbf{W}_{L_1})$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1\ldots,v_K), \ f_2(v_1\ldots,v_K), \ \mathbf{W}_1)$$

$$= \prod_k \Pr\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right), \ S\right)\right.$$

$$\left. , \ \mathbf{W}_{L_1}\right) \ \prod_t \Pr(\mathbf{T}_t \mid f_1(v_1\ldots,v_K), \ f_2(v_1\ldots,v_K), \ \mathbf{W}_1)$$

$$= \prod_k Q_M\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right), \ S\right)\right)$$

$$\prod_t Q_F(\mathbf{T}_t \mid f_1(v_1\ldots,v_K), \ f_2(v_1\ldots,v_K))$$

$$(45)$$

using the weights of the LSTM in the architecture is denoted as $\mathbf{W}_{L_1}$, w_i as words of sentences, v_i as regional image features, $a_i v_m$ as intermediate learnt parameters, $Q_M(.)$ and $Q_F(.)$ are the Image Caption generator and feature generator function respectively. $Q_F(.)$ derives $\mathbf{T}_i$ from $\mathbf{v}$ of $\mathbf{I}$, S is the semantic feature map.

1) Pseudo-Weights: The tags are pseudo-weights that are not learned but derived from a model and is expected to behave like weights. As an adaptive weight concept, these make different impacts on the feature composition. These are also called tag based weighted metrics and form a series of tensors working with each other. We can have any number of such tensors (parameters), however, for prove of concept and

feasibility of learning, we have considered three matrix product setup, reducing the learnable weights with high performing model derivatives. We consider $\mathbf{W}$ decomposition is given by,

$$\mathbf{W} = \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \mathbf{q}) \ \mathbf{W}_{p_s}$$
$$= \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \Phi(v_1, v_2, \ldots, v_k)) \ \mathbf{W}_{p_s} \quad (46)$$

where $\mathbf{W}_{p_r}$ and $\mathbf{W}_{p_s}$ are two matrices accessed by all possible training scenarios and captions, while $\text{diag}(\mathbf{W}_q \mathbf{q})$ or $\text{diag}(\mathbf{W}_q \Phi(v_1, v_2, \ldots, v_k))$ is a model derived transformation of the scenarios vectors. $\text{diag}(\mathbf{W}_q \mathbf{q})$ or $\text{diag}(\mathbf{W}_q \Phi(v_1, v_2, \ldots, v_k))$ also represents common linguistic topological features to establish the grammar in sentences and encourage appearance of adverb and adjective level complexity. $diag(.)$, accounts for k semantic aspects in the form of orthogonal spaces to capture the required features. Each slice of the weight tensor (diagonal matrices) corresponds to a concept and varies in size with the dataset.

2) Dynamic Factorization: In this work, we will discuss Dynamic Factorization, as this has the potential of demonstrating much better variation in representation for caption generation. Dynamic Factorization helps in better and optimized decomposition of the lower level components (regional features of the images) and creates the chances of rectification in case of low usefulness of the factorization context vectors (semantics) from images. Figure 5 provide a diagram differentiating Dynamic Factorization with Static Factorization. Mathematically, we can define this new factorization approach (Dynamic Factorization) as the following,

$$\mathbf{W} = \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \mathbf{q}) \ \mathbf{W}_{p_s}$$
$$= \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \Phi(v_1, v_2, \ldots, v_k)) \ \mathbf{W}_{p_s} \quad (47)$$
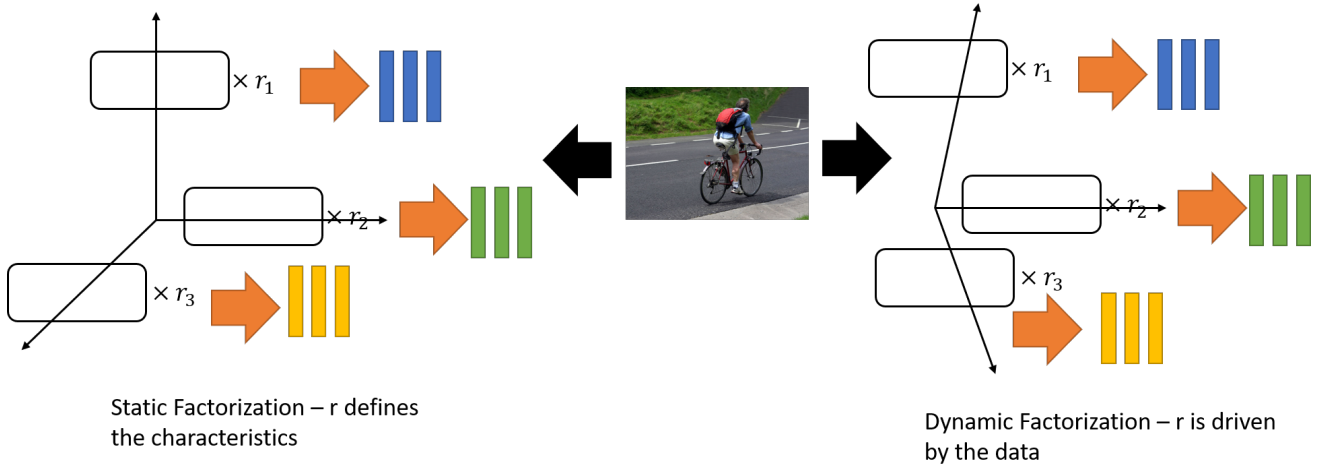
Fig. 5. Dynamic Factorization and Static Factorization

Compared to the previous factorization approach (Static Factorization), which were defined as,

$$\mathbf{W} = \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \mathbf{q}) \ \mathbf{W}_{p_s}$$
$$= \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q \Phi(\mathbf{v})) \ \mathbf{W}_{p_s} \quad (48)$$
$$= \mathbf{W}_{p_r} \ \text{diag}(\mathbf{W}_q S) \ \mathbf{W}_{p_s}$$

where $\Phi(.)$ is a functional transformation. Finally, we define the equation for the new $\mathbf{q}_{t,n}$ as,

$$\mathbf{q}_{t,n} = \mathbf{W}_{q,m} S \odot \mathbf{W}_{q,n} \mathbf{q}_t \quad (49)$$

where $\mathbf{q}_{t,n} \in \mathbb{R}^{b \times d}$, $\mathbf{q}_t \in \mathbb{R}^{b \times 999}$, $S \in \mathbb{R}^{b \times 999}$, $\mathbf{W}_{q,m} \in \mathbb{R}^{999 \times d}$, $\mathbf{W}_{q,n} \in \mathbb{R}^{2048 \times d}$. The final Language Decoder is defined as the followings and in place of $\mathbf{W}_{q*} \in \mathbb{R}^{2048 \times d}$, we have $\mathbf{W}_{q*} \in \mathbb{R}^{d \times d}$ for $* = i/f/o/g$ and for some value of dimension m_2.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_t + \mathbf{W}_{qi}\mathbf{q}_{t,n} + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \quad (50)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_t + \mathbf{W}_{qf}\mathbf{q}_{t,n} + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \quad (51)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_t + \mathbf{W}_{qo}\mathbf{q}_{t,n} + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \quad (52)$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_t + \mathbf{W}_{qg}\mathbf{q}_{t,n} + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \quad (53)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (54)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (55)$$

where $\mathbf{h}_t$ is pre-likelihood estimation of the topological sequence and is decoded as $\mathbf{x}_t$ as in Equation 21.

F. Full-Factorized-FDC-MRRC

Full-Factorized Feature Distribution Composition - Multiple Role Representation Crossover (Full-Factorized-FDC-MRRC) uses more factorized components for the language decoders for better enhancement of the features for the language decoder, where apart from decomposition of the weights for the image features FDC, the previous context embedding is also disintegrated. This disintegration of the language embedding brings stability in representation, which are unable to capture

(or approximated) by weights. Figure 6 provided the Full-Factorized-FDC-MRRC architecture details in a diagrammatic form for better perception of the network. The equations for the mathematical details of Full-Factorized-FDC-MRRC is denoted as the followings,

$$\overline{v} = \frac{1}{k} \sum_{i=1}^{i=k} v_i \quad (56)$$

$$\mathbf{h}_0, \ \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \quad (57)$$

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \quad (58)$$

$$\alpha_t = \text{softmax}(\mathbf{a}_t) \quad (59)$$

$$\widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \quad (60)$$

$$\mathbf{q}_t = \widehat{v}_t \quad (61)$$

$$\mathbf{p}_t = \mathbf{W}_e \mathbf{x}_{t-1} \quad (62)$$

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} f(\mathbf{q}_t) + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}} f(\mathbf{q}_t) + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$= \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} \mathbf{h}_{t-1} + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$

$$\tanh(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}} \mathbf{h}_{t-1} + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$

$$(63)$$

For every $* = i/f/o/g$, we define separate factorization of the different language components and can be defined as the followings.

$$\mathbf{p}_{*,t} = \mathbf{W}_{p,*m} S \odot \mathbf{W}_{p,*n} \mathbf{p}_t \quad (64)$$

$$\mathbf{q}_{*,t} = \mathbf{W}_{q,*m} S \odot \mathbf{W}_{q,*n} \mathbf{q}_t \quad (65)$$

where $\mathbf{W}_{p,*m} \in \mathbb{R}^{999 \times d}$, $\mathbf{W}_{p,*n} \in \mathbb{R}^{e \times d}$, e is the dimension of the embedding for $\mathbf{W}_e \in \mathbb{R}^{V \times e}$, V is the vocabulary size,
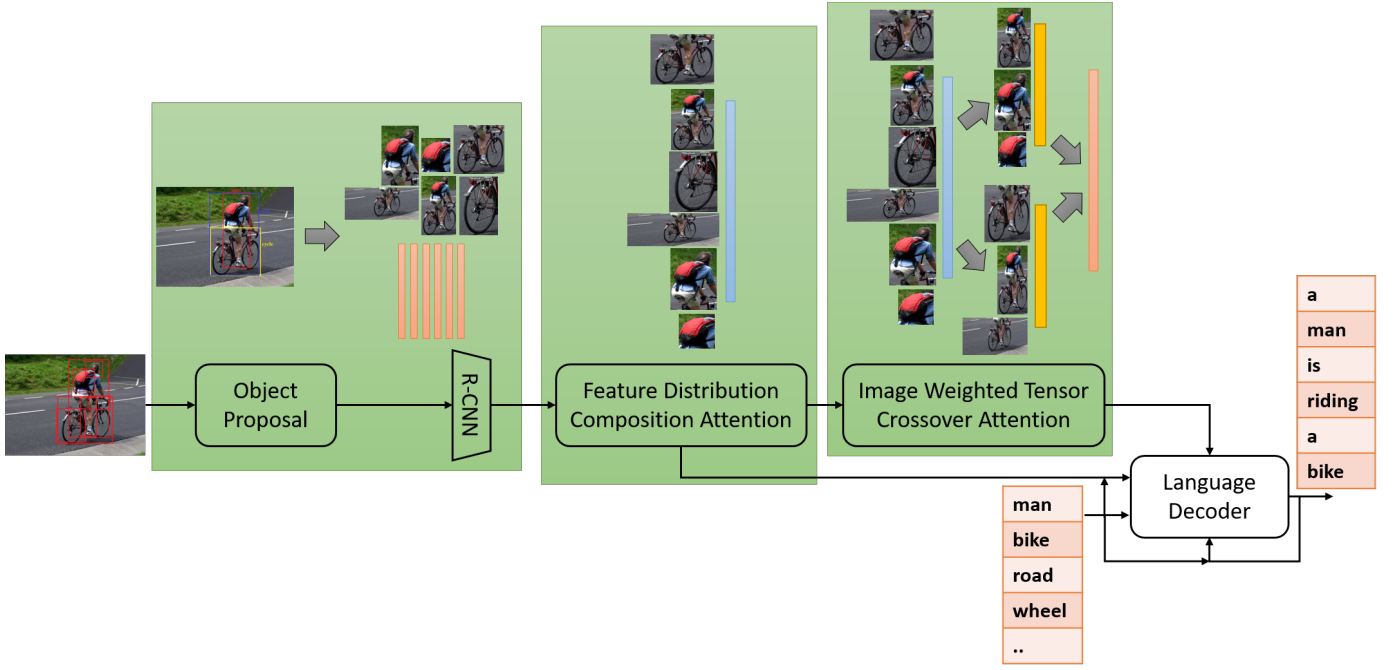
Fig. 6. Overall Full-Factorized-FDC-MRRC Architecture With FDC, IWTC, Semantic and RCNN Feature Fusion.

$\mathbf{W}_{q,*m} \in \mathbb{R}^{999 \times d}$, $\mathbf{W}_{q,*n} \in \mathbb{R}^{2048 \times d}$. It is a definite practice in neural network to factorize the language components. This is because of their concrete structures in feature space, acquired through transfer learning like word embedding and language distribution from images. However, these factorized weights help in better topic modeling and understanding the required direction of operation instead of estimating the representation into a definite direction through non-linear transformation based estimations of distribution. Unlike previous language decoders, we have $\mathbf{W}_{p*} \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_{q*} \in \mathbb{R}^{d \times d}$ for the following equations.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_{i,t} + \mathbf{W}_{qi}\mathbf{q}_{i,t} + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \quad (66)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_{f,t} + \mathbf{W}_{qf}\mathbf{q}_{f,t} + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \quad (67)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_{o,t} + \mathbf{W}_{qo}\mathbf{q}_{o,t} + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \quad (68)$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_{g,t} + \mathbf{W}_{qg}\mathbf{q}_{g,t} + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \quad (69)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (70)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (71)$$

where $\mathbf{h}_t$ is pre-likelihood estimation of the topological sequence and is decoded as $\mathbf{x}_t$ as in Equation 21.

Mathematically, Full-Factorized Feature Distribution Composition - Multiple Role Representation Crossover (Full-Factorized-FDC-MRRC), denoted as $f_{FDC_3}(.)$, can be described as the followings probability distribution estimation.

$$f_{FDC_3}(\mathbf{I}) = \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi(v_1 \ldots, v_K), \ S, \ f_L(\mathbf{W}_{L_1}, \ S))$$

$$\prod_i \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right), \ S),$$

$$f_L(\mathbf{W}_{L_1}, \ S)) \quad \prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k Q_M(w_k \mid \mathbf{T}_t, \ \Phi\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m \alpha_m\right), \ S))$$

$$\prod_t Q_F(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K))$$

$$(72)$$

using the weights of the LSTM in the architecture is denoted as $\mathbf{W}_{L_1}$, w_i as words of sentences, v_i as regional image features, $a_i v_m$ as intermediate learnt parameters, $Q_M(.)$ and $Q_F(.)$ are the Image Caption generator and feature generator function respectively. $Q_F(.)$ derives $\mathbf{T}_i$ from $\mathbf{v}$ of $\mathbf{I}$, S is the semantic feature map.

G. Semi-FDC-Factorized-Semantic-MRRC

Semi Feature Distribution Composition Factorized - Semantic - Multiple Role Representation Crossover (Semi-FDC-Factorized-Semantic-MRRC) operates on more refinement of the feature space decomposition through the utilization of semantic distribution for IWT Crossover Attention generation, which was previously dependent on regional image features.
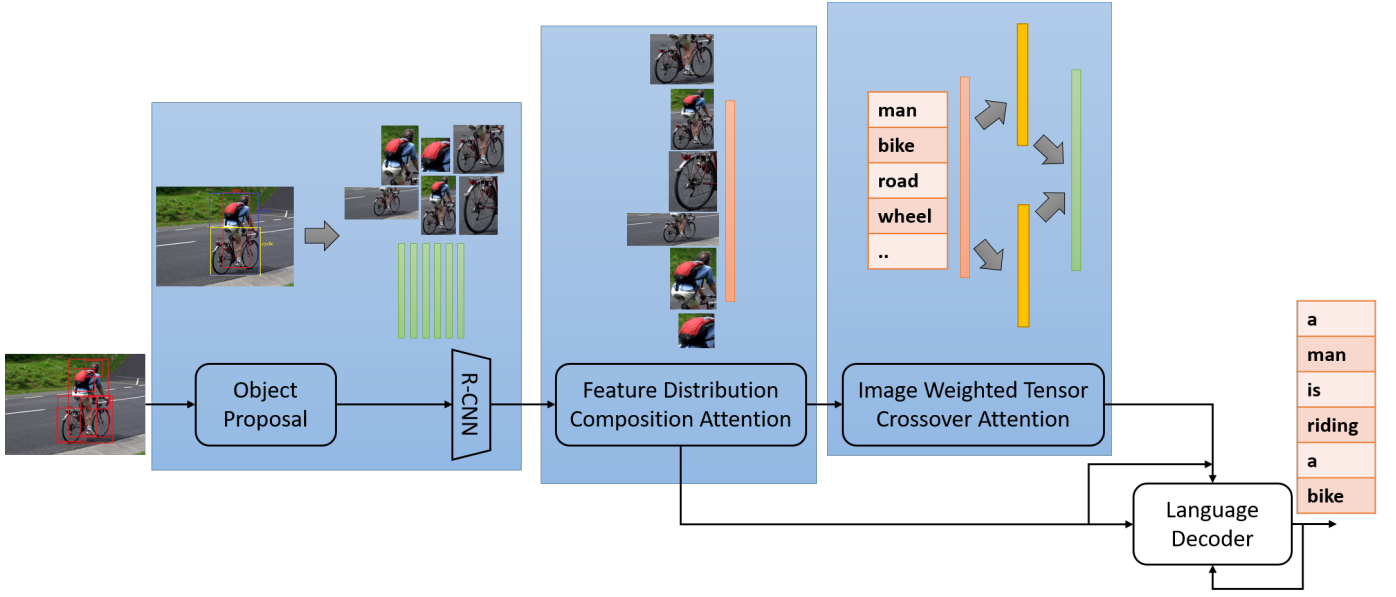
Fig. 7. Overall Semi-FDC-Factorized-Semantic-MRRC Architecture With FDC, IWTC, Semantic and RCNN Feature Fusion.

Regional image features are generated on the prediction of RPN and its position in the feature vector $\{v_1, v_2, \ldots, v_k\}$ is random as there is no criteria to establish the correctness of the sequence. However, since the features are extracted on a fine tuned model, it is expected that the model learned to establish the relative topological usefulness. Regional image features have this drawback and when we counter it through semantic based IWT Crossover attention, there ws enhancement in performance. Figure 7 provided the figure for Semi-FDC-Factorized-Semantic-MRRC architecture in details. Mathematical equations that govern the Semi-FDC-Factorized-Semantic-MRRC architecture can be denoted as the followings and most of the dimensions and notations are similar as before unless stated.

$$\overline{v} = \frac{1}{k} \sum_{i=1}^{i=k} v_i \tag{73}$$

$$\mathbf{h}_0, \ \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \tag{74}$$

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \tag{75}$$

$$\alpha_t = \mathrm{softmax}(\mathbf{a}_t) \tag{76}$$

$$\mathbf{q}_t = \widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \tag{77}$$

$$\mathbf{p}_t = \mathbf{W}_e \mathbf{x}_{t-1} \tag{78}$$

The MRRC equation has been replaced with a function of semantic level representation S as $f(S)$ has replaced $f(\mathbf{q}_t)$ to generate the MRRC attention. While, most of the factorization of vectors were helping in characterization, we introduce

MRRC attention as characterized attention and is defined as the following equation.

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} f(S) + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$
$$\tanh(\mathbf{W}_{s_{22}}(S \ \sigma(\mathbf{W}_{s_{21}} f(\mathbf{q}_t) + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3)$$
$$= \mathbf{W}_{s_{12}} \ \sigma(\mathbf{W}_{s_{11}} S + \mathbf{W}_{w_1} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_1) \otimes$$
$$\tanh(\mathbf{W}_{s_{22}}(S \ \sigma(\mathbf{W}_{s_{21}} \mathbf{h}_{t-1} + \mathbf{W}_{w_2} \sum_{i=0}^{t-1} \mathbf{W}_e \mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3) \tag{79}$$

The Language Decoder follows similar strategy as Semi-Factorized-FDC-MRRC, but the factorization is aided by the $\mathbf{q}_t$ instead of S. From this approach, we introduce the notion of a Dynamic Factorization scheme instead of Static Factorization. While Static Factorization is defined as Equation 48, Dynamic Factorization is defined as Equation 47. Here, we have used Dynamic Factorization where the factorized component changes with time and context and performed much better than previous models.

$$\mathbf{q}_{t,n} = \mathbf{W}_{h,m} f(\mathbf{q}_{t-1}) \odot \mathbf{W}_{h,n} \mathbf{q}_t = \mathbf{W}_{h,m} \mathbf{h}_{t-1} \odot \mathbf{W}_{h,n} \mathbf{q}_t \tag{80}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_t + \mathbf{W}_{qi}\mathbf{q}_{t,n} + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \tag{81}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_t + \mathbf{W}_{qf}\mathbf{q}_{t,n} + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \tag{82}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_t + \mathbf{W}_{qo}\mathbf{q}_{t,n} + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \tag{83}$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_t + \mathbf{W}_{qg}\mathbf{q}_{t,n} + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \tag{84}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{85}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{86}$$

where $\mathbf{h}_t$ is pre-likelihood estimation of the topological sequence and is decoded as $\mathbf{x}_t$ as in Equation 21.

Mathematically, Semi Feature Distribution Composition Factorized - Semantic - Multiple Role Representation Crossover (Semi-FDC-Factorized-Semantic-MRRC), denoted as $f_{FDC_4}(.)$, can be described as the followings probability distribution estimation.

$$f_{FDC_4}(\mathbf{I}) = \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi(v_1 \ldots, v_K), \ S, \ \mathbf{W}_{L_1})$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K, S), \ f_2(v_1 \ldots, v_K, S), \ \mathbf{W}_1)$$

$$= \prod_k \Pr\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right),\right.$$

$$\left.\left(\mathbf{W}_x\mathbf{h}_{t-1}\sum_{m=1}^{N} v_m\alpha_m \ \mathbf{W}_y\right), \ S\right), \ \mathbf{W}_{L_1})$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K, S), \ f_2(v_1 \ldots, v_K, S), \ \mathbf{W}_1)$$

$$= \prod_k Q_M\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m\alpha_m\right), \ S\right)\right)$$

$$\prod_t Q_F(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K, S), \ f_2(v_1 \ldots, v_K, S)) \tag{87}$$

using the weights of the LSTM in the architecture is denoted as $\mathbf{W}_{L_1}$, w_i as words of sentences, v_i as regional image features, $a_i v_m$ as intermediate learnt parameters, $Q_M(.)$ and $Q_F(.)$ are the Image Caption generator and feature generator function respectively. $Q_F(.)$ derives $\mathbf{T}_i$ from $\mathbf{v}$ of $\mathbf{I}$, S is the semantic feature map.

H. Semi-FDC-Factorized-FDC-MRRC

Semi Feature Distribution Composition - Multiple Role Representation Crossover (Semi-FDC-Factorized-FDC-MRRC) provided another architecture, where we explored Dynamic Factorization techniques along with MRRC with the regional image features. However, compared to Semi-FDC-Factorized-Semantic-MRRC, this model was outperformed due to the absence of semantic distribution component S. Figure 8 provided the diagram for the Semi-FDC-Factorized-FDC-MRRC model, used for the experiments. The mathematical representation of Semi-FDC-Factorized-FDC-MRRC is presented as the followings.

$$\overline{v} = \frac{1}{k}\sum_{i=1}^{i=k} v_i \tag{88}$$

$$\mathbf{h}_0, \ \mathbf{c}_0 = \mathbf{W}_{h_0}\overline{v}, \mathbf{W}_{c_0}\overline{v} \tag{89}$$

$$\mathbf{a}_t = \mathbf{W}_a \tanh(\mathbf{W}_h \mathbf{h}_{t-1}) \tag{90}$$

$$\alpha_t = \text{softmax}(\mathbf{a}_t) \tag{91}$$

$$\mathbf{q}_t = \widehat{v}_t = \sum_{i=1}^{i=k} v_i \alpha_{i,t} \tag{92}$$

$$\mathbf{p}_t = \mathbf{W}_e \mathbf{x}_{t-1} \tag{93}$$

There are $\mathbf{q}_t$ components for the MRRC representation $\mathbf{T}_t$ equation, while the rest Language Decoder equations are same as Semi-FDC-Factorized-Semantic-MRRC.

$$\mathbf{T}_t = \mathbf{W}_{s_{12}} \ \sigma\left(\mathbf{W}_{s_{11}}f(\mathbf{q}_t) + \mathbf{W}_{w_1}\sum_{i=0}^{t-1} \mathbf{W}_e\mathbf{x}_i + \mathbf{b}_1\right) \otimes$$

$$\tanh\left(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}}f(\mathbf{q}_t) + \mathbf{W}_{w_2}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3\right)$$

$$= \mathbf{W}_{s_{12}} \ \sigma\left(\mathbf{W}_{s_{11}}\mathbf{h}_{t-1} + \mathbf{W}_{w_1}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_1\right) \otimes$$

$$\tanh\left(\mathbf{W}_{s_{22}}(\mathbf{q}_t \ \sigma(\mathbf{W}_{s_{21}}\mathbf{h}_{t-1} + \mathbf{W}_{w_2}\sum_{i=0}^{t-1}\mathbf{W}_e\mathbf{x}_i + \mathbf{b}_2)) + \mathbf{b}_3\right) \tag{94}$$

$$\mathbf{q}_{t,n} = \mathbf{W}_{h,m}f(\mathbf{q}_{t-1})\odot\mathbf{W}_{h,n}\mathbf{q}_t = \mathbf{W}_{h,m}\mathbf{h}_{t-1}\odot\mathbf{W}_{h,n}\mathbf{q}_t \tag{95}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{pi}\mathbf{p}_t + \mathbf{W}_{qi}\mathbf{q}_{t,n} + \mathbf{W}_{Ti}\mathbf{T}_t + \mathbf{b}_i) \tag{96}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{pf}\mathbf{p}_t + \mathbf{W}_{qf}\mathbf{q}_{t,n} + \mathbf{W}_{Tf}\mathbf{T}_t + \mathbf{b}_f) \tag{97}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{po}\mathbf{p}_t + \mathbf{W}_{qo}\mathbf{q}_{t,n} + \mathbf{W}_{To}\mathbf{T}_t + \mathbf{b}_o) \tag{98}$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{pg}\mathbf{p}_t + \mathbf{W}_{qg}\mathbf{q}_{t,n} + \mathbf{W}_{Tg}\mathbf{T}_t + \mathbf{b}_g) \tag{99}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{100}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{101}$$

where $\mathbf{h}_t$ is pre-likelihood estimation of the topological sequence and is decoded as $\mathbf{x}_t$ as in Equation 21.

Mathematically, Semi Feature Distribution Composition - Multiple Role Representation Crossover (Semi-FDC-Factorized-FDC-MRRC), denoted as $f_{FDC_5}(.)$, can be described as the followings probability distribution estimation.

$$f_{FDC_4}(\mathbf{I}) = \prod_k \Pr(w_k \mid \mathbf{T}_t, \ \Phi(v_1 \ldots, v_K), \ S, \ \mathbf{W}_{L_1})$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k \Pr\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right),\right.$$

$$\left.\left(\mathbf{W}_x\mathbf{h}_{t-1}\sum_{m=1}^{N} v_m\alpha_m \ \mathbf{W}_y\right), \ S\right), \ \mathbf{W}_{L_1})$$

$$\prod_t \Pr(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K), \ \mathbf{W}_1)$$

$$= \prod_k Q_M\left(w_k \mid \mathbf{T}_t, \ \Phi\left(\left(\frac{1}{K}\sum_{m=1}^{K} v_m\right), \left(\sum_{m=1}^{N} v_m\alpha_m\right), \ S\right)\right)$$

$$\prod_t Q_F(\mathbf{T}_t \mid f_1(v_1 \ldots, v_K), \ f_2(v_1 \ldots, v_K)) \tag{102}$$

using the weights of the LSTM in the architecture is denoted as $\mathbf{W}_{L_1}$, w_i as words of sentences, v_i as regional image features, $a_i v_m$ as intermediate learnt parameters, $Q_M(.)$ and $Q_F(.)$ are the Image Caption generator and feature generator function respectively. $Q_F(.)$ derives $\mathbf{T}_i$ from $\mathbf{v}$ of $\mathbf{I}$, S is the semantic feature map.
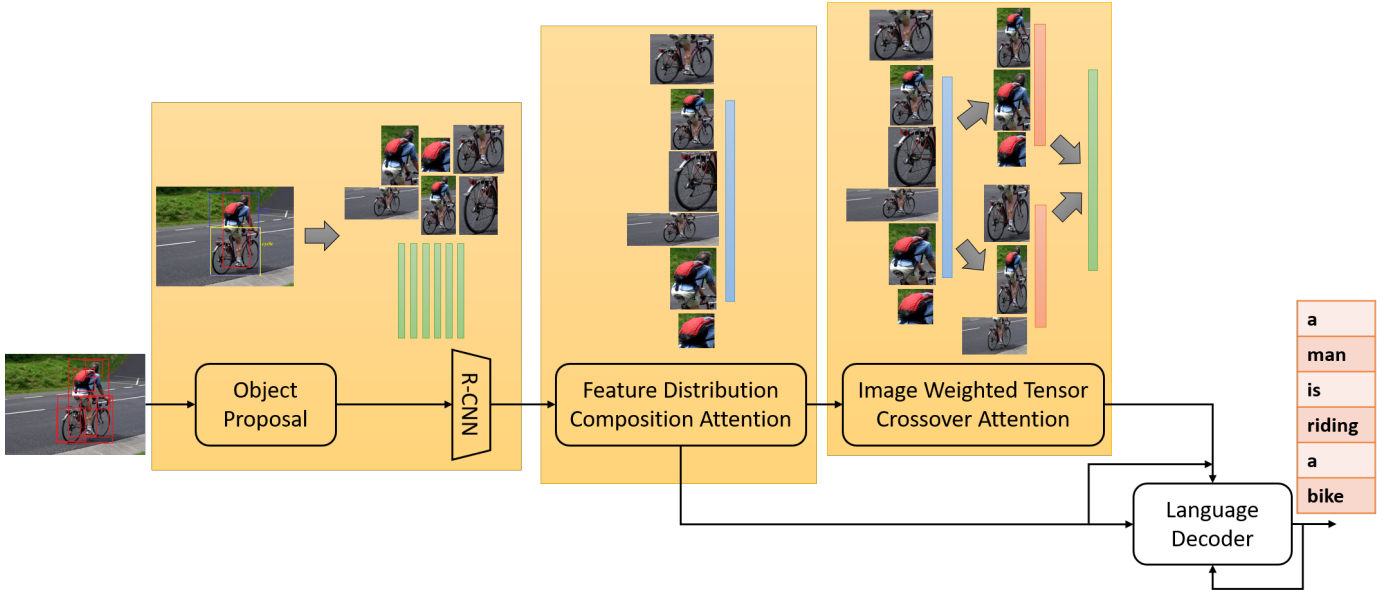
Fig. 8. Overall Semi-FDC-Factorized-FDC-MRRC Architecture With FDC, IWTC and RCNN Feature Fusion.

I. Reinforcement Learning

There are hardly any improvement with reinforcement learning [90] for bigger model as these architectures as the one-point error gradient can hardly provide scope of better optima for a system with so many variable weights. However, we did get some improvement for the best performing model Semi-FDC-Factorized-Semantic-MRRC. We define the Self-critical Sequence Training (SCST) [45] as Equation 103 or more precisely as Equation 104, where it utilizes the gradient of rewards associated with the sequences of generation for optimization and when the neural network is optimized with this reward difference, the procedure is known to be reinforcement learning, though the functional changes in neural network is not equivalent to the probability function determination. Mathematically, for image caption problem, the gradient of reward for reinforcement learning is defined as,

$$\frac{\delta L(\mathbf{w})}{\delta \mathbf{w}} = -\frac{1}{2b}\gamma \sum_i \Phi(\mathbf{y}, \mathbf{y}') \tag{103}$$

$$\frac{\delta L(\mathbf{w})}{\delta \mathbf{w}} = -\frac{1}{2b}\gamma \sum_i \Phi(\{y_1, \ldots, y_{c_1}\}, \{y'_1, \ldots, y'_{c_2}\}) \tag{104}$$

where $\Phi(.)$ is the evaluation function or the reward function that evaluates certain aspects of the generated captions $\{y_1, \ldots, y_c\} \in \mathbf{y}'$ and the baseline captions $\{y'_1, \ldots, y'_c\} \in \mathbf{y}$ and c_1 and c_2 is the length of the captions considered.

V. Results & Analysis

We performed different experiments with this data with varied dimension and different initialization approaches, along with different fine-tuning possibilities. The learning characteristics were better with $2-e$ learning rate while the best initialization strategy is with xavier initialization. Normal$(0, 0.05)$ strategy of initialization worked in many of our experiments, but it is time consuming and may be required to run the program for more than 20 epochs for better results. With resource and time constraints, we find xavier initialization had the best strategy, where the initialization is considered with a distribution spanned over the whole layer instead of the network. We used 100 dimension word embedding but got this transformed from 300 dimension word embedding from GloVe. The image features and the RCNN features used were derived from ResNet architecture, and we used the already extracted features from [73] and [41] works.

A. Dataset Description

In this part, we will provide some insights of the MSCOCO data and different tricks for training. MSCOCO consists of 123287 train images and 566747 train sentence, where each image is associated with at least five sentences from a vocabulary of 8791 words. There are 5000 images (with 25010 sentences) for validation and 5000 images (with 25010 sentences) for testing. We used the same data split as described in Karpathy et al [51]. For SCN network, two sets of image features are being used: one is ResNet features with 2048 dimension feature vector and another is a MLP transformed representation called tag features with feature vector of 999 dimension and consisted of the probability of occurrence for the most appearing set of tags as incidents in the dataset.

B. Training & Dimension Description

We trained the models, once at a time, mainly due to limited resources and establishing the concepts are more important than trying to outperform a model, whose data is not publicly available. we tried various combination of dimension, but found that 1024 was best and our K80 GPU can handle it with feasible time for experiments. However, there is very narrow improvement between the 512 and 1024 dimension model. We ran the experiments for around 20/25 epochs and stopped when there was no significant improvement on validation set.

C. Quantitative Analysis

Numerical results shows that our model outperformed much better than any of the older works, considering that we are using their extracted data. Gradual improvement in feature extraction will provide better improvements. There are many works being done, where they report improvements but never share the data. Our model is both theoretically and practically robust and given any data, it has the potential to outperform the base model. Table II provided the comparison of our results with all the previous works in this sector. We have only considered the works, where the data are available for comparison. Later, in Table I, we have provided another similar comparison, but with Reinforcement Learning technique for performance enhancement. However, Reinforcement Learning techniques like SCST never guarantee better performance, it keeps some scope open for experiments for these kinds of heuristic gradient to pull down the optimum to better optima, that can generate better captions based on a single comparison metrics.

D. Qualitative Analysis

Quantitative analysis can never the perfect judge for language generator, though the spectrum of the context can be judged statistically. Hence, we have included some instances of the generated captions, which has shown significant improvement in quality of the descriptions. Figure 9 and Figure 10 provided some instances of generated captions using all the architectures. It is difficult to judge from numerical results (in Table I or Table II) if a model is better than the other and didn't improved captions. The following qualitative analysis will reflect some of them as it is very difficult to conclude such claims from the quantitative numerical figures related to languages. Most of the caption generated work used diverse evaluation methods for the generated captions, but there are requirement of analysis of quality and grammatically correctness. There are captions which are true is some sense, but due to the limited proficiency of the language evaluation models, it has unable to capture them. Figure 9 and Figure 10 provided qualitative instances among various models.

VI. DISCUSSION

In this work, we analyzed the prospect of fusion of tensor product based structuring with semantic tag feature and RCNN features and demonstrated that these three combination can work well and create a state-of-the-art architecture for caption generation through adopting better representation generation and the ability to segregate the attrubutional features of the objects and activities from the images to the sentences. This architecture has adopted several existing features being offered through transfer learning and utilized these into a coherent model that is sensitive to different variations of the representation and thus help in better and longer sentence generation. The main contribution of this work is marked by the analysis of interaction of semantic features, RCNN features and tensor product. While RCNN is focused on regional objects, semantic features are focused on creation of a combination of objects and interaction useful for decoding of the word sequence,

and tensor product is used for orthogonal combination of the features that can be decoded for the segregation of the visual features. We were being able to create high performance with these combination and out-performed some of the present works.

REFERENCES

[1] Sur, C. (2019). Survey of deep learning and architectures for visual captioningtransitioning between media and natural languages. Multimedia Tools and Applications, 1-51.

[2] Sur, C. (2018). Representation for Language Understanding. Gainesville: University of Florida, pp.1-90. Available at: https://drive.google.com/file/d/15Fhmt5aM_b0J5jtE9mdWInQPfDS3TqVw .

[3] Q. Huang, L. Deng, D. Wu, C. Liu, X. He, "Attentive Tensor Product Learning," Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI), Honolulu, Hawaii, USA, January 27 – February 1, 2019.

[4] Palangi, H., Smolensky, P., He, X., & Deng, L. (2017). Question-answering with grammatically-interpretable representations. arXiv preprint arXiv:1705.08432.

[5] Lu, D., Whitehead, S., Huang, L., Ji, H., & Chang, S. F. (2018). Entity-aware Image Caption Generation. arXiv preprint arXiv:1804.07889.

[6] Lu, J., Yang, J., Batra, D., & Parikh, D. (2018, March). Neural Baby Talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7219-7228).

[7] You, Q., Jin, H., & Luo, J. (2018). Image Captioning at Will: A Versatile Scheme for Effectively Injecting Sentiments into Image Descriptions. arXiv preprint arXiv:1801.10121.

[8] Melnyk, I., Sercu, T., Dognin, P. L., Ross, J., & Mroueh, Y. (2018). Improved Image Captioning with Adversarial Semantic Alignment. arXiv preprint arXiv:1805.00063.

[9] Wu, J., Hu, Z., & Mooney, R. J. (2018). Joint Image Captioning and Question Answering. arXiv preprint arXiv:1805.08389.

[10] Kilickaya, M., Akkus, B. K., Cakici, R., Erdem, A., Erdem, E., & Ikizler-Cinbis, N. (2017). Data-driven image captioning via salient region discovery. IET Computer Vision, 11(6), 398-406.

[11] Chen, F., Ji, R., Su, J., Wu, Y., & Wu, Y. (2017, October). Structcap: Structured semantic embedding for image captioning. In Proceedings of the 2017 ACM on Multimedia Conference (pp. 46-54). ACM.

[12] Jiang, W., Ma, L., Chen, X., Zhang, H., & Liu, W. (2018). Learning to guide decoding for image captioning. arXiv preprint arXiv:1804.00887.

[13] Wu, C., Wei, Y., Chu, X., Su, F., & Wang, L. (2018). Modeling visual and word-conditional semantic attention for image captioning. Signal Processing: Image Communication.

[14] Fu, K., Li, J., Jin, J., & Zhang, C. (2018). Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs. IEEE Transactions on Neural Networks and Learning Systems, (99), 1-12.

[15] Chen, H., Ding, G., Lin, Z., Zhao, S., & Han, J. (2018). Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning. In IJCAI (pp. 606-612).

[16] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(2), 48.

[17] Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., & Qiao, Y. (2018). A Multi-task Learning Approach for Image Captioning. In IJCAI (pp. 1205-1211).

[18] Li, X., Wang, X., Xu, C., Lan, W., Wei, Q., Yang, G., & Xu, J. (2018). COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval. arXiv preprint arXiv:1805.08661.

[19] Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., & Han, J. (2017, February). Reference Based LSTM for Image Captioning. In AAAI (pp. 3981-3987).

[20] Tavakoliy, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017, October). Paying attention to descriptions generated by image captioning models. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2506-2515). IEEE.

[21] Chen, H., Zhang, H., Chen, P. Y., Yi, J., & Hsieh, C. J. (2017). Show-and-fool: Crafting adversarial examples for neural image captioning. arXiv preprint arXiv:1712.02051.

[22] Ye, S., Liu, N., & Han, J. (2018). Attentive Linear Transformation for Image Captioning. IEEE Transactions on Image Processing.

TABLE I
PERFORMANCE EVALUATION AND COMPARISON BETWEEN DIFFERENT ARCHITECTURES WITHOUT REINFORCEMENT LEARNING

Algorithm	CIDEr-D	Bleu_4	Bleu_3	Bleu_2	Bleu_1	ROUGE_L	METEOR	SPICE
Human [46]	0.85	0.22	0.32	0.47	0.66	0.48	0.2	–
Neural Talk [51]	0.66	0.23	0.32	0.45	0.63	–	0.20	–
MindsEye [53]	–	0.19	–	–	–	–	0.20	–
Google [47]	0.94	0.31	0.41	0.54	0.71	0.53	0.25	–
LRCN [55]	0.87	0.28	0.38	0.53	0.70	0.52	0.24	–
Montreal [49]	0.87	0.28	0.38	0.53	0.71	0.52	0.24	–
m-RNN [60]	0.79	0.27	0.37	0.51	0.68	0.50	0.23	–
[76]	0.81	0.26	0.36	0.49	0.67	–	0.23	–
MSR [50]	0.91	0.29	0.39	0.53	0.70	0.52	0.25	–
[57]	0.84	0.28	0.38	0.52	0.70	–	0.24	–
bi-LSTM [44]	–	0.244	0.352	0.492	0.672	–	–	–
MSR Captivator [54]	0.93	0.31	0.41	0.54	0.72	0.53	0.25	–
Nearest Neighbor [87]	0.89	0.28	0.38	0.52	0.70	0.51	0.24	–
MLBL [86]	0.74	0.26	0.36	0.50	0.67	0.50	0.22	–
ATT [70]	0.94	0.32	0.42	0.57	0.73	0.54	0.25	–
[46]	0.92	0.31	0.41	0.56	0.73	0.53	0.25	–
Adaptive [39]	1.085	0.332	0.439	0.580	0.742	–	0.266	–
MSM [38]	0.986	0.325	0.429	0.565	0.730	–	0.251	–
ERD [88]	0.895	0.298	–	–	–	–	0.240	–
Att2in [45]	1.01	0.313	–	–	–	–	0.260	–
NBT [6]	1.07	0.347	–	–	0.755	–	0.271	0.201
Attribute-Attention [15]	1.044	0.338	0.443	0.579	0.743	0.549	–	–
LSTM [73]	0.889	0.292	0.390	0.525	0.698	–	0.238	–
SCN [73]	1.012	0.330	0.433	0.566	0.728	–	0.257	–
Up-Down** [41]	1.054	0.334	–	–	0.745	0.544	0.261	0.192
Up-Down† [41]	1.135	0.362	–	–	0.772	0.564	0.270	0.203
FDC-MRRC	1.073	0.351	0.459	0.595	0.752	0.557	0.265	0.198
Semi-Factorized-FDC-MRRC	1.075	0.349	0.456	0.591	0.748	0.557	0.265	0.195
Full-Factorized-FDC-MRRC	1.078	0.349	0.456	0.592	0.751	0.556	0.266	0.196
Semi-FDC-Factorized-Semantic-MRRC	**1.080**	**0.352**	**0.460**	**0.598**	**0.755**	**0.559**	**0.265**	**0.197**
Semi-FDC-Factorized-FDC-MRRC	1.073	0.348	0.453	0.589	0.749	0.554	0.265	0.195

**Visual Feature from ResNet Architecture is used
†R-CNN Visual Feature from ResNet Architecture is Used

TABLE II
PERFORMANCE EVALUATION AND COMPARISON BETWEEN DIFFERENT ARCHITECTURES WITH REINFORCEMENT LEARNING FINE-TUNING

Algorithm	CIDEr-D	Bleu_4	Bleu_3	Bleu_2	Bleu_1	ROUGE_L	METEOR	SPICE
Adaptive [39]	1.085	0.332	0.439	0.580	0.742	–	0.266	–
MSM [38]	0.986	0.325	0.429	0.565	0.730	–	0.251	–
ERD [88]	0.895	0.298	–	–	–	–	0.240	–
Att2in [45]	1.01	0.313	–	–	–	–	0.260	–
NBT [6]	1.07	0.347	–	–	0.755	–	0.271	0.201
[15]	1.044	0.338	0.443	0.579	0.743	0.549	–	–
LSTM [73]	0.889	0.292	0.390	0.525	0.698	–	0.238	–
SCN [73]	1.012	0.330	0.433	0.566	0.728	–	0.257	–
Up-Down** [41]	1.054	0.334	–	–	0.745	0.544	0.261	0.192
Up-Down† [41]	1.135	0.362	–	–	0.772	0.564	0.270	0.203
Up-Down** + RL [41]	1.111	0.340	–	–	0.766	0.549	0.265	0.202
Up-Down† + RL [41]	1.201	0.363	–	–	0.798	0.569	0.277	0.214
FDC-MRRC	1.078	0.351	0.457	0.598	0.754	0.557	0.267	0.199
Semi-Factorized-FDC-MRRC	1.079	0.350	0.457	0.593	0.749	0.556	0.265	0.195
Full-Factorized-FDC-MRRC	1.075	0.351	0.457	0.593	0.751	0.555	0.266	0.196
Semi-FDC-Factorized-Semantic-MRRC	**1.082**	**0.353**	**0.460**	**0.597**	**0.753**	**0.557**	**0.266**	**0.197**
Semi-FDC-Factorized-FDC-MRRC	1.075	0.351	0.457	0.592	0.749	0.557	0.264	0.196

**Visual Feature from ResNet Architecture is used
†R-CNN Visual Feature from ResNet Architecture is Used

Fig. 9. Qualitative Analysis. Part 1. Here, 1 → Semi-FDC-Factorized-Semantic-MRRC : 2 → Semi-FDC-Factorized-FDC-MRRC 3 → FDC-MRRC 4 → Semi-Factorized-FDC-MRRC 5 → Full-Factorized-FDC-MRRC

[23] Wang, Y., Lin, Z., Shen, X., Cohen, S., & Cottrell, G. W. (2017). Skeleton key: Image captioning by skeleton-attribute decomposition. arXiv preprint arXiv:1704.06972.

[24] Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., & Luo, J. (2018). "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention. arXiv preprint arXiv:1807.03871.

[25] Chen, F., Ji, R., Sun, X., Wu, Y., & Su, J. (2018). GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1345-1353).

[26] Liu, C., Sun, F., Wang, C., Wang, F., & Yuille, A. (2017). MAT: A multimodal attentive translator for image captioning. arXiv preprint arXiv:1702.05658.

[27] Harzig, P., Brehm, S., Lienhart, R., Kaiser, C., & Schallner, R. (2018). Multimodal Image Captioning for Marketing Analysis. arXiv preprint arXiv:1802.01958.

[28] Liu, X., Li, H., Shao, J., Chen, D., & Wang, X. (2018). Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. arXiv preprint arXiv:1803.08314.

[29] Chunseong Park, C., Kim, B., & Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 895-903).

[30] Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 2556-2565).

[31] Yao, T., Pan, Y., Li, Y., & Mei, T. (2017, July). Incorporating copying mechanism in image captioning for learning novel objects. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5263-5271). IEEE.

[32] Zhang, L., Sung, F., Liu, F., Xiang, T., Gong, S., Yang, Y., & Hospedales, T. M. (2017). Actor-critic sequence training for image captioning. arXiv preprint arXiv:1706.09601.

[33] Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2017). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2321-2334.

[34] Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep

Fig. 10. Qualitative Analysis. Part 2. Here, 1 → Semi-FDC-Factorized-Semantic-MRRC : 2 → Semi-FDC-Factorized-FDC-MRRC 3 → FDC-MRRC 4 → Semi-Factorized-FDC-MRRC 5 → Full-Factorized-FDC-MRRC

reinforcement learning-based image captioning with embedding reward. arXiv preprint arXiv:1704.03899.

[35] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017, October). Improved image captioning via policy gradient optimization of spider. In Proc. IEEE Int. Conf. Comp. Vis (Vol. 3, p. 3).

[36] Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically Informative Image Captioning with Character-Level Reference. arXiv preprint arXiv:1804.05417.

[37] Liu, C., Mao, J., Sha, F., & Yuille, A. L. (2017, February). Attention Correctness in Neural Image Captioning. In AAAI (pp. 4176-4182).

[38] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017, October). Boosting image captioning with attributes. In IEEE International Conference on Computer Vision, ICCV (pp. 22-29).

[39] Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017, July). 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, p. 2).

[40] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[41] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In CVPR (Vol. 3, No. 5, p. 6).

[42] Zhang, M., Yang, Y., Zhang, H., Ji, Y., Shen, H. T., & Chua, T. S. (2018). More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining. IEEE Transactions on Image Processing.

[43] Park, C. C., Kim, B., & Kim, G. (2018). Towards Personalized Image Captioning via Multimodal Memory Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[44] Wang, Cheng, Haojin Yang, and Christoph Meinel. "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14.2s (2018): 40.

[45] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017, July). Self-critical sequence training for image captioning. In CVPR (Vol. 1, No. 2, p. 3).

[46] Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine

intelligence.

[47] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[48] Karpathy, Andrej, Armand Joulin, and Fei Fei F. Li. "Deep fragment embeddings for bidirectional image sentence mapping." Advances in neural information processing systems. 2014.

[49] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International Conference on Machine Learning. 2015.

[50] Fang, Hao, et al. "From captions to visual concepts and back." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[51] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[52] Anne Hendricks, Lisa, et al. "Deep compositional captioning: Describing novel object categories without paired training data." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[53] Chen, Xinlei, and C. Lawrence Zitnick. "Mind's eye: A recurrent visual representation for image caption generation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[54] Devlin, Jacob, et al. "Language models for image captioning: The quirks and what works." arXiv preprint arXiv:1505.01809 (2015).

[55] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[56] Gan, Chuang, et al. "Stylenet: Generating attractive visual captions with styles." CVPR, 2017.

[57] Jin, Junqi, et al. "Aligning where to see and what to tell: image caption with region-based attention and scene factorization." arXiv preprint arXiv:1506.06272 (2015).

[58] Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models." arXiv preprint arXiv:1411.2539 (2014).

[59] Kiros, Ryan, Richard Zemel, and Ruslan R. Salakhutdinov. "A multiplicative model for learning distributed text-based attribute representations." Advances in neural information processing systems. 2014.

[60] Mao, Junhua, et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)." arXiv preprint arXiv:1412.6632 (2014).

[61] Memisevic, Roland, and Geoffrey Hinton. "Unsupervised learning of image transformations." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

[62] Pu, Yunchen, et al. "Variational autoencoder for deep learning of images, labels and captions." Advances in Neural Information Processing Systems. 2016.

[63] Socher, Richard, et al. "Grounded compositional semantics for finding and describing images with sentences." Transactions of the Association for Computational Linguistics 2 (2014): 207-218.

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

[65] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[66] LTran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

[67] Tran, Kenneth, et al. "Rich image captioning in the wild." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016.

[68] Wu, Qi, et al. "What value do explicit high level concepts have in vision to language problems?." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[69] Yang, Zhilin, et al. "Review networks for caption generation." Advances in Neural Information Processing Systems. 2016.

[70] You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[71] Young, Peter, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions." Transactions of the Association for Computational Linguistics 2 (2014): 67-78.

[72] Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[73] Gan, Zhe, et al. "Semantic compositional networks for visual captioning." arXiv preprint arXiv:1611.08002 (2016).

[74] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[75] Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research47 (2013): 853-899.

[76] Jia, Xu, et al. "Guiding the long-short term memory model for image caption generation." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[77] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International Journal of Computer Vision 123.1 (2017): 32-73.

[78] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

[79] Li, Siming, et al. "Composing simple image descriptions using web-scale n-grams." Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011.

[80] Kuznetsova, Polina, et al. "TREETALK: Composition and Compression of Trees for Image Descriptions." TACL 2.10 (2014): 351-362.

[81] Mao, Junhua, et al. "Learning like a child: Fast novel visual concept learning from sentence descriptions of images." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[82] Mathews, Alexander Patrick, Lexing Xie, and Xuming He. "SentiCap: Generating Image Descriptions with Sentiments." AAAI. 2016.

[83] Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[84] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." Advances in Neural Information Processing Systems. 2011.

[85] Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[86] Kiros, R., Salakhutdinov, R., Zemel, R. (2014, January). Multimodal neural language models. In International Conference on Machine Learning (pp. 595-603).

[87] Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C. L. (2015). Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467.

[88] Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., Cohen, W. W. Encode, Review, and Decode: Reviewer Module for Caption Generation. arXiv 2016. arXiv preprint arXiv:1605.07912.

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

[90] Sur, Chiranjib. "UCRLF: unified constrained reinforcement learning framework for phase-aware architectures for autonomous vehicle signaling and trajectory optimization." Evolutionary Intelligence (2019): 1-24.