ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

EgoCap and EgoFormer: First-Person Image Captioning with Context Fusion

Zhuangzhuang Dai[a], Vu Tran[b], Andrew Markham[c], Niki Trigoni[c], M Arif Rahman[d], L. N. S. Wijayasingha[d], John Stankovic[d], Chen Li[e]

[a]*Dept. of Applied AI and Robotics, Aston University, United Kingdom*
[b]*Nordic Semiconductor, United Kingdom*
[c]*Dept. of Computer Science, University of Oxford, United Kingdom*
[d]*Dept. of Computer Science, University of Virginia, USA*
[e]*Dept. of Materials and Production, Aalborg University, Denmark*

Article history:
Received 23 Dec 2023

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Image Captioning, Storytelling, Dataset

ABSTRACT

First-person captioning is significant because it provides veracious descriptions of egocentric scenes in a unique perspective. Also, there is a need to caption the scene, a.k.a. life-logging, for patients, travellers, and emergency responders in an egocentric narrative. Ego-captioning is indeed non-trivial since (1) Ego-images can be noisy due to motion and angles; (2) Describing a scene in a first-person narrative involves drastically different semantics; (3) Empirical implications have to be made on top of visual appearance because the cameraperson is often outside the field of view. We note we humans make good sense out of casual footage thanks to our contextual awareness in judging when and where the event unfolds, and whom the cameraperson is interacting with. This inspires the infusion of such "contexts" for situation-aware captioning. We create *EgoCap* which contains 2.1K ego-images, over 10K ego-captions, and 6.3K contextual labels, to close the gap of lacking ego-captioning datasets. We propose *EgoFormer*, a dual-encoder transformer-based network which fuses both contextual and visual features. The context encoder is pre-trained on ImageNet before fine tuning with context classification tasks. Similar to visual attention, we exploit stacked multi-head attention layers in the captioning decoder to reinforce attention to the context features. The *EgoFormer* has realized state-of-the-art performance on *EgoCap* achieving a CIDEr score of 125.52. The *EgoCap* dataset and *EgoFormer* are publicly available at `https://github.com/zdai257/EgoCap-EgoFormer`.

1. Introduction

Vision data collected by body-worn cameras has seen a dramatic surge in the past decade. These data contain valuable information about the cameraperson's status as well as the surroundings. Nevertheless, these data are yet too unstructured, and sometimes irrelevant, to extract salient objects or activities, upon which object-led scene understanding has accomplished tremendous success, Anderson et al. (2018). Egocentric vision data are typically contaminated by motion blurring, hand occlusion, and awkward camera angles, Grauman et al. (2021). Interestingly, we humans tell a good story from footage captured in poor camera angles. From a pair of shoes pointing inwards on featureless ground could we tell "I" am in a social occasion with somebody in front. This is accredited to our empirical knowledge of quickly judging - Whom is the cameraperson interacting with? Where and when does the scene unfold? With the help of such context, we make incredibly accurate inference

[**] Corresponding author: Z. Dai
e-mail: `z.dai1@aston.ac.uk` (Zhuangzhuang Dai), `vu.huy@nordicsemi.no` (Vu Tran), `andrew.markham@cs.ox.ac.uk` (Andrew Markham), `niki.trigoni@cs.ox.ac.uk` (Niki Trigoni), `mir6zw@virginia.edu` (M Arif Rahman), `lnw8px@virginia.edu` (L. N. S. Wijayasingha), `jas9f@virginia.edu` (John Stankovic), `cl@mp.aau.dk` (Chen Li)

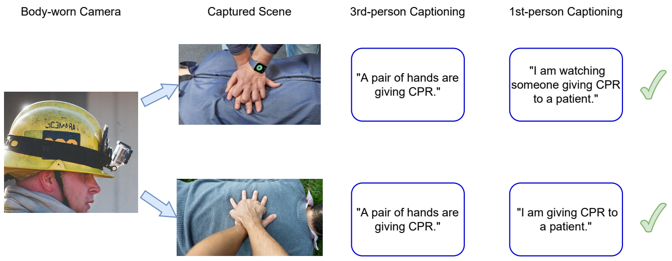

Body-worn Camera
Captured Scene
3rd-person Captioning
1st-person Captioning
"A pair of hands are giving CPR."
"I am watching someone giving CPR to a patient."
✓
"A pair of hands are giving CPR."
"I am giving CPR to a patient."
✓

Fig. 1. First-person captioning resolves ambiguity where third-person fails.

of the situation even from poorly presented angles, objects, or blurriness.

In this work, we tackle the under-explored problem of egocentric image captioning. Ego-captioning aims at human-understandable interpretation of vision data which is crucial for various life-logging applications. Prospective use cases include auto calorie intake recording for people on a diet, Bolanos et al. (2017), daily activity tracking for patients, Fan et al. (2018), and event summarization for emergency responders, Dai et al. (2022). As is shown in Fig. 1, we note 1st-person captioning provides a precise perspective in storytelling, whereas, a 3rd-person narrative poses ambiguity. Moreover, a first-person narrator places the viewer at the centre of the action and lends credence to the narration, Fan et al. (2018). Thereby, ego-captioning is also critical for artificial intelligence to establish the notion of “self”. It enables a machine to distinguish itself as a participant or an observer, which is vital to avoid misconceptions in scene understanding.

Nonetheless, the state-of-the-art data-driven captioning has largely focused on describing the contents objectively, i.e., in a 3rd-person narrative, Anderson et al. (2018). This results in most captioning datasets being labelled in the 3rd person, such as COCO, Lin et al. (2014), and MSR-VTT, Xu et al. (2016). We note first-person captions cannot be easily created from 3rd-person captions. Specific challenges are: (i) Syntax of a 1st-person narrative is semantically different from a 3rd-person narrative; (ii) Egocentric vision data are often contaminated by noise caused by motion, occlusion, and awkward camera angles which affect the accuracy of extracting object attributes or key features; (iii) Empirical implications of the cameraperson’s status have to be made as she/he is usually outside the field of view. Despite the widespread use of wearable cameras, video stream recording remains particularly power-and-memory-consuming. Existing life-logging datasets, e.g., *EDUB-SegDesc*, Bolanos et al. (2017), or *Deepdiary*, Fan et al. (2018), record sequences of images shot at 2 frames per minute. In this paper we scope our research to ego-image captioning.

The deep learning approach to 3rd-person captioning has drawn increasing attention for its success in learning visual-semantic representations. In contrast, previous attempts to generate ego-captions using templates or beam search based on mixed 1st- and 3rd-person captions, Fan et al. (2018), deliver inconsistent results. On the other hand, we notice contextual cues, such as combination of instances often seen indoor or outdoor (*where*), illumination condition (*when*), and interacting human/objects (*whom*), play a significant role in facilitating human-alike scene understanding. We are, thus, inspired to fuse such “contexts” in the caption generation process. We argue an *attention* mechanism should be applied equally to the visual features and the contextual knowledge to tackle the perception challenges in an ego-perspective. Transformers, Vaswani et al. (2017), originally initiated for sequential Natural Language Processing (NLP), emerges as a natural baseline for its strength in addressing long-term syntactic dependencies, Dosovitskiy et al. (2021).

To this end, we create a new dataset, *EgoCap*, comprising life-logging images with five ego-captions each to generate 1st-person captions consistently. We select source images from prevailing datasets (COCO, Lin et al. (2014), MSVD, Chen and Dolan (2011), MSR-VTT, Xu et al. (2016), and Ego4D, Grauman et al. (2021)) to avoid privacy issues and to increase scene diversity. *EgoCap* incorporates contextual labels, namely *where*, *when*, and *whom*, through querying surveyors. We propose an enhanced transformer network, *EgoFormer*, that fuses the contextual knowledge using a stacked cross-attention layer alongside visual features. The *EgoFormer* comprises feature extractor backbones, a visual encoder, a context encoder, and a captioning decoder. We take a strategy of pre-training the visual encoder and decoder on the COCO dataset to grant the model visual-semantic and syntactic capabilities before fine tuning it on *EgoCap*. Likewise, we take the context encoder pre-trained on ImageNet, Deng et al. (2009), to broaden its horizon of concept recognition before fine tuning it with context classifications. We evaluate the performance of *EgoFormer* on *EgoCap*, and on a released set from Deepdiary, Fan et al. (2018). Our proposed model demonstrates state-of-the-art performance improvement based on various machine translation metrics.

Our major contributions are: (1) We release a dataset for egocentric image captioning which, to the best of our knowledge, is the first that can support end-to-end learning; (2) We propose a transformer-based network with visual-context fusion modules to conduct ego-captioning with enhanced contextual awareness; (3) Extensive experiments demonstrate the superiority of the proposed approach to egocentric image captioning.

2. Related Work

2.1. Data-driven Captioning

Ever since Vaswani et al. (2017) proposed the transformer architecture, it has surpassed convolution-based models, Donahue et al. (2015); Anderson et al. (2018), to become state of the art for visual-captioning tasks. Numerous augmented transformers then emerged to improve performance further, Zhao et al. (2019); Zhang et al. (2021a). Multiple self-attention heads and hierarchies of attention were utilized in Huang et al. (2019). Cornia et al. (2020) introduced memory cells and skip connections to fully exploit low- and high-level features of the attention layers.

Transformers are outstanding in capturing long-term dependencies. To date egocentric image captioning has stagnated in the CNN-RNN era, Singh et al. (2016). It is found allowing multi-pass attention to the greedy decoded caption, Barraco et al. (2022), produces better results. However, this not

Table 1. A comparison of existing egocentric or captioning datasets.

Datasets	Size	Labels					
		Diverse	*OD*$^\diamond$	*HAC**	3rd-cap	1st-cap	Context
COCO Lin et al. (2014)	118*K*	✓	✓		✓		
MSVD Chen and Dolan (2011)	1.9*K*	✓			✓		
MSR-VTT Xu et al. (2016)	10*K*	✓			✓		
Charades-Ego Sigurdsson et al. (2018)	4*K*			✓			
EPIC-Kitchens Damen et al. (2018)	100 *h*		✓	✓			
Deepdiary Fan et al. (2018)	7.7$K^\triangle$			✓		✓	
EDUB-SegDesc Bolanos et al. (2017)	1.3$K^\dagger$					✓	
Ego4D Grauman et al. (2021)	3025 *h*	✓		✓			
EgoCap	2.1*K*	✓				✓	✓

◇ Object Detection.

★ Human Activity Classification.

△ Fewer than 300 images are released for privacy concerns.

† Unavailable for download.

only increases the inference complexity but presumes all texts are rich in semantics and significance, which does not apply to ego-captions. The recently published CoCa captioner, Yu et al. (2022), provides a paradigm of empowering a feature encoder with image recognition capabilities via ImageNet, Deng et al. (2009), pre-training. Meanwhile, many works have extracted and adopted additional features from the images, a.k.a., the bottom-up and top-down approach, Anderson et al. (2018). Geometric features, object-relation features, and semantic groundings, Sen et al. (2020), are used to enhance captioning. However, these additional features rely on the success of object detection from the image, Ren et al. (2015). We notice viability of detecting objects for egocentric images is impeded owing to mere object cues available, e.g., in bad camera angles. In NLP, it is recognized that coherent texts can be synthesized through attention, Zhao et al. (2019). This lends captioning models the power of comprehending the scene using external sources of information. Hence, this research is inspired by fusing additional contexts to enhance egocentric visual captioning in a CoCa-alike way.

2.2. Existing Egocentric or Captioning Datasets

Although popular 3rd-person captioning datasets, such as COCO, Lin et al. (2014), are valuable sources, they cannot be directly used for ego-captioning. Current egocentric visual captioning datasets are limited in either scale or diversity as shown in Table 1. Charades-Ego, Sigurdsson et al. (2018), and EPIC-Kitchens, Damen et al. (2018), provide class labels of Human Activity Classification (HAC) only, and are constrained in scene diversity. Deepdiary, Fan et al. (2018), and EDUB-SegDesc, Bolanos et al. (2017), combined release fewer than 300 ego-image samples in total due to privacy concerns. Ego4D, Grauman et al. (2021), is a large-scale egocentric video dataset collected across the globe. Unfortunately, annotations of Ego4D only provide HAC labels and template-based captions like "*A* interacts with *B*". We contrast *EgoCap* with existing datasets in Table 1. In sum, there is currently a lack of sizeable datasets supporting egocentric captioning studies.

3. EgoCap

We create a first egocentric image captioning dataset, *EgoCap* [1], each image of which comprises five captions in only first-person narrative alongside three contextual labels (*where*, *when*, and *whom*). We collect source images randomly from datasets widely acknowledged in visual-semantic studies, including COCO, Lin et al. (2014), MSVD, Chen and Dolan (2011), MSR-VTT, Xu et al. (2016), and Ego4D, Grauman et al. (2021). This not only maximizes scene diversity but evades privacy concerns for release. The context labels, which are obtained through polling five surveyors with expertise in vision data processing, are regarded as probability distributions for contextual representation learning.

EgoCap is comprised of 2079 images (1252 from Ego4D; 289 from COCO; 218 from MSVD; 320 from MSR-VTT) and over 10K egocentric captions alongside 6.3K contextual tags. We also retrieved weak labels of 3rd-person captions or HACs from their source datasets, and associated them for reference. Figure 2 shows the composition of *EgoCap* from multiple viewpoints. To the best of our knowledge, this is a first sizable dataset, with labelled contextual information, that allows end-to-end ego-caption learning.

4. EgoFormer

We observe object detection constantly fails in extracting useful information from ego-images. This prevents the conventional "bottom-up" method, Anderson et al. (2018), from resolving the caption through identifying key attributes in the scene. In fact, COCO dataset reports 3.5 categories of objects and 7.7 instances on average per image, Lin et al. (2014). However, after applying Faster R-CNN, Ren et al. (2015), a *de facto* approach to acquiring bottom-up features, we notice only 1.87 categories of objects and 2.98 instances are detected in average from *EgoCap*. Furthermore, many ego-images reveal 0 detected objects which utterly prevents "bottom-up" feature integration.

[1] Available at: `https://drive.google.com/drive/folders/10u8kBlrqi9sFiXZrouP6FChypen4dcFz`

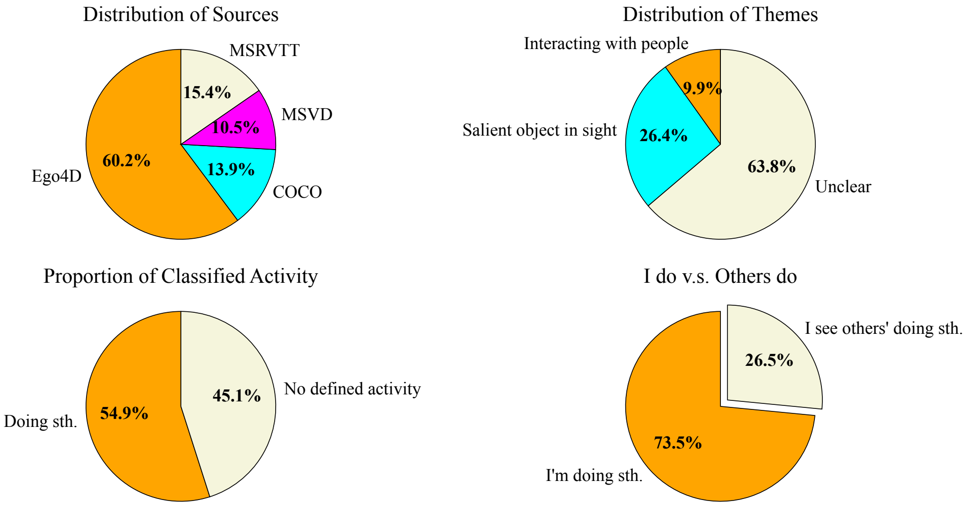


Fig. 2. The composition of *EgoCap* (top left). Distribution of image themes (top right). Proportion of images with explicitly labelled activity or not (bottom left). Proportion of "I do" versus "I see others do" in those with classified activity labels (bottom right).

Now that fine-grained attribute features of images are unavailable in egocentric footage, we argue the *where*, *when*, and *whom* knowledge beyond the field-of-view play a key role in ego-captioning. Specifically, we provide the captioning engine with such contextual knowledge to be able to judge the spatial-temporal conditions and interacting object(s) in which the scene unfolds. To this end, we propose *EgoFormer*, a two-stream transformer based network to generate ego-caption with enhanced contextual awareness. One encoder (visual ViT, Dosovitskiy et al. (2021)) learns visual features from the ego-image patches. This ViT is pre-trained on the COCO dataset, together with the decoder, to establish visual recognition and semantic capabilities. The other encoder (context ViT), pre-trained on ImageNet, is trained to master multi-label classification of *where*, *when*, and *whom* contexts. Obtained context features and visual features are fused in a stacked MHA (Multi-Head Attention) module. The design of *EgoFormer* is sketched in Fig. 3 and explained in detail in the following subsections.

4.1. Visual Encoder

The transformer model has demonstrated state-of-the-art performance in various visual-semantic tasks since its proposal, Vaswani et al. (2017). We follow the encoder-decoder paradigm for captioning. We first input the images to feature extractor backbones for visual token extraction. In order to emulate 1D sequences from images, we add 2D positional encoding to the evenly split image patches. We take the outputs, $I_{CNN} \in R^{H' \times W' \times C'}$ from the activation of the last convolutional layer, and flatten the first two dimensions to produce a 1D sequence. Average pooling is used to derive flattened representations from the convolutional layer. We use a convolutional layer of 1×1 kernal size to rescale the channel size to the transformer hidden state size, d_m (d_m = 256 in our experiments).

The visual ViT encoder consists of N stacked modules of MHA (i.e., $SelfAttn$) and feed forward networks (FFN) to learn visual representations, $\mathbf{V}_N$, of ego-images. The first MHA layer takes as input the queries, $\mathbf{Q} = XW_Q$, keys, $\mathbf{K} = XW_K$, and values, $\mathbf{V} = XW_V$, which are split into h heads for joint attention in the sub-spaces. Each head, h_i, conducts MHA among $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$. The FFN layer is comprised of two linear layers with a *ReLU* activation and dropout (of rate 0.1) after the first. The visual representations can be expressed as;

$$\mathbf{V}_N = FFN\left(MHA_N(\mathbf{Q}, \mathbf{K}, \mathbf{V})\right) \tag{1}$$

In each sub-layer, there is a residual connection and a normalization layer. All stacked layers follow the same design.

4.2. Context Encoder

Ego-images are often corrupted by occlusion or bad camera angles as seen in *EgoCap* (examples are shown in Supp. Materials). We humans have the ability to judge the situation from partially visible matters and casual camera angles, Xiao et al. (2020). Thus, it is of great importance for ego-captioning models to possess such contextual awareness. We propose a separate context encoder and utilize a novel decoder architecture to fuse the contextual representations to facilitate attention to both visual and contextual cues.

We design a context ViT encoder to perform contextual representation learning. This context ViT encoder is pre-trained on image classification tasks (ImageNet, Deng et al. (2009)) before fine tuning on the context multilabel classifications of *EgoCap*. The loss function of the classifier is the sum of the three-head cross-entropy losses,

$$Loss(\theta_C) = \sum_{i}^{K} CrossEntropy(\rho_i, \varphi_i) \tag{2}$$

where $K = 3$ for the three-head classification; θ_C stands for context ViT parameters; ρ are the prediction probabilities; and φ are probability distributions of the contextual labels.

The context ViT encoder uses ResNet-101, He et al. (2015), as backbone feature extractor. The ViT consists of M stacked layers of MHA, each containing h' attention heads, followed

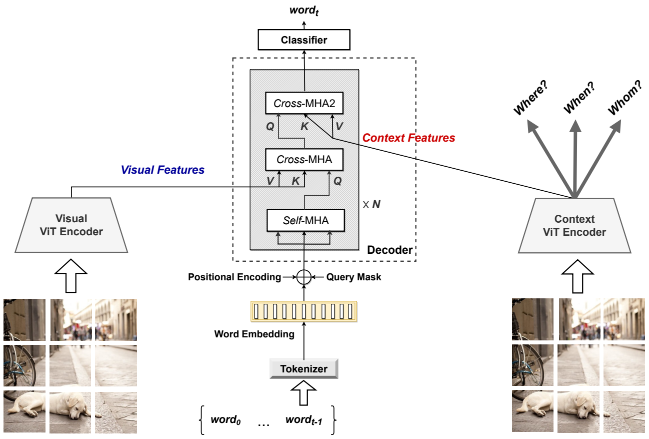


Fig. 3. Network architecture of *EgoFormer* decoder with context fusion.

by FFNs to learn contextual representations, $\mathbf{C}_M$, of the ego-images. We utilize a ViT encoder of hidden state size 768 pre-trained on ImageNet-21K. It is attached to a 3-layer fully-connected classifier of 768, $d_m \times 197$, and 512 neurons, respectively. Prior to the 2nd layer, context feature tokens are flatten from a 2D matrix into 1D vectors.

4.3. Context-Aware Caption Decoder

To utilize context, we want to grant greater attention to the domains in which the contextual knowledge lies. The MHA stands out for this task as it is elaborated for attention. Besides visual feature attention, we propose to conduct attention to the context features extracted from the first hidden layer of the context classifier. This is accomplished through a stacked MHA layer post visual MHA. The network architecture is illustrated in Fig. 3. Such an enhanced visual-context attention module can be expressed as;

$$\mathbf{X'}_j = MHA_j^V \left(SelfAttn(\mathbf{X}_{j-1}) W_Q^V, \mathbf{V}_N W_K^V, \mathbf{V}_N W_V^V \right) \quad (3)$$

$$\mathbf{X}_j = FFN_j \left(MHA_j^C(\mathbf{X'}_j W_Q^C, \mathbf{C}_M W_K^C, \mathbf{C}_M W_V^C) \right) \quad (4)$$

where $\mathbf{X}_j$ are the decoder hidden states of the j-th module. Each sub-layer is followed by a residual connection and a normalization layer. The visual-context attention layer duplicates in all N stacked decoder modules, each of which comes with h attention heads.

We use the BERT tokenizer, Devlin et al. (2019), as a word embedding layer to tokenize the decoder input sentence, which is a sequence starting with a start-of-sentence symbol and paddings to make up a maximum of 128 words. A word classifier is designed to predict the next word based on the last hidden states in an auto-regressive manner. We use three fully-connected layers of output sizes 512, 512, and 30522 (i.e., vocabulary size of the BERT tokenizer), respectively, with *ReLU* activations and dropout (of rate 0.1) after the first two layers. The next-word predictor can be written as;

$$Token_{id} = \arg\max_{id} FC_3(ReLU(FC_2(ReLU(FC_1(\mathbf{X}_N))))) \quad (5)$$

The loss function is cross-entropy loss of the next-word classifier (context classification heads are dropped in this end-to-end caption training phase). We follow the design of the original transformer, Vaswani et al. (2017), for the rest.

5. Evaluation

For *EgoFormer* and all comparison models, we use the Karpathy split, Karpathy and Fei-Fei (2017), of the COCO dataset to pre-train with recommended hyperparameters in the original works for a fair comparison. Rich categories of objects in COCO help the visual encoder and decoder learn the visual-semantic representations comprehensively. We use ImageNet 21K, Deng et al. (2009), to pre-train the context ViT encoder before fine tuning on the contextual labels of *EgoCap*. We eventually train the *EgoFormer* model from end-to-end on the image-caption pairs of *EgoCap* with 10% of the pre-training learning rates. We split the *EgoCap* dataset with 1579 samples for training, 200 for validation, and 300 for testing. Our quantitative and qualitative analysis, and an ablation study, are presented below.

5.1. Quantitative Comparison with Other Methods

We utilize common machine translation metrics for benchmarkings. Note bottom-up approaches are impractical on *EgoCap* due to many samples coming with 0 detected objects, which deny object attributes. This rules out multiple state-of-the-art methods that work well on COCO (COCO samples guarantee object presence), such as Meshed-Memory transformer, Cornia et al. (2020), object relation transformer, Sen et al. (2020), or AoANet, Huang et al. (2019). We compare our *EgoFormer* with CNN-RNN, Vinyals et al. (2015), CNN-RNN with attention, Xu et al. (2015), the original transformer ($Trans$), Vaswani et al. (2017), transformer with concatenated context and visual features ($Trans_{Concat}$), Libovický et al. (2018), transformer with gated information fusion ($Trans_{GIF}$), Zhao et al. (2019), RSTNet, Zhang et al. (2021a), CaMEL, Barraco et al. (2022), COSNet, Li et al. (2022), and CLIP-ViL, Shen et al. (2022). These models are fine-tuned on *EgoCap* with the same testing set. Contexts are not infused in testing these models as they do not have context encoders. The evaluation results are summarized in Table 2.

Our proposed model scores the highest over other candidates, especially on CIDEr which is deemed producing better human consensus. The *EgoFormer* exceeds the original transformer by 38 units and sees a twofold improvement over those based on a conventional CNN-RNN with attention. This proves the efficacy of transformers exploiting contextual knowledge. Recent studies of fusing additional information for captioning show promising results, Sen et al. (2020). We experimented by concatenating visual and contextual sequences presented in Libovický et al. (2018), or using the paradigm proposed in Zhao et al. (2019) to fuse the context representations as external information through a gated soft switch. Their performance ends up poorer than *EgoFormer*. We tested RSTNet, CaMEL, COSNet, and CLIP-ViL without self critical sequence training on *EgoCap* due to a limited size for the reinforcement learning phase. Although RSTNet integrates grid

Table 2. Evaluation of *EgoFormer* in comparison to other captioning models on *EgoCap*.

$Model$	$BLEU1$	$BLEU2$	$BLEU3$	$BLEU4$	$METEOR$	$ROUGE-L$	$CIDEr$	$SPICE$
$CNN\,RNN$ Vinyals et al. (2015)	42.70	28.96	21.15	15.30	-	41.09	33.14	-
$CNN\,RNN_{Atten}$ Xu et al. (2015)	52.68	38.27	28.95	21.43	-	47.25	51.26	-
$Trans$ Vaswani et al. (2017)	63.04	48.37	37.39	28.53	28.54	54.87	87.23	14.16
$Trans_{Concat}$ Libovický et al. (2018)	63.95	50.51	41.04	32.33	32.34	54.91	86.55	16.45
$Trans_{GIF}$ Zhao et al. (2019)	70.77	59.74	47.87	**38.10**	37.51	61.77	124.79	18.18
$RSTNet$ Zhang et al. (2021a)	69.60	57.18	44.61	34.31	35.66	60.68	114.05	18.88
$CaMEL$ Barraco et al. (2022)	**71.98**	**60.12**	47.31	38.09	37.91	61.54	123.76	**20.09**
$COSNet$ Li et al. (2022)	70.45	58.80	47.19	37.52	37.25	61.63	123.20	19.06
$CLIP-ViL$ Shen et al. (2022)	69.42	57.19	44.68	34.49	35.96	60.17	116.34	19.35
$EgoFormer$	70.93	59.69	**47.89**	37.93	**38.01**	**61.91**	**125.52**	19.93
$EgoFormer_{small}$	69.27	57.16	44.73	34.54	35.85	60.48	110.71	16.92
$EgoFormer_{tiny}$	64.03	51.56	39.33	29.82	33.46	57.75	91.88	14.41

Table 3. *EgoFormer* benchmarks on COCO. *Bk* short for *BLEUk*. R_L short for *ROUGE-L*. C short for *CIDEr*. S short for *SPICE*.

$Model$	$B1$	$B4$	R_L	C	S
$Trans$ Vaswani et al. (2017)	65.2	24.8	45.6	74.9	19.9
$AoANet$ Huang et al. (2019)	77.4	37.2	57.5	119.8	21.3
$M2$ Cornia et al. (2020)	80.8	39.1	58.6	131.2	22.6
$RSTNet$ Zhang et al. (2021a)	81.8	40.1	59.5	135.6	23.3
$CaMEL$ Barraco et al. (2022)	82.7	40.9	60.1	138.9	23.9
$COSNet$ Li et al. (2022)	82.7	42.0	60.6	141.1	24.6
$CLIP-ViL$ Shen et al. (2022)	-	39.2	-	130.3	23.0
$EgoF_{Blind\,ctx}$	66.2	25.6	45.8	76.0	19.7
$EgoF_{Raw\,ctx}$	66.2	25.6	45.8	75.9	19.7

Table 4. Ablation studies of *EgoFormer* on *EgoCap*.

$Model$	$B1$	$B4$	R_L	C
$EgoF$	**70.93**	**37.93**	61.91	**125.52**
$EgoF_{Ctx\,fuse\ in\ prior}$	70.32	36.70	**62.19**	121.47
$EgoF_{Ctx\,pos\,encoding}$	61.53	25.59	53.51	78.82
$EgoF_{Ctx\,word\,embedding}$	65.88	30.87	55.36	93.73
$EgoF_{Backbone\,ViT\ grad}$	62.92	30.50	55.08	94.30
$EgoF_{Backbone\,ViT\,no-grad}$	65.79	34.53	56.93	97.87
$EgoF_{Blind\,ctx}$	65.06	31.19	55.45	91.33
$EgoF_{Raw\,ctx}$	64.60	31.74	55.42	93.63
$EgoF_{Where\,only}$	65.48	32.58	56.76	102.11
$EgoF_{When\,only}$	65.89	32.96	57.16	100.86
$EgoF_{Whom\,only}$	65.59	33.18	56.86	101.24

features, poor camera angles may undermine efficacy of fusing spatial geometry features. Leveraging pre-trained CLIP models, Radford et al. (2021) , COSNet and CLIP-ViL turn out sub-optimised because, similarly, CLIP cannot predict objects and subjects accurately in severe visual degradation. CaMEL realizes optimum with several metrics thanks to its dual language decoders that learn first-person means of expression well. Our proposed *EgoFormer* delivers state-of-the-art performance with BLEU-3, METEOR, ROUGE-L, and CIDEr. Lesser performance is noticed with *EgoFormer* of more compact visual backbones such as ResNet-50 ($EgoFormer_{small}$) and ResNet-18 ($EgoFormer_{tiny}$).

5.2. Qualitative Analysis

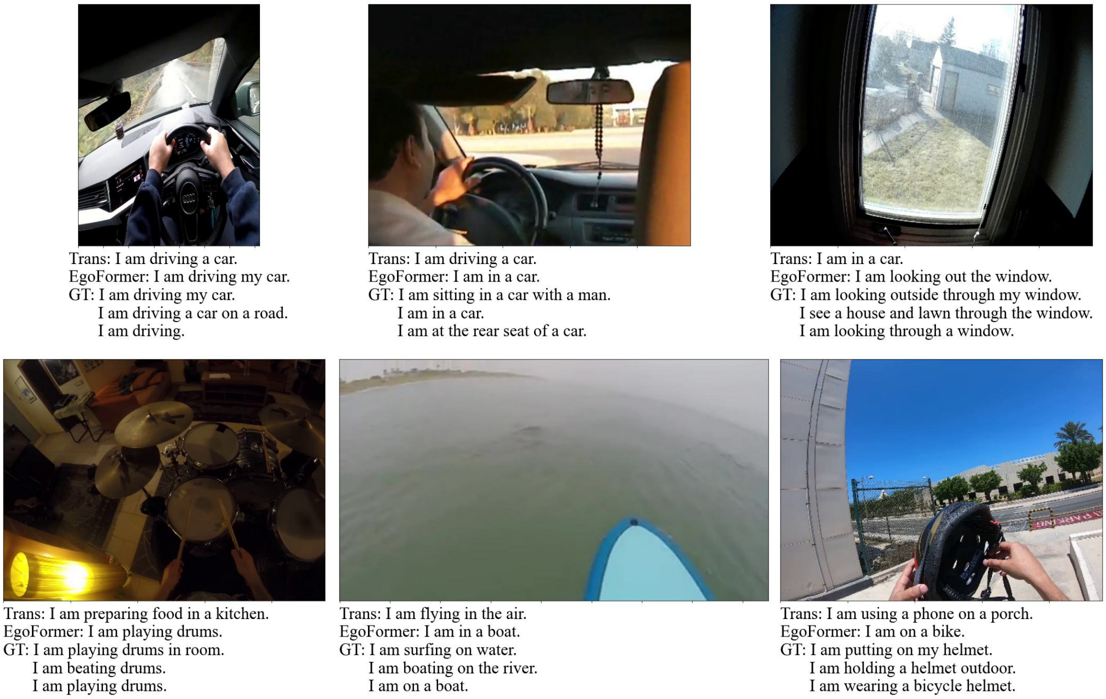
Qualitative studies are demonstrated in Fig. 4. In the first row, although both *EgoFormer* and the baseline recognize *driving*, the baseline cannot tell the fact that someone else is driving while "I" am a *passenger*. The baseline also tends to suffer from overfitting toward *car* wherever *window* occurs as is seen in the 3rd sample. It can be seen from the second row the *EgoFormer* is capable of recognizing correct activities or concepts in challenging camera angles. In the last sample, it is interesting *EgoFormer* reveals a *bike* from holding a *bike helmet* without actually seeing a *bike*. We refer the reader to Supp. Materials for more qualitative analysis, inclusive of Deepdiary, Fan et al. (2018), for out-of-domain tests.

5.3. EgoFormer for Third-Person Caption

Although contexts for COCO dataset are not available, we examine *EgoFormer* architecture's generalization in resolving third-person captioning. We evaluate *EgoFormer* on COCO without pre-training ($EgoF_{Blind\,ctx}$), or with ImageNet pre-trained context ViT ($EgoF_{Raw\,ctx}$), in contrast to original transformer and state-of-the-art transformers (with object attributes) shown in Table 3. Since "contexts" do not undergo proper representation learning in both cases, $EgoF_{Blind\,ctx}$ and $EgoF_{Raw\,ctx}$ unsurprisingly fall behind state-of-the-art methods.

5.4. Ablation Study

In *EgoFormer*, the context fusion takes place post visual attention. We investigate the effect of fusing the context prior to visual (*Ctx fuse in prior*). Since the context tags are in the form of 1-out-of-3 options, a question arises whether they could be regarded as some sort of encoding where these options are simply differentiated via adding positional encoding. We experiment by adopting a learnable embedding layer (*Ctx pos encoding*) with an input size of 3×3 to map the context selections into features in which the dimension is equivalent to transformer hidden states, d_m. Transformer with a knowledge graph, Zhang et al. (2021b), proposes to embed semantic keywords together with their neighbours to enrich the caption generation. We evaluate this by fusing the wording embedding of the top-4 synonyms of *indoor/outdoor*, *daytime/night*, and *object/human* (masked tokens for *ambiguous*) using a stacked MHA (*Ctx word embedding*). An important question is whether ego-captioning relies on the same visual representations as 3rd-person captioning (pre-trained on COCO) which are extracted by a frozen backbone CNN. Whereby, we test this by allowing gradient descents for the backbone CNN parameters when



Trans: I am driving a car.
EgoFormer: I am driving my car.
GT: I am driving my car.
I am driving a car on a road.
I am driving.

Trans: I am driving a car.
EgoFormer: I am in a car.
GT: I am sitting in a car with a man.
I am in a car.
I am at the rear seat of a car.

Trans: I am in a car.
EgoFormer: I am looking out the window.
GT: I am looking outside through my window.
I see a house and lawn through the window.
I am looking through a window.

Trans: I am preparing food in a kitchen.
EgoFormer: I am playing drums.
GT: I am playing drums in room.
I am beating drums.
I am playing drums.

Trans: I am flying in the air.
EgoFormer: I am in a boat.
GT: I am surfing on water.
I am boating on the river.
I am on a boat.

Trans: I am using a phone on a porch.
EgoFormer: I am on a bike.
GT: I am putting on my helmet.
I am holding a helmet outdoor.
I am wearing a bicycle helmet.

Fig. 4. Qualitative analysis of *EgoFormer* compared with original transformer (*Trans*) on *EgoCap* dataset. GT is short for ground truth.

training *EgoFormer* (*Backbone ViT grad*), and freezing both the backbone CNN and the ViT encoders (*Backbone ViT no-grad*), respectively. We also investigate the necessity of two-stage pre-training of context ViT encoder by skipping the ImageNet transfer learning (*Blind ctx*) or skipping the *EgoCap* context classifications (*Raw ctx*). Lastly, we test the contribution of contexts by fine tuning the classifier with *where*, *when*, or *whom* only as context.

Following an identical training strategy, the ablative outcomes are shown in Table 4. Fusing the contexts prior to visual features appears sub-optimal regarding all metrics except R_L. The context positional encoding method turns out the worst most likely because the semantics of the context information are overlooked. We find the performance is degraded when fusing word embeddings of contextual keywords and their synonyms directly. The overall evaluation improves in order from allowing the backbone and encoders to learn, to freezing both backbone and encoders, to only freezing the backbone (*EgoFormer*) during training. This implies the ViT encoders have to adapt which proves *EgoFormer* is learning unique visual-context representations to arrive at optimized ego-captions. Nonetheless, the backbone CNN outputs had better stay constant as sequences of feature tokens. The CIDEr scores deteriorate by 34 units and 32 units without ImageNet and *EgoCap* context pre-training phases, respectively. Nonetheless, fine tuning with any of the contexts appears to optimize the contextual representations to an extent. In summary, the *EgoFormer* accomplishes top-2 in every metric if not the best.

5.5. Computational Resources

We use an NVIDIA RTX A6000 GPU with 48 GB memory to train *EgoFormer* including all pre-training phases, comparative models, and ablative studies. It takes 20 minutes to perform fine tuning of the ViT context encoder. Pre-training the captioner's encoder-decoder backbone network on COCO costs 30 hours. End-to-end training of *EgoFormer* takes about 4 hours.

6. Conclusion

Body-worn camera footage from life-logging, patients, or emergency responders, have witnessed a great surge. Nonetheless, first-person captioning which reflects a veracious perspective of depicting ego-images has been under-explored compared to 3rd-person ones. We create a dataset of egocentric image-caption pairs with contexts. We propose a novel transformer network to fuse the contextual knowledge which brings about state-of-the-art captioning on *EgoCap*. Crucially, our findings shed light on image captioning when object cues are absent.

Upon successfully fusing the contexts, we plan to expand "contexts" in *EgoCap* with greater concreteness, such as more fine-grained details about location ("kitchen/bedroom/in transportation/shopping mall" etc.) and time ("sunny dawn/rainy day/stormy night" etc.). It is planned to leverage crowdsourcing to constantly grow the size and diversity of *EgoCap*. We want to integrate hand gesture, eye contact, and auditory inputs to enrich contexts in broader senses. We think ego-image captioning is a vital step towards the more challenging ego-video captioning which we will address in future work.

Acknowledgments

This research is supported by EPSRC project "*ACE-OPS: From Autonomy to Cognitive assistance in Emergency OPerationS*" (EP/S030832/1) and by NIST project "*Pervasive, Accurate, and Reliable Location Based Services for Emergency*

Responders". This work was also supported, in part, by the award 70NANB21H029 from the U.S. Department of Commerce, NIST, and by AI for the People and Aalborg Robotics Challenge (ARC) Bridging Project.

-Experiments were run on Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Barraco, M., Stefanini, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R., 2022. Camel: Mean teacher learning for image captioning. 2022 26th International Conference on Pattern Recognition (ICPR) , 4087–4094URL: `https://api.semanticscholar.org/CorpusID:247025790`.

Bolanos, M., Peris, A., Casacuberta, F., Soler, S., Radeva, P., 2017. Egocentric video description based on temporally-linked sequences. Journal of Visual Communication and Image Representation 50.

Chen, D.L., Dolan, W.B., 2011. Collecting highly parallel data for paraphrase evaluation, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics), pp. 190–200.

Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2020. Meshed-Memory Transformer for Image Captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Dai, Z., Saputra, M.R.U., Lu, C.X., Tran, V., Wijayasingha, L.N.S., Rahman, M.A., Stankovic, J.A., Markham, A., Trigoni, N., 2022. Deep odometry systems on edge with ekf-lora backend for real-time indoor positioning, in: 2022 Workshop on Cyber Physical Systems for Emergency Response (CPS-ER), pp. 1–6.

Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M., 2018. Scaling egocentric vision: The epic-kitchens dataset, in: European Conference on Computer Vision (ECCV).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Ieee. pp. 248–255.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186.

Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16×16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR).

Fan, C., Zhang, Z., Crandall, D., 2018. Deepdiary: Lifelogging image captioning and summarization. Journal of Visual Communication and Image Representation 55, 40–55.

Grauman, K., Westbury, A., Byrne, E., et. al., 2021. Ego4d: Around the world in 3,000 hours of egocentric video. `arXiv:2110.07058`.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. Computing Research Repository (CoRR) abs/1512.03385. `arXiv:1512.03385`.

Huang, L., Wang, W., Chen, J., Wei, X., 2019. Attention on attention for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Karpathy, A., Fei-Fei, L., 2017. Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 664–676.

Li, Y., Pan, Y., Yao, T., Mei, T., 2022. Comprehending and ordering semantics for image captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi:`10.48550/arXiv.2206.06930`.

Libovický, J., Helcl, J., Mareček, D., 2018. Input combination strategies for multi-source transformer decoder, in: Proceedings of the Third Conference on Machine Translation (WMT).

Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2014. Microsoft coco: Common objects in context. CoRR abs/1405.0312.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 8748–8763.

Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks., in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Neural Information Processing Systems (NIPS), pp. 91–99.

Sen, H., Wentong, L., Tavakoli, H.R., Yang, M.Y., Rosenhahn, B., Pugeault, N., 2020. Image captioning through image transformer, in: Asian Conference on Computer Vision (ACCV).

Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K., 2022. How much can clip benefit vision-and-language tasks?, in: International Conference on Learning Representations (ICLR).

Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K., 2018. Actor and Observer: Joint Modeling of First and Third-Person Videos, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, Utah, United States. pp. 7396–7404.

Singh, S., Arora, C., Jawahar, C.V., 2016. First person action recognition using deep learned descriptors, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Neural Information Processing Systems (NIPS).

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator., in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society. pp. 3156–3164.

Xiao, H., Xu, J., Shi, J., 2020. Exploring diverse and fine-grained caption for video by incorporating convolutional architecture into lstm-based model. Pattern Recognition Letters 129, 173–180.

Xu, J., Mei, T., Yao, T., Rui, Y., 2016. Msr-vtt: A large video description dataset for bridging video and language, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning (ICML), JMLR.org. p. 2048–2057.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y., 2022. Coca: Contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res. 2022. URL: `https://api.semanticscholar.org/CorpusID:248512473`.

Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R., 2021a. Rstnet: Captioning with adaptive attention on visual and non-visual words, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15460–15469. doi:`10.1109/CVPR46437.2021.01521`.

Zhang, Y., Shi, X., Mi, S., Yang, X., 2021b. Image captioning with transformer and knowledge graph. Pattern Recognition Letters 143.

Zhao, S., Sharma, P., Levinboim, T., Soricut, R., 2019. Informative image captioning with external sources of information, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6485–6494.