ClipSitu: Effectively Leveraging CLIP for Conditional Predictions in Situation Recognition

Debaditya Roy*,1, Dhruv Verma*,2, and Basura Fernando1,2

[1]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

Abstract

*Situation Recognition is the task of generating a structured summary of what is happening in an image using an activity verb and the semantic roles played by actors and objects. In this task, the same activity verb can describe a diverse set of situations as well as the same actor or object category can play a diverse set of semantic roles depending on the situation depicted in the image. Hence a situation recognition model needs to understand the context of the image and the visual-linguistic meaning of semantic roles. Therefore, we leverage the CLIP foundational model that has learned the context of images via language descriptions. We show that deeper-and-wider multi-layer perceptron (MLP) blocks obtain noteworthy results for the situation recognition task by using CLIP image and text embedding features and it even outperforms the state-of-the-art CoFormer, a Transformer-based model, thanks to the external implicit visual-linguistic knowledge encapsulated by CLIP and the expressive power of modern MLP block designs. Motivated by this, we design a cross-attention-based Transformer using CLIP visual tokens that model the relation between textual roles and visual entities. Our cross-attention-based Transformer known as ClipSitu XTF outperforms existing state-of-the-art by a large margin of 14.1% on semantic role labelling (value) for top-1 accuracy using imSitu dataset. Similarly, our ClipSitu XTF obtains state-of-the-art situation localization performance. We will make the code publicly available**.

1. Introduction

Situation recognition was first introduced to computer vision in pioneering work [34]. Situation recognition is an important problem in scene understanding, activity understanding, and action reasoning as it provides a structured representation of the main activity depicted in the image. The key component in situation recognition is the task of semantic role labeling. Semantic role labeling is complex as the same activity verb may have different functional meanings and purposes depending on the context of the image. For example, the verb “spray” can be used to describe a firefighter spraying water on a fire, someone spraying oil on salad, someone spraying perfume on their face, and someone spraying hairspray on their hair. Hence, semantic role labeling requires a detailed understanding of the event in the image using contextual information from the image and how it relates to the linguistic definition of the event in terms of the activity (verb) and activity-specific roles.

Multimodal Foundation Models such as CLIP [23] and ALIGN [11] provide context as they are trained on many millions of image/text pairs to capture cross-modal dependencies between images and text. In these millions of examples, CLIP model might encounter different usages of the same verb that describe visually different but semantically similar images. Hence, CLIP is an excellent multimodal foundation model for solving image semantic role labeling tasks as it provides a grounded understanding of visual and linguistic information. In [7], CLIP is shown to be trainable for complex vision and language tasks termed Structured Vision and Language Concepts. Another way to leverage multimodal foundation models is to apply an MLP on top of the image encoder in works such as VL-Adapter [26], AIM [32], EVL [19] and wise-ft [29]. These approaches can be applied for predicting the main activity in the image i.e. for image classification [29] or action detection [32]. However, semantic role labeling is a conditional classification task that needs verb and role along with the image. Therefore, in [16], authors convert situation recognition to a text-prompt-based prediction problem by finetuning a CLIP image encoder with the text outputs from a large language model – GPT-3 [2] called CLIP-Event. The

[*] These authors contributed equally to this work
[*] https://github.com/LUNAProject22/CLIPSitu

verbs are ranked using the prompt "An image of ⟨verb⟩" based on image CLIP embeddings. After predicting the verb, each noun is predicted using another text prompt "The ⟨name⟩ is a ⟨role⟩ of ⟨verb⟩", i.e. "The firefighter is an agent in spraying". Even with the world knowledge in GPT-3, CLIP-Event performs worse on semantic role labeling than state-of-the-art CoFormer [3] which is directly trained on the images. The reason is that finetuning CLIP on semantic role labeling is not effective as the dataset imSitu [34] is not massive containing only 126,102 images yet it contains a massive amount of nouns (11,538) that are related to 190 unique roles. Therefore, the mapping between roles and nouns becomes an extremely challenging task.

We show that a well-designed multimodal MLP that consists of a modern MLP block design is able to solve semantic role labeling using CLIP embeddings and it outperforms the state-of-the-art without finetuning the CLIP model. This multimodal MLP is trained on a combination of image and text embedding from the verb and the role obtained from the CLIP model. Multimodal MLP predicts the entity corresponding to the role using a simple loss function. Motivated by the effectiveness of CLIP-based multimodal MLP, we adopt a Transformer encoder to leverage the connection across semantic roles in an image. Each semantic role is represented using a multimodal input of image and text embedding of the verb and the role. We show that sharing information across semantic roles using a Transformer leads to slightly improved performance. Through the multimodal MLP and Transformer we find that CLIP-based image, verb and role embeddings are effective for role prediction and predicting all roles for a verb by sharing information across them further improves the efficacy. Motivated by these two findings, we design a cross-attention Transformer to learn the relation between semantic role queries and CLIP-based visual token representations of the image to further enhance the connection between visual and textual entities. We term this model as ClipSitu XTF and it obtains state-of-the-art results for Situation Recognition on imSitu dataset outperforming state-of-the-art CoFormer [3] by 14.1% on top-1 value performance. Similarly, we leverage the cross-attention scores to localize the role in the image. Using ClipSitu XTF, we obtain state-of-the-art results for situation localization.

2. Related Work

Situation Recognition. To understand the relationship between different entities in an image, tasks such as image captioning [13, 15, 10], scene graph generation [31, 5], and human-object interaction detection [9, 18] have been proposed in the literature. In situation recognition [34], the situational verbs and their roles are obtained based on the meaning of the activity in each image from FrameNet [8]. The entities for each role are populated using the large object dataset ImageNet. Recently, situation recognition has also been extended to videos with the VidSitu dataset [24] where each video spans multiple events each of which is described using a situational verb, semantic roles, and their nouns. The VidSitu dataset is extended with grounded entities in [14] while [30] proposes a contrastive learning objective framework for video semantic role labeling. We limit the scope of this work to situation recognition in images.

One-stage prediction approaches predict the situational verb from the image and then the nouns associated with the roles of those verbs. In [34], a conditional random field model is proposed that decomposes the task of situation recognition into verb prediction and semantic role labeling (SRL). For SRL, they optimize the log-likelihood of the ground-truth nouns corresponding to each role for an image over possible semantic role-noun pairs from the entire dataset. In [33], a tensor decomposition model is used on top of CRF that scores combinations of role-noun pairs. They also perform semantic augmentation to provide extra training samples for rarely observed noun-role combinations. In [20], a predefined order for semantic roles is decided to predict the nouns for an image, and a recurrent neural network is used to predict the nouns in that order. Authors in [17] propose a gated graph neural network (GGNN) to capture all possible relations between roles instead of a predefined order as in [20]. In [25], a mixture kernel is applied to relate the nouns predicted for one role with respect to the noun predicted for another role. These relations provide a prior for the GGNN [17] to predict nouns.

In [22], imSitu is extended with grounded entities in each image to create Situations With Grounding (SWiG) dataset. They propose two models – Independent Situation Localizer (ISL) and Joint Situation Localizer (JSL). Both ISL and JSL use LSTMs to predict nouns in a predefined sequential order similar to [20] while RetinaNet estimates the locations of entities. A transformer encoder-decoder architecture is proposed in [4] where the encoder captures semantic features from the image for verb prediction and the decoder learns the role relations. In [12], situational verbs are predicted using a CLIP encoder on the image and the detected objects in the image.

Two-stage prediction approaches introduce an additional stage to enhance the verb prediction using the predicted nouns of the roles. In [6], transformers are used to predict semantic roles using interdependent queries that contain the context of all roles. The context acts as the key and values while the verb and the role form the query to predict the noun. They also consider the nouns of two predefined roles along with the image to enhance the verb prediction using a CNN. In [28], a coarse-to-fine refinement of verb prediction is proposed by re-ranking verbs based on the nouns predicted for the roles of the verb. CoFormer[3] combines ideas from [28] and [4] with transformer encoder and decoder predicting verbs and nouns, respectively. They add another encoder-decoder to refine the verb prediction based on the decoder outputs from the noun decoder.

3. CLIPSitu Models and Training

In this section, first, we present how we extract CLIP [23] embedding (features) for situation recognition. After that we present the verb prediction model. Then, we present three models for Situation Recognition using the CLIP embeddings. Afterward, we present a loss function that we use to train our models.

3.1. Extracting CLIP embedding

Every image I has a situational action associated with it, denoted by a verb v. For this verb v, there is a set of semantic roles $R_v = \{r_1, r_2, \cdots, r_m\}$ each of which is played by an entity denoted by its noun value $N = \{n_1, n_2, \cdots, n_m\}$. We use CLIP [23] visual encoder $\psi_v()$, and the text encoder $\psi_t()$ to obtain representations for the image, verb, roles, and nouns denoted by X_I, X_V, X_{R_v} and, X_N respectively. Here $X_{R_v} = \{X_{r_1}, X_{r_2}, \cdots, X_{r_m}\}$ for m roles and $X_N = \{X_{n_1}, X_{n_2}, \cdots, X_{n_m}\}$ for corresponding m nouns where $X_{r_i} = \psi_t(r_i)$ and $X_{n_i} = \psi_t(n_i)$ are obtained using text encoder. Similarly, the $X_I = \psi_v(I)$ and $X_V = \psi_t(v)$ is obtained using vision encoder. Note that all representations X_I, X_V, X_{r_i} and X_{n_i} have the same dimensions. In imSitu dataset[34], we have 504 unique verbs, 190 unique roles, and 11538 unique nouns. We extract CLIP text embeddings each verb, role, and noun separately.

3.2. ClipSitu Verb MLP

The first task in situation recognition is to predict the situational verb correctly from the image. We design a simple MLP with CLIP embeddings of the image X_I as input called ClipSitu Verb MLP as follows:

$$\hat{v} = \phi_V(X_I). \tag{1}$$

where ϕ_V contains l linear layers of a fixed dimension with ReLU activation to predict the situational verb. Just before the final classifier, there is a Dropout layer with a 0.5 rate. We train ClipSitu Verb MLP with standard cross-entropy loss.

3.3. ClipSitu MLP

Here we present a modern multimodal MLP block design for semantic role labeling for Situation Recognition that predicts each semantic role of a verb in an image. We term this method as **ClipSitu MLP**. Specifically, given the image, verb, and role embedding, the ClipSitu MLP predicts the embedding of the corresponding noun value for the role. In contrast to what has been done in the literature, ClipSitu MLP obtains contextual information by conditioning the information from the image, verb, and role embeddings. While the image embedding provides context about the possible nouns for the role, the verb provides the context on how to interpret the image situation.

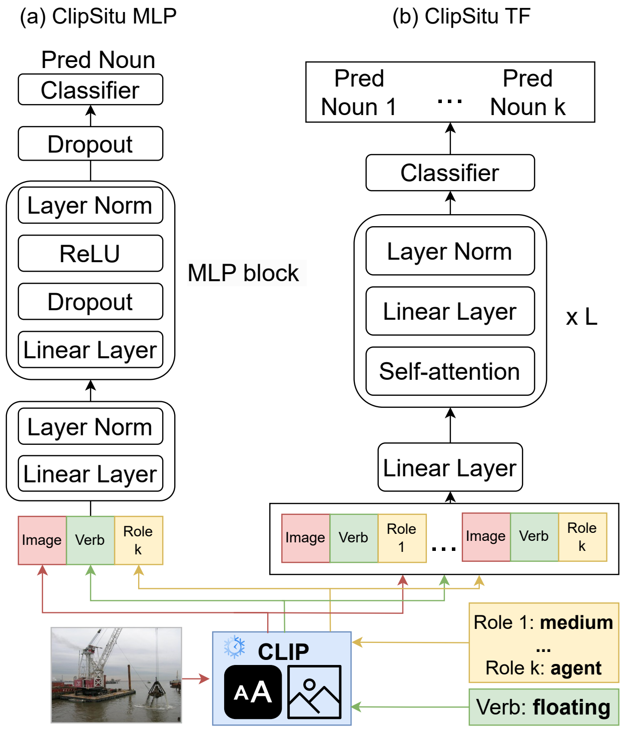


Figure 1. Architecture of the ClipSitu MLP and TF models. We use pooled image embedding from the CLIP image encoder for ClipSitu MLP and TF. In ClipSitu TF, all the roles for the verb are predicted simultaneously.

We concatenate the role embedding for each role r_i to the image and verb embedding to form the multimodal input X_i where $X_i = [X_I, X_v, X_{r_i}]$. Then, we stack l MLP blocks to construct CLIPSitu MLP and use it to transform the multimodal input X_i to predict the noun embedding $\hat{X}_{n_i}$ as follows:

$$\hat{X}_{n_i} = \phi_{MLP}(X_i). \tag{2}$$

In ϕ_{MLP}, the first MLP block projects the input feature X_i to a fixed hidden dimension using a linear projection layer followed by a LayerNorm [1]. Each subsequent MLP block consists of a Linear layer followed by a Dropout layer (with a dropout rate of 0.2), ReLU [21], and a LayerNorm as shown in Fig. 1(a). We predict the noun class from the predicted noun embedding using a dropout layer (rate 0.5) followed by a linear layer which we name as classifier ϕ_c as

$$\hat{y}_{n_i} = \texttt{argmax}\ \phi_c(\hat{X}_{n_i}) \tag{3}$$

where $\hat{y}_{n_i}$ is the predicted noun class. We use cross-entropy loss between predicted $\hat{y}_{n_i}$ and ground truth nouns y_{n_i} as

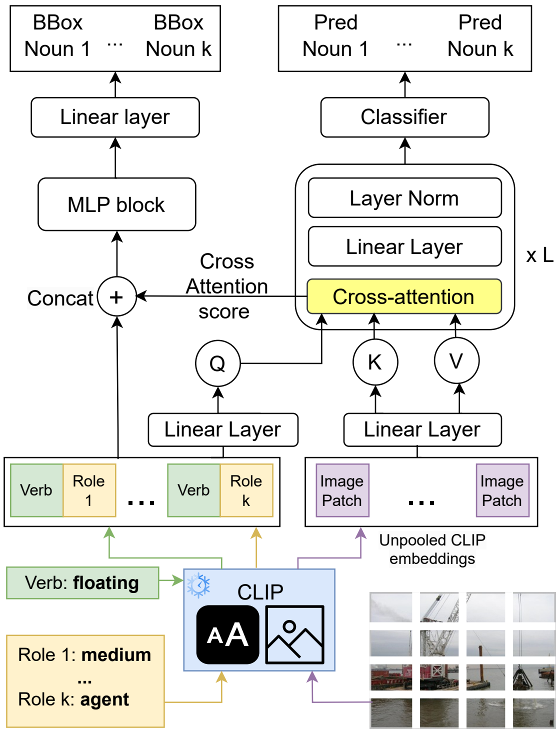


Figure 2. Architecture of ClipSitu XTF. We use embeddings from each patch of the image obtained from CLIP image encoder.

explained later in Section 3.6 to train the model.

3.4. ClipSitu TF: ClipSitu Transformer

The role-noun pairs associated with a verb in an image are related as they contribute to different aspects of the execution of the verb. Hence, we extend our ClipSitu MLP model using a Transformer [27] to exploit the interconnected semantic roles and predict them in parallel. The input to the Transformer is similar to ClipSitu MLP (i.e. $X_i = [X_I, X_v, X_{r_i}]$), however, we build a set of vectors using $\{X_1, X_2, \cdots, X_m\}$ where m denotes the number of roles of the verb. Each vector in the set is further processed by a linear projection to reduce dimensions. We initialize a Transformer model ϕ_{TF} with l encoder layers and multi-head attention with h heads. Using the Transformer model, we predict the value embedding of the m roles as output tokens of the transformer

$$\{\hat{X}_{n_1}, \hat{X}_{n_2}, \cdots, \hat{X}_{n_m}\} = \phi_{TF}(\{X_1, X_2, \cdots, X_m\}). \tag{4}$$

Similar to the MLP, we predict the noun classes using a classifier on the value embedding as $\hat{y}_i = \texttt{argmax}\ \phi_c(\hat{X}_{n_i})$ where $i = \{1, \cdots, m\}$ as shown in Fig. 1(b).

3.5. ClipSitu XTF: Cross-Attention Transformer

Each semantic role in a situation is played by an object located in a specific region of the image. Therefore, it is important to pay attention to the regions of the image which has a stronger relationship with the role. Such a mechanism would allow us to obtain better noun prediction accuracy. Hence, we propose to use the encoding for each patch of the image obtained from the CLIP model. We design a cross-attention Transformer called **ClipSitu XTF** to model how each patch of the image is related to every role of the verb through attention as shown in Fig. 2.

Let the patch embedding of an image be denoted by $X_{I,p} = \{X_I^1, X_I^2, \cdots, X_I^p\}$ where p is the number of image patches. These patch embeddings form the key and values of the cross-attention Transformer while the verb-role embedding is the query in Transformer. The verb embedding is concatenated with each role embedding to form m verb-role embeddings $X_{vr} = \{[X_V; X_{r_1}], [X_V; X_{r_2}], \cdots, [X_V; X_{r_m}]\}$. We project each verb-role embedding to the same dimension as the image patch embedding using a linear projection layer. Then the cross-attention operator in a Transformer block is denoted as follows:

$$Q = W_Q X_{vr}, K = V = W_I X_{I,p}$$
$$\hat{X} = softmax\frac{QK^T}{\sqrt{d_K}}V \tag{5}$$

where W_Q and W_I represent projection weights for queries, keys, and values and d_K is the dimension of the key token K. As with ClipSitu TF, we have l cross-attention layers in ClipSitu XTF. The predicted output from the final cross-attention layer contains m noun embeddings $\hat{X} = \{\hat{X}_{n_1}, \hat{X}_{n_2}, \cdots, \hat{X}_{n_m}\}$. Similar to the transformer in Section 3.4, we predict the noun classes using a classifier on the noun embeddings as $\hat{y}_i = \phi_c(\hat{X}_{n_i})$ where $i = \{1, \cdots, m\}$.

Next, we use ClipSitu XTF to perform localization of roles that requires predicting a bounding box $\mathbf{b}_i$ for every role r_i in the image. The cross-attention scores from the first layer of ClipSitu XTF $A_{m\times p}$ are rearranged into m score vectors $\{A_1, \cdots, A_m\}$. Each score vector A_i is p-dimensional and shows how each patch in the image is related to the verb and role. To incorporate the verb and role context to the score, we concatenate each score vector to its corresponding verb-role embedding from X_{vr} to obtain input for localization $X_l = \{[X_V; X_{r_1}; A_1], [X_V; X_{r_2}; X_2], \cdots, [X_V; X_{r_m}; X_m]\}$. We pass X_l through a single MLP block (designed for ClipSitu MLP) followed by a linear layer and a sigmoid function to obtain the predicted bounding box $\hat{b}_i \in [0, 1]^4$ for every role r_i. The four elements in the predicted bounding box indicate the center coordinates, height and width relative to the input image size. Though ClipSitu XTF uses cross-attention as CoFormer [3], the verb role tokens in the query and the image tokens are obtained from CLIP and not learned. Leveraging the power of CLIP embeddings allows us to design a simpler one-stage

ClipSitu XTF compared to the two-stage CoFormer [3].

3.6. Losses

Minimum Annotator Cross Entropy Loss. The imSitu dataset employs three annotators to label each noun for a role. In some instances, annotators may not provide the same annotation. Existing approaches [6, 3] make multiple predictions instead of one to tackle this issue. However, this can confuse the network during training as for there are multiple annotations for the same example. Furthermore, the loss function should not penalize a prediction that is close to any of the annotators' ground truth but further away from others. We propose minimum cross-entropy loss that considers the ground truth from each annotator. For a prediction $\hat{y}_i$, ground truth from all the annotators $\mathcal{A} = \{A_1, \cdots A_q\}$ is used to train our network as follows

$$\mathcal{L}_{MAXE} = \min_{\mathcal{A}} - \sum_{c=1}^{C} y_{i,c}^{(A_j)} log(\hat{y}_{i,c}) \text{ where } \forall A_j \in \mathcal{A}. \tag{6}$$

Here, C denotes the total number of classes and $\mathcal{L}_{MAXE}$ stands for Minimum Annotator Cross Entropy Loss. To train ClipSitu XTF for localization of roles, we employ $L1$ loss to compare the predicted and ground-truth bounding boxes

$$\mathcal{L}_{L1} = \frac{1}{m} \sum_{i=1}^{m} \|b_i - \hat{b}_i\|_1. \tag{7}$$

We train ClipSitu XTF for noun prediction and localization using the combined loss $\mathcal{L} = \mathcal{L}_{MAXE} + \mathcal{L}_{L1}$.

4. Experiments

4.1. Evaluation Details

We perform our experiments on imSitu dataset [34] and the augmented imSitu dataset called SWiG [22] for situation recognition and localization, respectively termed as grounded situation recognition. The dataset has a total of 125k images with 75k train, 25k validation, and 25k test images. The metrics used for semantic role labeling are *value* and *value-all* [34] which predict the accuracy of noun prediction for a given role. For a given verb with k roles, *value* measures whether the predicted noun for at least one of k roles is correct. On the other hand, *value-all* measures whether all the predicted nouns for all k roles are correct. A prediction is correct if it matches the annotation of any one of the three annotators. Situation localization metrics *grnd value* and *grnd value-all* compute the accuracy of bounding box prediction [22] similar to value and value-all. A predicted bounding box is correct if the overlap with the ground truth is $\geq$ 0.5. The metrics value, value-all, grnd value and grnd value-all are evaluated in three settings based on whether we are using ground truth verb, top-1 predicted verb, or top-5 predicted verbs. For our model ablation on semantic role labeling and situation localization, we use the ground truth verb setting for measuring value, value-all, grnd value, and grnd value-all. All experiments are performed on the dev set unless otherwise specified.

4.2. Implementation Details

We use the CLIP model with ViT-B32 image encoder to extract image features unless otherwise specified. The input to ClipSitu MLP is a concatenation of the CLIP embeddings of the image, verb, and role, each of 512 dimensions leading to 1536 dimensions. For both ClipSitu TF and XTF, we set the sequence length to be 6 which refers to the maximum number of roles possible for a verb following [6]. Each verb has a varying number of roles and we mask the inputs that are not required. For ClipSitu TF, each input token in the sequence is the concatenated image, verb, and role CLIP embedding same as the MLP above which is projected to 512 dimensions using a linear layer. For the patch-based cross-attention Transformer (ClipSitu XTF), we obtain the embedding for input image patches from CLIP image encoder (ViT-B32 model) which results in 50 tokens ($224/32 \times 224/32$ + 1 class) of 512 dimensions that are used as key and value. The query tokens are concatenated verb and role CLIP embeddings that are projected to 512 dimensions using a linear layer. Unless otherwise mentioned, we train all our models with a batch size of 64, a learning rate of 0.001, and an ExponentialLR scheduler with Adamax optimizer, on a 24 GB Nvidia 3090.

4.3. Analysis with CLIP Image Encoders

In Table 1, we compare the proposed ClipSitu Verb MLP model against zero-shot and linear probe performance of CLIP. We also compare against a state-of-the-art CLIP finetuning model called weight-space ensembles (wise-ft) [29] that leverages both zero-shot and fine-tuned CLIP models to make verb predictions. We compare ClipSitu Verb MLP and wise-ft using 4 CLIP image encoders - ViT-B32, ViT-B16, ViT-L14, and ViT-L14@336px. The image clip embeddings for ViT-B32 and ViT-B16 are 512 dimensions and for ViT-L14, and ViT-L14@336px are 768 dimensions. These four encoders represent different image patch sizes, different depths of image transformers, and different input image sizes. The hidden layer is 1024 dimensional in the ClipSitu Verb MLP. Zero-shot performance of CLIP suggests that CLIP image features are beneficial for situation recognition tasks. Increasing the number of hidden layers does not improve performance for ClipSitu Verb MLP as it obtains the best top-1 and top-5 verb prediction even with a single hidden layer. ClipSitu Verb MLP performs better than wise-ft and linear probe for all image encoders which shows that MLP based finetuning on CLIP image features

Image Encoder	Verb Model	Hidden Layer	Top-1	Top-5
ViT-B32	zero-shot	-	29.20	65.21
	linear probe	-	44.63	78.35
	wise-ft	-	46.51	74.30
	ClipSitu Verb MLP	1	46.69	76.11
		2	46.51	76.08
		3	44.51	74.15
ViT-B16	zero-shot	-	31.90	67.89
	linear probe	-	49.27	78.76
	wise-ft	-	48.77	83.45
	ClipSitu Verb MLP	1	50.91	89.57
		2	50.83	89.40
		3	48.63	88.55
ViT-L14	zero-shot	-	38.18	79.34
	linear probe	-	52.39	87.67
	wise-ft	-	51.51	84.30
	ClipSitu Verb MLP	1	56.70	84.61
		2	56.63	84.49
		3	53.80	82.44
ViT-L14 @336px	zero-shot	-	39.70	79.21
	linear probe	-	53.40	81.45
	wise-ft	-	52.22	82.95
	ClipSitu Verb MLP	1	**57.86**	**86.11**
		2	56.22	84.55
		3	54.35	82.85

Table 1. Comparing performance of ClipSitu Verb MLP with zero-shot and finetuned CLIP (linear probe and wise-ft [29]).

Image Encoder	Model	Top-1		Top-5		Ground truth	
		value	v-all	value	v-all	value	v-all
ViT-B32	MLP	45.65	27.06	66.27	37.55	76.91	43.22
	TF	45.67	27.33	66.28	37.98	76.77	42.97
	XTF	44.54	25.94	64.93	35.56	75.25	40.79
ViT-B16	MLP	46.33	28.29	67.37	39.45	77.88	44.78
	TF	46.41	28.65	67.39	39.75	77.23	43.82
	XTF	45.67	27.44	66.09	37.42	75.43	40.58
ViT-L14	MLP	46.46	28.39	67.61	39.71	77.63	43.94
	TF	46.95	29.56	68.19	41.22	78.02	45.25
	XTF	46.95	29.49	68.08	40.61	77.84	44.54
ViT-L14 @336px	MLP	46.74	29.06	67.90	40.54	77.93	44.88
	TF	46.97	29.66	68.27	41.41	78.30	45.79
	XTF	**47.17**	**30.06**	**68.44**	**41.66**	**78.49**	**45.81**

Table 2. Comparison of CLIP Image Encoders on noun prediction task using top-1 and top-5 predicted verb from the best-performing Verb MLP model obtain from Table 1. All models' performance improves by increasing the number of patch tokens either by reducing patch size (32→16→14) or increasing image size (224→336). v-all stands for value all.

works better than finetuning the CLIP image encoder itself or using regression (linear probe) for situational verb prediction. Our best performing ClipSitu Verb MLP outperforms linear probe by 4.46% on Top-1 and wise-ft by 3.2% on Top-5 when using the same ViT-L14 image encoder.

Next, we study the effect of using different CLIP image encoders for noun prediction with ClipSitu MLP, TF and XTF. We compare ViT-B32, ViT-B16, ViT-L14, and ViT-L14@336px. For ClipSitu XTF, the number of image patch tokens used as key and value changes based on patch size and image size. We have 197 tokens ($224/16 \times 224/16$ + 1 class token) for ViT-B16, 257 tokens for ViT-L14 ($224/14 \times 224/14$ + 1 class token), and 577 tokens for ViT-L14@336px ($336/14 \times 336/14$ + 1 class token). For ViT-L14, and ViT-L14@336px image encoders, we obtain 768-dimensional embeddings which are projected using a linear layer to 512. We choose the best hyperparameters for ClipSitu MLP, TF, and XTF whose ablations are presented in Section 4.5.

In Table 2, we observe that the value and value-all using ground truth verbs steadily improve for all three models as the number of patches increases from 32 to 16 to 14 or the image size increases from 224 to 336. For ViT-B32 and ViT-B16, the best performance is obtained by ClipSitu MLP but it drops with ViT-L14. On the other hand, the maximum improvement is seen in ClipSitu XTF i.e. 5.1% for value-all compared to 1.6% and 2.8% for ClipSitu MLP and TF, respectively. ClipSitu XTF is able to extract more relevant information when attending to more image patch tokens to produce better predictions. To compare noun prediction using top-1 and top-5 predicted verbs, we use the best ClipSitu Verb MLP (ViT-L14@336px) from Table 1. For both Top-1 and Top-5 predicted verbs, we observe a similar trend as the ground truth verb. ClipSitu XTF again shows the most improvement in value and value-all to obtain the best performance among the three models across ground truth, Top-1 and Top-5 predicted verbs.

4.4. Comparison with SOTA

In Table 3, we compare the performance of proposed approaches with state-of-the-art approaches on situation recognition. We use ViT-L14@336px image encoder for all models – ClipSitu Verb MLP, ClipSitu MLP, ClipSitu TF, and ClipSitu XTF. ClipSitu Verb MLP outperforms SOTA method CoFormer on Top-1 and Top-5 verb prediction by a large margin of 12.6% and 12.4%, respectively, on the test set, which shows the effectiveness of using CLIP image embeddings over directly predicting the verb from the images. The comparison with existing works shows that with a well-designed MLP network, ClipSitu MLP outperforms state-of-the-art CoFormer [3] in all metrics comprehensively. ClipSitu MLP, TF, and XTF also handily outperform the only other CLIP-based semantic role labeling method, CLIP-Event [16]. ClipSitu XTF performs the best for noun prediction based on both the predicted top-1 verb and top-5 verb for value and value-all matrices. ClipSitu XTF outperforms state-of-the-art CoFormer by a massive margin of 14.1% on top-1 value and by 9.6% on top-1 value-all using the Top-1 predicted verb on the test set. Furthermore, on situation localization, ClipSitu XTF performs significantly better than state-of-the-art for top-1 grnd value by 11% while showing improvements on all metrics.

4.5. Ablations on hyperparameters

In Fig. 3, we explore combinations of MLP blocks and the hidden dimensions of each block to obtain the best MLP network for semantic role labeling. Increasing the number of MLP blocks and hidden dimensions steadily improves performance as the number of unique nouns to be predicted

Set	Method	Top-1 predicted verb					Top-5 predicted verb					Ground truth verb			
		verb	value	value-all	grnd value	grnd value-all	verb	value	value-all	grnd value	grnd value-all	value	value-all	grnd value	grnd value-all
dev	CRF [34]	32.25	24.56	14.28	-	-	58.64	42.68	22.75	-	-	65.90	29.50	-	-
	CRF w/ DataAug [33]	34.20	26.56	15.61	-	-	62.21	46.72	25.66	-	-	70.80	34.82	-	-
	RNN w/ Fusion [20]	36.11	27.74	16.60	-	-	63.11	47.09	26.48	-	-	70.48	35.56	-	-
	GraphNet [17]	36.93	27.52	19.15	-	-	61.80	45.23	29.98	-	-	68.89	41.07	-	-
	CAQ w/ RE-VGG [6]	37.96	30.15	18.58	-	-	64.99	50.30	29.17	-	-	73.62	38.71	-	-
	Kernel GraphNet [25]	43.21	35.18	19.46	-	-	68.55	56.32	30.56	-	-	73.14	41.68	-	-
	ISL [22]	38.83	30.47	18.23	22.47	07.64	65.74	50.29	28.59	36.90	11.66	72.77	37.49	52.92	15.00
	JSL [22]	39.60	31.18	18.85	25.03	10.16	67.71	52.06	29.73	41.25	15.07	73.53	38.32	57.50	19.29
	GSRTR [4]	41.06	32.52	19.63	26.04	10.44	69.46	53.69	30.66	42.61	15.98	74.27	39.24	58.33	20.19
	SituFormer [28]	44.32	35.35	22.10	29.17	13.33	71.01	55.85	33.38	45.78	19.77	76.08	42.15	**61.82**	24.65
	CoFormer [3]	44.41	35.87	22.47	29.37	12.94	72.98	57.58	34.09	46.70	19.06	76.17	42.11	61.15	23.09
	ClipSitu XTF	**58.19**	**47.23**	**29.73**	**41.30**	**13.92**	**85.69**	**68.42**	**41.42**	**49.23**	**23.45**	**78.52**	**45.31**	55.36	**32.37**
test	CRF [34]	32.34	24.64	14.19	-	-	58.88	42.76	22.55	-	-	65.66	28.96	-	-
	CRF w/ DataAug [33]	34.12	26.45	15.51	-	-	62.59	46.88	25.46	-	-	70.44	34.38	-	-
	RNN w/ Fusion [20]	35.90	27.45	16.36	-	-	63.08	46.88	26.06	-	-	70.27	35.25	-	-
	GraphNet [17]	36.72	27.52	19.25	-	-	61.90	45.39	29.96	-	-	69.16	41.36	-	-
	CAQ w/ RE-VGG [6]	38.19	30.23	18.47	-	-	65.05	50.21	28.93	-	-	73.41	38.52	-	-
	Kernel GraphNet [25]	43.27	35.41	19.38	-	-	68.72	55.62	30.29	-	-	72.92	42.35	-	-
	ISL [22]	39.36	30.09	18.62	22.73	07.72	65.51	50.16	28.47	36.6	11.56	72.42	37.10	52.19	14.58
	JSL [22]	39.94	31.44	18.87	24.86	09.66	67.60	51.88	29.39	40.6	14.72	73.21	37.82	56.57	18.45
	GSRTR [4]	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.5	15.88	74.11	39.00	57.45	19.67
	SituFormer [28]	44.20	35.24	21.86	29.22	13.41	71.21	55.75	33.27	46.00	20.10	75.85	42.13	**61.89**	24.89
	CoFormer [3]	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
	CLIP-Event [16]	45.60	33.10	20.10	21.60	10.60	-	-	-	-	-	-	-	-	-
	ClipSitu XTF	**58.19**	**47.23**	**29.73**	**40.01**	**15.03**	**85.69**	**68.42**	**41.42**	**49.78**	**25.22**	**78.52**	**45.31**	54.36	**33.20**

Table 3. Comparison with state-of-the-art on Grounded Situation Recognition. Robustness of ClipSitu MLP, TF, and XTF is demonstrated by the massive improvement for value and value-all with Top-1 and Top-5 predicted verbs over SOTA.

Heads		1				2				4				8			
Layers		1	2	4	6	1	2	4	6	1	2	4	6	1	2	4	6
ClipSitu TF	value	75.73	75.78	**76.87**	24.68	75.80	75.95	75.97	18.28	75.71	76.77	75.87	05.20	75.74	75.93	76.07	75.94
	value-all	41.40	41.52	41.84	00.21	41.64	41.60	41.83	00.21	41.43	42.10	41.75	00.00	41.35	41.58	**42.97**	41.72
ClipSitu XTF	value	72.70	74.33	**75.27**	74.35	53.11	53.17	53.11	53.16	53.18	53.51	53.45	53.49	53.13	53.44	53.38	53.54
	value-all	36.61	39.11	**40.79**	39.06	16.58	16.64	16.38	16.46	16.50	16.77	16.90	16.97	16.42	16.89	16.85	17.02

Table 4. Ablation on Transformer hyperparameters. 1 head with 4 layers is sufficient to obtain best value and value-all performance for ClipSitu XTF. For TF, 1 head and 4 layers produces best value whereas 8 heads and 4 layers produces best value-all performance.

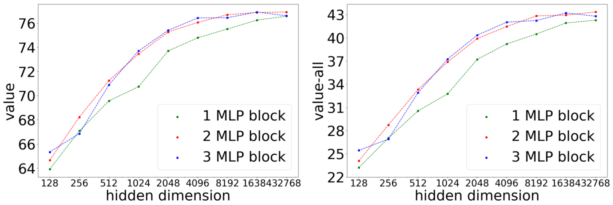


Figure 3. Effect of the number of MLP blocks and hidden dimensions on value and value-all. We train with very large hidden dimensions such as 8192, 16384, and 32768 to obtain state-of-the-art value and value-all results.

Model	Top-1		Top-5		Ground truth	
	grnd value	grnd v-all	grnd value	grnd v-all	grnd value	grnd v-all
XAtt. L1	39.30	10.54	46.46	19.70	53.87	30.22
XAtt. L4	33.30	09.34	42.55	17.53	51.32	31.32
XAtt. L1 + L4	36.56	09.88	44.23	11.43	54.56	34.71
XAtt L1 + verb-role emb	**41.30**	**13.92**	**49.23**	**23.45**	**55.36**	**32.37**

Table 5. Comparing different ClipSitu XTF inputs for situation localization. XAtt. – cross-attention scores. L1 – first XTF layer and L4 – last XTF layer of best performing model (1 head, 4 layers, ViT-L14@336px). Concatenating (+) verb role embedding (emb) improves the performance of cross-attention scores. We use the best Verb MLP from Table 1. v-all stands for value-all.

is 11538. We train MLP with small to very large hidden dimensions i.e. 128→16384 which results in a steady improvement in both value and value-all. No improvement in value and value-all is seen when we increase the layer dimension further to 32768 for 3 MLP blocks which demonstrates that we have reached saturation. Our best ClipSitu MLP for semantic role labeling obtains 76.91 for value and 43.22 for value-all with 3 MLP blocks with each block having 16,384 hidden dimensions which beats the state-of-the-art CoFormer [3]. The main reason our ClipSitu MLP performs so well on semantic role labeling is our modern MLP block design that contains large hidden dimensions along with LayerNorm which have not been explored in existing MLP-based CLIP finetuning approaches. We also compare the performance of ClipSitu MLP with the proposed minimum annotator cross-entropy loss ($\mathcal{L}_{MAXE}$) versus applying cross-entropy using the noun labels of each annotator separately. We find that $\mathcal{L}_{MAXE}$ produces better value and value-all performance (76.91 and 43.22) compared to cross-entropy (76.57 and 42.88).

In Table 4, we explore the number of heads and layers needed to obtain the best-performing hyperparameters for semantic role labeling using ClipSitu TF and XTF. We find that a single head with 4 transformer layers performs the best in terms of value for both ClipSitu TF and XTF while for value-all, an 8-head 4-layer ClipSitu TF performs the best and we use this for subsequent evaluation. For both ClipSitu TF and XTF, increasing the number of layers be-

Model	# Parameters	GFlops	Inference Time(ms)
CoFormer [3]	93.0M	1496.67	30.62
ClipSitu Verb MLP	1.3M	0.17	0.08
ClipSitu MLP	580.2M	443.18	32.33
ClipSitu TF	**20.2M**	**8.65**	**1.55**
ClipSitu XTF	45.3M	116.01	11.17

Table 6. Comparison of parameters, flop count and inference time for CoFormer [3], Verb MLP, ClipSitu MLP, TF and XTF models.

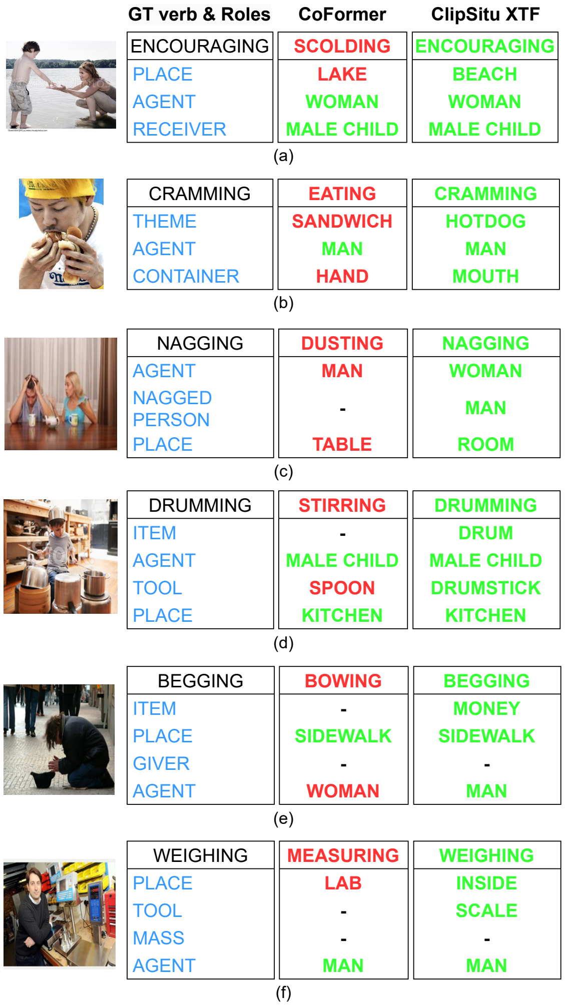
yond 4 does not yield any improvement in value or value-all when using less number of heads (1,2,4). Similarly, for ClipSitu XTF, increasing the number of heads and layers leads to progressively deteriorating performance. Both of these performance drops can be attributed to the fact that we have insufficient samples for training larger Transformer networks [23]. In ClipSitu XTF, we have fixed role tokens obtained from CLIP. We found this produces better noun prediction performance than learning the role tokens for each verb. Details are in Supplementary section 1. In Table 5, ablation on situation localization shows that cross-attention scores from the first XTF layer performs better than the last XTF layer. We concatenated verb and role embeddings to the cross-attention score of first XTF layer to provide more context about the role which further improves localization performance.

Complexity. We compare the number of parameters, computation, and inference time for ClipSitu MLP, TF, and XTF using the ViT-L14-336 image encoder and CoFormer [3] in Table 6. Please also see supplementary material section.

Qualitative Results. In Fig. 4, we compare the qualitative results of ClipSitu XTF with CoFormer. ClipSitu XTF is able to correctly predicts verbs such as cramming (Fig. 4(b)) while CoFormer focuses on the action of eating and hence incorrectly predicts the verb which also makes its noun predictions for the container and theme incorrect. CoFormer predicts the place as table and predicts the verb as dusting (Fig. 4(c)) instead of focusing on the action of nagging. Finally, we see in Fig. 4(d) that CoFormer is confused by the visual context of kitchen as it predicts stirring instead of identifying the action which is drumming. On the other hand, ClipSitu XTF correctly predicts drumming and the tool as drumsticks while still predicting place as kitchen. More qualitative results are in Supplementary section 2.

5. Conclusion

We propose to leverage CLIP embeddings for semantic role labeling. We show that multimodal ClipSitu MLP with large hidden dimensions outperforms the state-of-the-art semantic role labeling approach. We propose a ClipSitu XTF model that employs cross-attention between image patch embeddings from the CLIP image encoder and text embeddings. ClipSitu XTF sets the new state-of-the-art in semantic role labeling improving the current results by a large margin of 14.1% on top-1 value and by 9.6% on top-1 value-all. We also show that our approach of using CLIP embeddings is much more effective than finetuning CLIP, given the relatively small size of the dataset. Unlike, VL-Adapter [26], AIM [32], EVL [19] and wise-ft [29], our models can handle conditional inputs to solve Situation Recognition task. Despite the simplicity, our work shows that a traditional approach of freeze and finetune can be still relevant when used with modern neural network designs especially when using Foundational models.



GT verb & Roles	CoFormer	ClipSitu XTF
ENCOURAGING	SCOLDING	ENCOURAGING
PLACE	LAKE	BEACH
AGENT	WOMAN	WOMAN
RECEIVER	MALE CHILD	MALE CHILD

(a)

GT verb & Roles	CoFormer	ClipSitu XTF
CRAMMING	EATING	CRAMMING
THEME	SANDWICH	HOTDOG
AGENT	MAN	MAN
CONTAINER	HAND	MOUTH

(b)

GT verb & Roles	CoFormer	ClipSitu XTF
NAGGING	DUSTING	NAGGING
AGENT	MAN	WOMAN
NAGGED PERSON	-	MAN
PLACE	TABLE	ROOM

(c)

GT verb & Roles	CoFormer	ClipSitu XTF
DRUMMING	STIRRING	DRUMMING
ITEM	-	DRUM
AGENT	MALE CHILD	MALE CHILD
TOOL	SPOON	DRUMSTICK
PLACE	KITCHEN	KITCHEN

(d)

GT verb & Roles	CoFormer	ClipSitu XTF
BEGGING	BOWING	BEGGING
ITEM	-	MONEY
PLACE	SIDEWALK	SIDEWALK
GIVER	-	-
AGENT	WOMAN	MAN

(e)

GT verb & Roles	CoFormer	ClipSitu XTF
WEIGHING	MEASURING	WEIGHING
PLACE	LAB	INSIDE
TOOL	-	SCALE
MASS	-	-
AGENT	MAN	MAN

(f)

Figure 4. ClipSitu XTF vs. CoFormer [3] predictions. green refers to correct prediction while red refers to incorrect prediction. '-' refers to predicting blank (a noun class) for this role.

Acknowledgment This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001). This research is also supported by funding allocation to B.F. by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF), as well as its Centre for Frontier AI Research (CFAR).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19659–19668, 2022.

[4] Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. Grounded situation recognition with transformers. In *British Machine Vision Conference (BMVC)*, 2021.

[5] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021.

[6] Thilini Cooray, Ngai-Man Cheung, and Wei Lu. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4736–4745, 2020.

[7] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. 2023.

[8] Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. Background to framenet. *International journal of lexicography*, 16(3):235–250, 2003.

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018.

[10] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[12] Tianyu Jiang and Ellen Riloff. Exploiting commonsense knowledge about objects for visual activity recognition. 2023.

[13] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897, 2019.

[14] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded video situation recognition. In *Advances in Neural Information Processing Systems*.

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[16] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022.

[17] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4173–4182, 2017.

[18] JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*, 32:964–979, 2023.

[19] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 388–404. Springer, 2022.

[20] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463, 2017.

[21] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[22] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[24] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021.

[25] Mohammed Suhail and Leonid Sigal. Mixture-kernel graph attention network for situation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10363–10372, 2019.

[26] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2651–2658, 2022.

[29] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[30] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736, 2022.

[31] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[32] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*.

[33] 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*, pages 7196–7205, 2017.

[34] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016.