Attention-based Context Aware Reasoning for Situation Recognition

Thilini Cooray, Ngai-Man Cheung*, Wei Lu
Singapore University of Technology and Design (SUTD)
thilini_cooray@mymail.sutd.edu.sg, {ngaiman_cheung, luwei}@sutd.edu.sg

Abstract

Situation Recognition (SR) is a fine-grained action recognition task where the model is expected to not only predict the salient action of the image, but also predict values of all associated semantic roles of the action. Predicting semantic roles is very challenging: a vast variety of possibilities can be the match for a semantic role. Existing work has focused on dependency modelling architectures to solve this issue. Inspired by the success achieved by query-based visual reasoning (e.g., Visual Question Answering), we propose to address semantic role prediction as a query-based visual reasoning problem. However, existing query-based reasoning methods have not considered handling of ***inter-dependent queries*** *which is a unique requirement of semantic role prediction in SR. Therefore, to the best of our knowledge, we propose the first set of methods to address inter-dependent queries in query-based visual reasoning. Extensive experiments demonstrate the effectiveness of our proposed method which achieves outstanding performance on Situation Recognition task. Furthermore, leveraging query inter-dependency, our methods improve upon a state-of-the-art method that answers queries separately. Our code: https://github.com/thilinicooray/context-aware-reasoning-for-sr*

1. Introduction

Visual reasoning is the process of analyzing visual information in order to achieve a final conclusion. There are a variety of visual reasoning tasks being researched in the computer vision domain beginning with the basic building blocks of object [14, 24, 22, 8] and action [4, 23, 25] classification. Scene Graph Generation [11, 17, 28] was introduced in order to expand the visual reasoning capabilities of computer vision models beyond mere object and action classification and brought visual reasoning to the next level by combining all the predicted visual relations in an image and constructing a knowledge graph out of it.

Brushing			
Role	**Value**	**Role**	**Value**
Agent	Woman	Agent	Man
Target	Hair	Target	Teeth
Tool	Brush	Tool	Toothbrush
Substance	-	Substance	Toothpaste

Figure 1. Situation recognition (SR) [30]: Two different situations for the same action (verb). The SR task is to predict the action (verb) and the values of all the associated semantic roles.

However, these relations in scene graphs were captured in a triplet (subject-predicate-object) manner which limits the expressibility when it comes to describe actions, as the objects participate in an action expand beyond subject and object elements. In order to address this limitation, Yatskar et al. [30] introduced *Situation Recognition (SR)*. In SR, the model is expected to not only predict the salient action of the image, but also predict all the objects that participate in the action. Relationships between individual objects and the action are indicated by a concept called *semantic roles*. A situation is a structure which comprises of an action along with its semantic roles making this a structured prediction task.

Figure 1 shows two instances of action "Brushing" in the *imSitu* dataset [30], the prime dataset for SR. Semantic roles of "Brushing" are *agent* (person who is brushing), *target* (entity or object the agent is brushing), *tool* (the tool being used for brushing), *substance* (any substance being used for brushing). Note that *place* is also a semantic role for brushing, but we omit it in this example for clarity as it is not significant here. Also note that different actions may have different semantic roles. For example, action "eating"

*Corresponding Author

has roles: *food*, *place*, *container*, *agent*, *tool*. SR is a very challenging reasoning task, as the number of different role types and possible values are very large [20, 16]. Furthermore, even for the same action (verb), the possible values for individual roles can be very different as illustrated in Figure 1.

Semantic role prediction has drawn the most attention compared to action prediction due to its more challenging requirement of capturing all action related objects in the image, regardless of its visible salience. Existing work has focused on modelling inter-dependency among semantic roles using Recurrent Neural Networks [20] and Graph Neural Networks [16].

In this work, we take a radically different approach for SR. Inspired by query-based visual reasoning models [7, 10, 33, 13] which have proven to be successful in analyzing an image conditioned on a given query (natural language question, object name etc.) to obtain an answer, we propose to model SR as a query-based visual reasoning task. In particular, we propose a novel visual reasoning model which focuses on reasoning the image based on given queries rather than emphasizing object co-occurrence patterns during training. However, one major challenge that SR introduces (which does not exist for conventional query-based visual reasoning tasks) is that, while other tasks require single output answer (e.g. Visual Question Answering [7, 10, 9, 1, 12, 3]), SR expects answers to *multiple inter-dependent queries* which finally forms a structure.

To fill this gap of handling *inter-dependent queries*, we make the first effort by proposing a novel contextualization module to incorporate information from related queries to address inter-query relational reasoning. Our contextualization mechanism explicitly allows both multi-modal reasoning and neighbour information integration together. This enables the model to dynamically combine the information for optimal predictions. We propose a method to generate the context using attention, and propose different mechanisms to incorporate the generated context to improve reasoning. Our contributions are:

- We propose to address SR via query-based visual reasoning.
- We propose novel methods to handle inter-dependent queries that arise in semantic role prediction in SR
- We perform extensive experiments to validate our methods.

2. Related Work

Yatskar et al. introduced the SR task along with the *imSitu* dataset whose actions and frames are based on FrameNet [2]. They proposed a baseline model which consists of a Convolutional Neural Network (CNN) [15] for image encoding followed by a Conditional Random Field to predict actions and labels for semantic roles. As mentioned by Yatskar et al. [30], this dataset suffers from huge sparsity issues in both object labels as well as situations because some objects can participate in many roles while other objects can only be seen few times. To address this sparsity issue, Yatskar et al. [29] later proposed another model which maps roles and labels to a lower dimensional vector space and have also used additional images to reduce data sparsity. Then two models were presented by Mallya and Lazebnik [20] and Li et al. [16] focusing on improving role predictions by explicitly modelling dependency among semantic roles. Mallya and Lazebnik [20] use a Recurrent Neural Net to model role dependencies and predict labels as a sequence labelling problem while using a Fusion Network [19] for action prediction. Li et al. [16] argue that all roles in a frame should depend on each other without manually assigning any priority to roles like in sequence labelling. Therefore they propose a Gated Graph Neural Network (GGNN) [18] based role modelling method. These two models achieve the highest results for frame prediction emphasizing the importance of modelling role inter-dependency for this task.

On the subject of improving multi-modal reasoning for independent query predictions, Visual Question Answering (VQA) [7, 10, 9, 1, 12, 3] task leads the way with numerous highly capable multi-modal reasoning methods. Inspired by these, we utilize a very simple, but effective VQA method by Anderson et al. [1] to fill the lack of sophisticated multi-modal reasoning application in SR. However, existing VQA tasks only require answering questions independently or use answers from previous questions to answer the current question (ex: Visual Dialog [6] and Visual Commonsense Reasoning (VCR) [32]). SR stands out from these as mentioned earlier that each role (the query to which we try to find an answer) depends on all other roles of its action without any defined order like in Visual Dialog or VCR.

Inter-dependent question answering is a novel requirement in SR which has not been raised before. We believe this has the potential to be useful for other tasks such as Embodied Question Answering [5] in multi-agent environments where agents can utilize information from each-other along with its own surrounding to answer questions. Therefore in this work, we propose several models which are capable of inter-dependent VQA, aiming to solve semantic role prediction in SR.

3. Context Aware Visual Reasoning for Situation Recognition

3.1. Task Definition

Situation Recognition defines a space which consists of a discrete set of verbs V, nouns N, roles R and frames F. Each verb $v \in \{1, \ldots, |V|\}$ is mapped with a frame $f \in F$

which consists of semantic roles $R_v \subset R$. Each semantic role is paired with a noun value $n \in N \cup \{\emptyset\}$. An instance of an action v in an image I forms a realized frame $F_{(I,v)} = \{(r_i, n_i) : r_i \in R_v, n_i \in N \cup \{\emptyset\}, i = 1, \ldots, \mid R_v \mid\}$. Given an image, the full task of SR is to predict the pair of action and its associated realized frame which is called a situation $S = \{v, F_{(I,v)}\}$. Action prediction is considered as a separate classification task independent from role prediction in existing work [20, 16]. As our focus is on inter-dependent query answering, we only aim at predicting the realized frame $F_{(I,v)}$ when the action v is given as action classification is not inter-dependent with roles. Therefore we call our task of role prediction formally as Frame Recognition (FR) from here onwards.

3.2. Frame Recognition and Backbone Model

We formulate FR as a Visual Question Answering (VQA) problem; Given an image I and query q, we want to find the most relevant information from the image to answer q. We formulate queries for each semantic role of the frame as the joint embedding of current frame's verb name and semantic role name. The model needs to answer all of them to retrieve the final realized frame.

We adopt the Top-Down Attention (TDA) model proposed by Anderson et al. [1] as our backbone VQA mechanism due to its simplicity and effectiveness as well as its less dependency towards the structure of the query compared to other state-of-the-art VQA models such as BAN [12], which relies on multiple channel query representations. Hence TDA allows us to use VQA with simple single channel queries which is sufficient for FR.

Given a set of region features of an image and a query embedding, TDA calculates the relevancy score for each image region feature with respect to query embedding. Then all image region features are weighted according to relevancy scores and summed together and fused with the query embedding. This creates the feature representation of the answer to the current query which then be sent through the classifier to obtain the final answer label.

3.3. Top-Down Attention for Frame Recognition

Figure 2 visualizes how we utilize TDA model for semantic role prediction to obtain the final frame $F_{(I,v)}$. First we consider each semantic role in the current frame as a separate query to our TDA model (handling inter-dependent queries will be discussed next). In the model, first we obtain image region features $\mathbf{E}_I = \{\mathbf{e}_n\}_{n=1}^{N_e}$ by encoding the image $\mathbf{I}$ using a CNN and obtaining the grid features just after the last pooling layer. N_e is the number of regions of the image. We use word embeddings for semantic role r and verb v of the current frame to generate the query encoding $\mathbf{q}$.

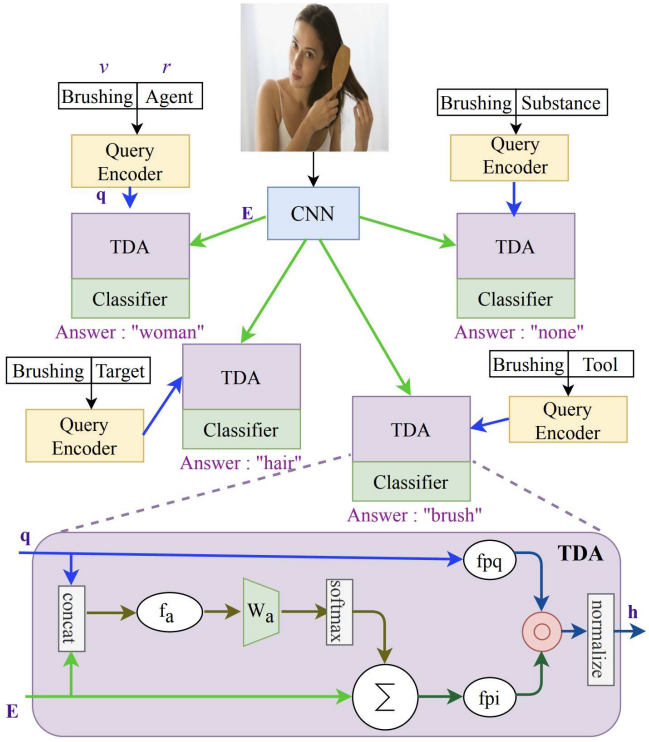

Figure 2. Top-Down Attention (TDA) model for Frame Recognition in SR. Each role of the verb "Brushing" forms a query, receives the image encoding and goes through the TDA network and the classifier as an independent query to obtain the final noun prediction. Nodes with the same colour indicates the same network which shares parameters.

$$\mathbf{E}_I = \text{CNN}(\mathbf{I}), \tag{1}$$

$$\mathbf{q} = f_q([\mathbf{w}_v, \mathbf{w}_r]), \tag{2}$$

where $\mathbf{E}_I \in \mathbb{R}^{N_e \times d\_img}$ and f_q is a non-linear layer. $[\cdot]$ is used to denote the concatenation. $\mathbf{q} \in \mathbb{R}^{d\_q}$ and embedding vectors for verb and role are $\mathbf{w}_v, \mathbf{w}_r \in \mathbb{R}^{d\_wemb}$. These embeddings are randomly initialized and learnt during model training. (Details of all networks (e.g., f_q) are provided in Supplementary).

Then we calculate the image region-level attention weights based on the query encoding, and derive updated image encoding,

$$s_n = \mathbf{w}_a f_a([\mathbf{e}_n, \mathbf{q}])^T, \tag{3}$$

$$\alpha_n = \frac{\exp(s_n)}{\sum_{i=1}^{N_e} \exp(s_i)}, \quad \tilde{\mathbf{E}} = \sum_{n=1}^{N_e} \alpha_n \mathbf{e}_n, \tag{4}$$

s_n denotes un-normalized region-level attention weights obtained for current query $\mathbf{q}$. α_n denotes the normalized attention weight for region n, and $\tilde{\mathbf{E}}$ is the aggregated im-

age encoding for the query. $\mathbf{w}_a \in \mathbb{R}^{d\_hidden}$ are model parameters and f_a is a non-linear layer.

Then updated image encoding $\tilde{\mathbf{E}}$ and query encoding $\mathbf{q}$ are fused together to obtain the un-normalized hidden representation $\mathbf{h_u} \in \mathbb{R}^{d\_hidden}$,

$$\mathbf{h_u} = f_{pq}(\mathbf{q}) \circ f_{pi}(\tilde{\mathbf{E}}), \tag{5}$$

where f_{pq} and f_{pi} non-linear layers are used to project query and image encoding to a different space and $\circ$ denotes element-wise multiplication.

Element-wise multiplication can cause model convergence to an unsatisfactory local minimum [31]. In order to avoid this Yu et al. [31] have used the power normalization ($z \leftarrow \text{sign}(z)|z|^{0.5}$) and ℓ_2 normalization ($z \leftarrow z/\|z\|$) layers. Following their approach, we also modified the original TDA model by adding a Dropout [26] layer and normalization after element-wise multiplication to produce the *normalised hidden representation* $\mathbf{h}$:

$$\mathbf{h} = \ell_2\text{Norm}(\text{PowerNorm}(\text{Dropout}(\mathbf{h_u}))), \tag{6}$$

Classifier Finally the normalized hidden representation is sent through a non linear network $f_{classifier}$ followed by a SoftMax function to obtain final probability distributions of each role label prediction.

$$p = \text{SoftMax}(f_{classifier}(\mathbf{h})), \tag{7}$$

Learning and Inference We use cross entropy loss to train the model as follows:

$$Loss = \sum_{j=1}^{F_I} \left(-\sum_{i=1}^{|N|} y_{(j,i)} \log(p_i) \right) \tag{8}$$

$y_{(j,i)} \in \{0, 1\}$ is the ground truth encoding from the j^{th} realized frame for the noun i, where we can have F_I realized frames for each image. Also note that $p_i \in p$. This Situation Recognition dataset *imSitu* [30] contains three realized frame annotations for each image.

For the complete frame prediction, first we obtain the required role list R_v for the given verb v to be queried in the model to retrieve noun label predictions $\hat{i} = \arg\max_i p_i^r$ for each role $r \in R_v$.

4. Handling Inter-dependent Semantic Roles

As we mentioned, the above system answers role queries independently. However, a semantic role not only depends on its action but also on its fellow semantic roles of the current frame, which we refer as its *neighbor roles*. For example in Figure 1, for the action "Brushing", *neighbor roles* for semantic role *Tool* are *Agent, Target* and *Substance*.

Existing query-based visual reasoning approaches [7, 1, 12] aim at answering questions individually. It has not been investigated how to incorporate information from inter-dependent queries to improve single query performance. Hence our backbone TDA model also suffers from this limitation. However, for structured prediction tasks like FR, modelling inter-dependency is important. Therefore, to address the gap between existing query-based visual reasoning approaches and inter-dependency models, we propose three different novel methods: (i) Context Aware Query (CAQ), (ii) Context Aware Image (CAI), and (iii) Context Aware Image Reconstruction (CAIR).

4.1. Context Aware Query (CAQ) for Inter-dependent Semantic Role Prediction

CAQ proposes to update the original query encoding with information from neighbour roles as a mechanism to incorporate structure to the existing TDA model. We call the aggregated information retrieved from *neighbour roles* as ***context***. Figure 3 depicts the system.

Context Generation We use hidden representations of all the roles of current verb v, $\mathbf{h}^r$, where $r = \{r_1, \ldots, r_{|R_v|}\}$ from TDA model, for the context generation. When generating context for role r, we calculate attention for all other roles in the current frame based on the hidden representation of r to decide how much each neighbour role is important to the current role. Then we weigh hidden representation of each neighbour role and aggregate all of them to generate the context for r.

$$d_k^b = \frac{\mathbf{h}\mathbf{W}_Q^b(\mathbf{h}^{r_k}\mathbf{W}_K^b)^T}{\sqrt{d\_hidden}}, \quad r_k \in R_v \setminus \{r\}, \tag{9}$$

$$\alpha_k^b = \frac{\exp(d_k^b)}{\sum_{i;r_i \neq r}^{|R_v|} \exp(d_i^b)}, \quad \mathbf{c}^b = \sum_{r_k \in R_v \setminus \{r\}} \alpha_k^b \mathbf{h}^{r_k}\mathbf{W}_V^b, \tag{10}$$

$$\mathbf{c} = [\mathbf{c}^1, \ldots, \mathbf{c}^b, \ldots, \mathbf{c}^B]\mathbf{W}_O, \quad b \in \{1 \ldots B\} \tag{11}$$

We use multi-head attention [27] for this to calculate the context in different representation sub-spaces and join them together to obtain the final context $\mathbf{c}$ (for the current role r). B is the number of heads. $\mathbf{W}_K^b \in \mathbb{R}^{d\_hidden \times d\_head}$, $\mathbf{W}_Q^b \in \mathbb{R}^{d\_hidden \times d\_head}$ and $\mathbf{W}_V^b \in \mathbb{R}^{d\_hidden \times d\_head}$ are model parameters to project hidden representations of key, query and value to a smaller B different subspaces. In our case, key and value are equal and they represent neighbour roles while query is the current role. $d\_head = d\_hidden/B$.

Context Aware Query Generation and Reasoning Now we incorporate the obtained context to query as follows and get the context aware query encoding $\mathbf{q}_c$.

$$\mathbf{q}_c = f_{cq}([\mathbf{c}, \mathbf{w}_v, \mathbf{w}_r]) \tag{12}$$

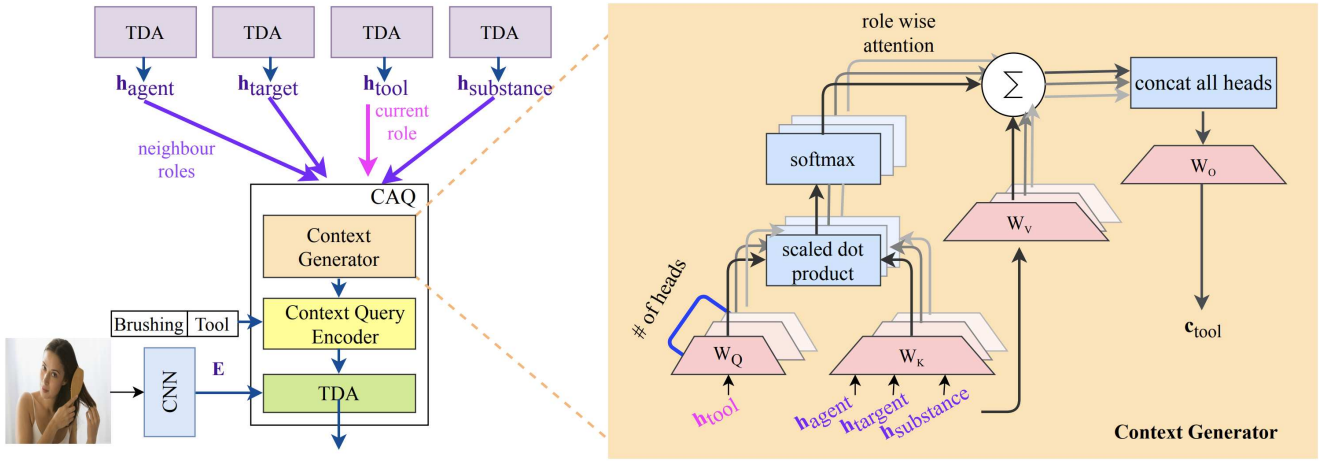

Figure 3. Context Aware Query (CAQ) based reasoning. In this example, the context is generated for the query of semantic role "tool", using its neighbour roles "agent", "target" and "substance", in the frame of verb "brushing". The context generator is discussed in Sec. 4.1. Diagram best viewed in colored version. Inputs to original TDA components (depicted in purple) are same as Figure 2.

Comparing with Equation 2, Equation 12 can be seen as adapting the query encoding using context $\mathbf{c}$ which is derived from hidden representations $\mathbf{h}^{r_k}$ of neighbor roles of current role r.

Then we input updated query encoding $\mathbf{q}_c$ and original image encoding $\mathbf{E}_I$ to Equation 3. Similar reasoning process to TDA is carried out until Equation 6 to obtain the new hidden representation $\mathbf{h}_c$. Finally $\mathbf{h}_c$ will be sent to the classifier for final prediction.

4.2. Context aware image (CAI)

In CAI, we add context $\mathbf{c}$ obtained in Equation 11 by adding it to the image instead of the question. This allows us to only extract information from image directly related to the context. This approach provides a way to highlight information now seems important at the presence of context prior to the reasoning. We use the following Equation 13 to incorporate context information generated in Equation 11 to image region encoding:

$$\mathbf{e}_n^c = \sigma([\mathbf{c}, \mathbf{e}_n]\mathbf{W}_{ic}) \circ \mathbf{e}_n, \quad \mathbf{e}_n \in \mathbf{E}_I \tag{13}$$

First, we concatenate the context with all the $\mathbf{e}_n$, $n \in N_e$ regions of the original image $\mathbf{E}_I$ and do a linear transformation using $\mathbf{W}_{ic} \in \mathbb{R}^{(d\_hidden+d\_img)\times d\_img}$. Finally, this is passed through a *sigmoid* gate to determine how much information of each region needs to be sent for the reasoning step based on the context. Once we obtained the updated image regions, we input it to Equation 3 instead of original image regions along with original query encoding $\mathbf{q}$ and continue the TDA mechanism.

4.3. Context Aware Image Re-construction (CAIR)

CAIR aims at improving inter-role agreement in the frame by encouraging the model to reconstruct the original image using hidden state $\mathbf{h}$ of all roles. If at least one of the role label representations is incorrect, the image reconstructed by the predicted realized frame differs from the original image. Therefore to construct an image similar to the original, the entire frame needs to be accurate. We use a non-linear layer f_{recon} to generate the reconstructed image from hidden representations $\mathbf{h}$ output from Equation 6 for all the roles of the current frame and send the original grid features $\mathbf{E}_I$ of the image through a linear network $f_{flatten\_img}$ to obtain the vector representation of the original image.

$$\hat{\mathbf{E}} = f_{recon}([\mathbf{h}_1, \ldots, \mathbf{h}_{|R_v|}]) \tag{14}$$

$$\mathbf{E}_{org} = f_{flatten\_img}(\mathbf{E}_I) \tag{15}$$

We add an auxiliary ℓ_2 loss to the original cross entropy loss in Equation 8 to encourage the model to make role label predictions which the combined frame prediction can reconstruct the original image as correctly as possible.

$$Loss_{recon} = \left\| \mathbf{E}_{org} - \hat{\mathbf{E}} \right\|_2 \tag{16}$$

When using this approach, the final loss for training the model is as following. β is a hyperparameter.

$$L = Loss + \beta Loss_{recon} \tag{17}$$

Verb Model	Top 1 Verb	Top 5 Verb
VGG Classifier [20, 16]	36.83	63.48
Predicted Query Model	35.70	62.19
RE-VGG Classifier	**37.96**	**64.99**

Table 1. Verb only prediction performance in accuracy %. For model using *gold queries*, Top-1: 43.21, Top-5: 68.83.

FR Model	Value	Value-all
TDA	72.96	37.60
CAQ	**73.62**	**38.71**
CAI	73.17	37.95
CAIR	73.30	38.17

Table 2. Frame recognition only performance in accuracy % of proposed context aware methods.

5. Evaluation

5.1. Dataset and Implementation Details

We use *imSitu* [30] dataset for our experiments and we follow the experiment setup and evaluation criteria from Yatskar et al. [30]. Here we report results for three metrics. *Verb*: verb prediction, *Value*: role-label tuple is considered correct given the verb, if it matches any of the F_I annotations, *Value-all*: when the entire frame is correct, meaning all role-value tuples of the predicted frame matches at least one ground truth annotation. Accuracy % of each of the three metrics is used to compare performance. *imSitu* dataset contains 75K train, 25K development and 25K test set samples which spreads across $V = 504$ verbs, $R = 190$ roles and $N = 2001$ nouns including *UNK* token for unknowns. Each image has $F_I = 3$ realized frames.

We implemented our models using PyTorch [21] framework. We use VGG-16 [24] as our backbone CNN architecture to encode images following all existing work [30, 29, 20, 16] for SR. We extract grid features of size $7 \times 7 \times 512$ after the final max pooling layer as our image regions where $N_e = 49$. Complete details of the entire implementation and all network architectures are provided in the supplementary materials.

5.2. Reasoning Enhanced Verb Prediction

In this section, we discuss experiments for verb prediction only. Main experiments on FR using our proposed context-aware reasoning will be discussed in the next section.

We analyse the performance of verb prediction when visual reasoning is expanded beyond CNN. Table 1 shows performance of multiple approaches we followed. First we report results for the CNN [15] verb classifier, the model which was used by many of the existing work [30, 20, 16] as the baseline. For reasoning enhanced predictions, we use the same TDA architecture explained in Section 3.3 and use *Agent* and *Place* role labels as the query in Equation 2 to reason the image for verb. We use ground truth *Agent* and *Place* label annotations to form *gold queries* in our reference gold query model. In the Predicted Query Model model, *predicted queries* are formulated using *Agent* and *Place* label predictions from our TDA based pre-trained FR model. Due to FR model's prediction errors, we observe a considerable performance drop in results. Finally, we have our Reasoning Enhanced verb prediction model (RE-VGG) in which we incorporate visual reasoning capabilities of the predicted-role based TDA verb model to the VGG classifier by summing verb wise scores output from the last FC layer of both models to obtain our best verb model.

5.3. Context Aware Reasoning for Frame Recognition

In this section we discuss results for the main contribution of this work on how well the context incorporation helps to improve Frame Recognition and results are shown in Table 2.

Our TDA model answers queries independently without considering its neighbour roles of the current frame. Next we have performance of our three proposed models for handling inter-dependent queries. CAQ has outperformed both CAI and CAIR becoming the best approach for inter-dependent query answering. The reason is that it only uses context information as a guidance for the reasoning and if the model feels original image's features are more important to answer query than the context, CAQ allows that too. But in CAI, as the original image is altered using the context, it does not have the opportunity to use original image information at all. CAIR only distantly encourages for role inter-dependency and does not explicitly force like CAQ, hence it cannot perform as good as CAQ.

5.4. Comparison with Existing Work

Table 3 and Table 4 show the performance comparison of our models against existing work. The results of different methods are obtained by either running the authors' provided implementation if they are available, or taking from their papers if the implementations are not available. However, for GGNN based model [16], the authors' provided implementation could not converge. After communicating with the authors, we have re-implemented the model ourselves, and our results are similar to the reported ones by the authors except for "value-all", which we observed lower accuracy than what was reported in [16].

We report results for both TDA model and our best inter-dependent query handling CAQ model. Our TDA model which handles role predictions independently has already outperformed all existing work including models which explicitly model role dependencies [20, 16]. This not only proves the effectiveness of sophisticated multi-modal reasoning but also shows how visual reasoning tasks other than VQA can benefit from adopting to query-based reasoning methods. We further improve our performance with CAQ and achieves the new state-of-the-art results for FR. We report verb prediction results for both our CNN based verb classifier *VGG Verb* as well as our reasoning enhanced *RE-*

	top-1 predicted verb			top-5 predicted verbs			ground truth verbs		mean
	verb	value	value-all	verb	value	value-all	value	value-all	
CNN + CRF [30]	32.25	24.56	14.28	58.64	42.68	22.75	65.90	29.50	36.32
Tensor Composition [29]	32.91	25.39	14.87	59.92	44.50	24.04	69.39	33.17	38.02
Above + DataAug [29]	34.2	26.56	15.61	62.21	46.72	25.66	70.80	34.82	39.57
RNN [20]	36.11	27.74	16.60	63.11	47.09	26.48	70.48	35.56	40.40
VGG Verb, GGNN † [16]	36.83	28.31	16.55	63.48	47.27	25.77	69.63	33.58	40.18
VGG Verb, TDA (Ours)	36.83	29.01	17.52	63.48	48.82	27.91	72.96	37.60	41.77
VGG Verb, CAQ (Ours)	36.83	29.24	18.02	63.48	49.22	28.62	**73.62**	**38.71**	42.22
RE-VGG, CAQ (Ours)	**37.96**	**30.15**	**18.58**	**64.99**	**50.30**	**29.17**	**73.62**	**38.71**	**42.94**

Table 3. Situation prediction results on *imSitu* development set. † denotes results of our implementation. Best performance in each column is highlighted in **bold** and second best is underlined.

	top-1 predicted verb			top-5 predicted verbs			ground truth verbs		mean
	verb	value	value-all	verb	value	value-all	value	value-all	
CNN + CRF [30]	32.34	24.64	14.19	58.88	42.76	22.55	65.66	28.96	36.25
Tensor Composition [29]	32.96	25.32	14.57	60.12	44.64	24.00	69.20	32.97	37.97
Above + DataAug [29]	34.12	26.45	15.51	62.59	46.88	25.46	70.44	34.38	39.48
RNN [20]	35.90	27.45	16.36	63.08	46.88	26.06	70.27	35.25	40.16
VGG Verb, GGNN † [16]	36.97	28.21	16.27	63.62	47.16	25.32	69.34	33.29	40.02
VGG Verb, TDA (Ours)	36.97	29.04	17.56	63.62	48.81	27.80	72.80	37.46	41.75
VGG Verb, CAQ (Ours)	36.97	29.29	17.98	63.62	49.22	28.45	**73.41**	**38.52**	42.18
RE-VGG, CAQ (Ours)	**38.19**	**30.23**	**18.47**	**65.05**	**50.21**	**28.93**	**73.41**	**38.52**	**42.88**

Table 4. Situation prediction results on *imSitu* test set. Best performance in each column is highlighted in **bold** and second best is underlined.

VGG models and we achieve new state-of-the-art results for verb prediction as well.

5.5. Qualitative Analysis

Figure 4 shows two sample predictions from the *imSitu* development set for verbs "Assembling" and "Igniting" with predicted attention heat maps output from Equation 4 for all roles in both TDA and CAQ models. Role dependency matrices were generated by combining unnormalized neighbour role weights generated for all roles from Equation 9. For verb "Assembling", TDA model has predicted role *Tool* incorrectly. When CAQ model generates the context for role *Tool*, roles *Component* and *Goal Item* provide the most impact according to the second row of the matrix. We can see the correct predictions of those roles have guided *Tool* in the CAQ model to correct its prediction by adjusting the attention directly to the "Drill". In the second sample also the correct prediction of role *Item* (most important neighbour for *Tool* in verb "Igniting") has guided to correct the attention error of *Tool* happened in TDA via the context information in CAQ. These results show both the effectiveness of our model as well as its interpretability.

5.6. Ablation Study

We discuss our analysis on combining proposed context incorporation approaches in this section and results are reported in Table 5. Even-though TDA was able to benefit from CAIR according to Table 2, CAQ and CAI were unable to achieve improvement from combining with CAIR. This is because the generated context in these models already implicitly facilitates inter-role agreement in order to maintain the stability of predictions across the frame. Hence CAIR is just an ineffective repetition. Performance has degraded when CAQ combined with CAI. The reason for this is that, when both image and query are incorporated with context, there is no room left for individual reasoning to incorporate important information from the original image which might be particularly important for the current role. This result shows an important message on how important it is to allow models some space for independent reasoning as

CAQ	CAI	CAIR	Value	Value-all
Proposed approach	-	-	73.62	38.71
✓	-	✓	73.62	38.63
-	✓	✓	73.17	37.99
✓	✓	-	72.94	37.38
✓	✓	✓	73.41	38.21

Table 5. Model performance after combining Context Incorporation Methods. First row contains our final proposed CAQ only model as the reference.

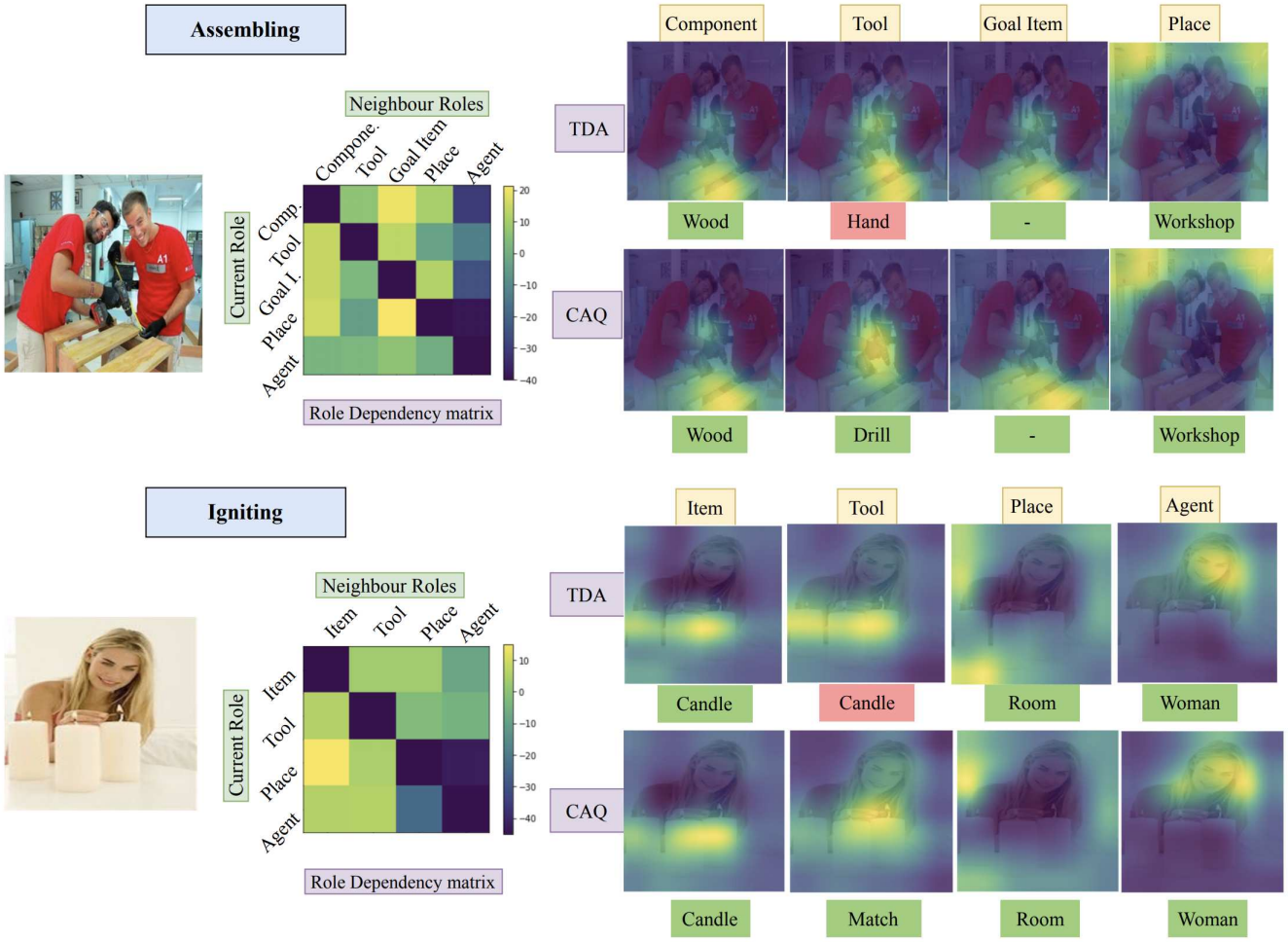

Figure 4. Visualization of attention maps for multi-modal reasoning and role dependency matrices for two verbs. In both attention maps and matrices, lighter the colour represents higher the value. Diagonal elements of the matrix are indicated in the darkest color to show that own value of current role is not considered as a neighbour role in context generation. Predicted nouns for each role is indicted after each attention map and coloured in green if its correct, red otherwise. **Note the improved attention in "Tool" prediction using context from neighbor roles.** We have removed attention maps for the least important *Agent* role of verb "Assembling" due to the space limitation. Best viewed in colored version.

well without completely relying on role inter-dependency, which can cause bias for object co-occurrences in training set. However, this particular issue has been solved for a certain extend after adding CAIR to this model. This is because the $Loss_{recon}$ in Equation 16 pushes all predicted objects in the frame to generate an image representation closer to the original image regulating the model from biasing to training set object co-occurrences.

6. Conclusion

We address the task of Situation Recognition as a query-based visual reasoning problem. We further extend our work by proposing novel mechanisms to enable query-based visual reasoning models to handle inter-dependent queries which is a unique requirement of Situation Recognition. For the best of our knowledge, this is the first attempt in incorporating inter-dependent query handling capabilities to query-based visual reasoning models. Our methods achieve new state-of-the-art results for Situation Recognition.

Acknowledgement This work was supported by ST Electronics and the National Research Foundation(NRF), Prime Minister's Office, Singapore under Corporate Laboratory @ University Scheme (Programme Title: STEE Infosec - SUTD Corporate Laboratory), National Research Foundation Singapore under its AI Singapore Programme [Award Number: AISG-100E2018-005] and partially supported by the Energy Market Authority (EP award no. NRF2017EWT-EP003-061).

References

[1] 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 *CVPR*, volume 3, page 6, 2018. 2, 3, 4

[2] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pages 86–90, 1998. 2

[3] Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. MUREL: multimodal relational reasoning for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1989–1998, 2019. 2

[4] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[5] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1–10, 2018. 2

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4

[8] Yiluan Guo and Ngai-Man Cheung. Efficient and deep person re-identification using multi-level similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[9] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[10] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997, 2017. 2

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015. 1

[12] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 1571–1581, 2018. 2, 3, 4

[13] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012. 1

[15] Yann Lecun and Yoshua Bengio. *Convolutional networks for images, speech, and time-series*. MIT Press, 1995. 2, 6

[16] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4183–4192, 2017. 2, 3, 5, 6, 7

[17] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1261–1270, 2017. 1

[18] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015. 2

[19] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 414–428, 2016. 2

[20] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 455–463, 2017. 2, 3, 5, 6, 7

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 1

[23] Gaurav Sharma, Frédéric Jurie, and Cordelia Schmid. Expanded parts model for human attribute and action recognition in still images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 652–659, 2013. 1

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 6

[25] Sibo Song, Ngai-Man Cheung, V Chandrasekhar, and B Mandal. Deep adaptive temporal pooling for activity recognition. In *ACM Multimedia*, 2018. 1

[26] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. 4

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017. 4

[28] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. *arXiv preprint arXiv:1701.02426*, 2017. 1

[29] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6, 7

[30] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 4, 6, 7

[31] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018. 4

[32] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[33] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. *CoRR*, abs/1811.07662, 2018. 2