

Hierarchical Context-aware Network for Dense Video Event Captioning

Lei Ji^{1,2,3}*, Xianglin Guo⁴*, Haoyang Huang³, Xilin Chen^{1,2}

¹Institute of Computing Technology, CAS, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

³Microsoft Research Asia, Beijing, China

⁴New York University, New York, USA

lei.ji@microsoft.com, xg893@nyu.edu

haoyang.huang@microsoft.com, xlchen@ict.ac.cn

Abstract

Dense video event captioning aims to generate a sequence of descriptive captions for each event in a long untrimmed video. Video-level context provides important information and facilitates the model to generate consistent and less redundant captions between events. In this paper, we introduce a novel **Hierarchical Context-aware Network** for dense video event captioning (HCN) to capture context from various aspects. In detail, the model leverages local and global context with different mechanisms to jointly learn to generate coherent captions. The local context module performs full interaction between neighbor frames and the global context module selectively attends to previous or future events. According to our extensive experiment on both Youcook2 and Activitynet Captioning datasets, the video-level HCN model outperforms the event-level context-agnostic model by a large margin. The code is available at <https://github.com/KirkGuo/HCN>.

1 Introduction

With the increase of video data uploaded online every day, the acquisition of knowledge from videos especially for Howto tasks is indispensable for people’s daily life and work. However, watching a whole long video is time-consuming. Existing technologies focus on two main research directions to compact video information: video summarization to trim long videos to short ones and (dense) video captioning to generate a textual description of the key events in the video. Typically for long untrimmed videos, dense video event captioning generates fine-grained captions for all events to facilitate users quickly skimming the video content and enables various applications e.g. video chaptering and search inside a video.

*Equal contribution

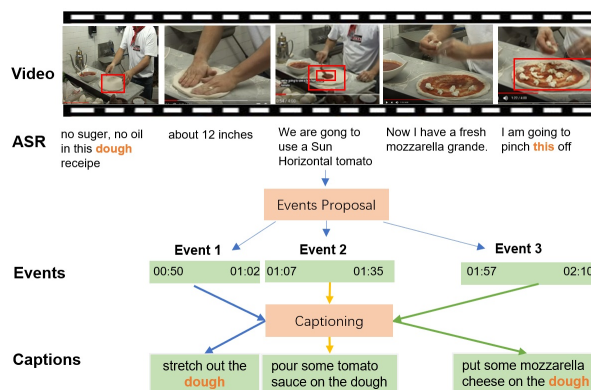


Figure 1: A showcase of dense video event captioning. Given a video and the speech text, the task is to generate event proposals and captions.

Dense video event captioning (Krishna et al., 2017) and multi-modal video event captioning (Iashin and Rahtu, 2020b) aims to generate a sequence of captions for all events regarding to uni-modality (video) or multi-modality (video + speech) inputs. Figure 1 presents a showcase, which demonstrates the challenges of this task from both vision and speech text perspective. For *vision* understanding, the fine-grained objects are hard to recognize due to ambiguity, occlusion, or state change. In this case, the object "dough" is occluded in event 1 and is hard to recognize from the video. However, it can be recognized from the previous neighbor video frame with a clear appearance. From *speech text* perspective, although the speech text offers semantic concepts (Shi et al., 2019; Iashin and Rahtu, 2020b), it brings another challenge of co-reference and ellipsis in speech text due to the informal utterance of oral speeches. In the case of Figure 1, the entity "dough" in event 3 is an ellipsis in the text. Nonetheless, it is capable of generating consistent objects "dough" in event 3 with the contextual information from other events such as event 1 in this example. To sum up, both

local neighbor-clip and global inter-event contexts are important for event-level captioning to generate coherent and less duplication descriptions between events.

Previous endeavors widely used recurrent neural network (Krishna et al., 2017) which suffers from capturing long dependency, while recently attention-based model (Zhou et al., 2018b; Sun et al., 2019b,a) is becoming the new paradigm for dense video event captioning and effective for multi-modal video captioning (Shi et al., 2019; Iashin and Rahtu, 2020b). However, existing attention-based models generate the captioning only relying on the video clip inside each event, and ignore video-level local and global context. Motivated by this, we mainly investigate how to effectively and jointly leverage both local and global context for video captioning.

In this paper, we propose a novel hierarchical context-aware model for dense video event captioning (HCN) to capture both the local and global context simultaneously. In detail, we first exploit a local context encoder to embed the visual and linguistic features of the source and surrounding clips, then design a global context encoder to capture relevant features from other events. Specifically, we apply different mechanisms: a flat attention module between the source and local context; a cross attention module for the source to select the global context. With regards to the neighbor frames (temporally close) usually alike, e.g. with the same objects, the flat attention is a full interaction to generate *accurate* and *coherent* captions. Contemporaneously, the cross attention on global context can selectively attend to the relevant events and capture prior temporal dependency between events to generate *coherent* and *less duplicate* captions. The experimental results demonstrate the effectiveness of our model. Our contributions can be summarized as:

- 1) We propose a hierarchical context-aware model for dense video event captioning to capture video-level context.
- 2) We carefully design different mechanisms to capture both local and global context: a flat attention model with full interaction between neighbor frames and a cross attention model to selectively capture inter-event features.
- 3) Experimental results on both Youcook2 and Activitynet Captions dataset demonstrate the effectiveness of our models and outperforms the context-

agnostic model to a large extent.

2 Preliminary

The dense video event captioning task is to produce a sequence of events and generate a descriptive sentence for each event given a long untrimmed video. In this work, we focus only on the task to generate captions and directly apply the ground-truth event proposals similar to (Hessel et al., 2019; Iashin and Rahtu, 2020b). The paradigm for video captioning is an encoder-decoder network, which inputs video features and outputs descriptions for each event. In this section, we describe the task formulation including the context-agnostic model as well as the context-aware model in one framework.

2.1 Overview

Problem Definition We define a sequence of event segment proposals as $\mathbf{e} = \{e_i | i \in [1, m]\}$, representing the video with m proposals, e_i is the feature of the i -th event including both video and text feature, $e_i = \{v_i, t_i\}$, where v_i is video feature and t_i is transcript text feature (if available) of the i -th event. We take all the video frames and transcript tokens of the event between the start and end time. The number of video frames is likely to be different from the number of text tokens depending on the actual video clip. Given all events e , the goal is to predict the target descriptive sentences $\mathbf{Y} = \{y_i | i \in [1, m]\}$. Each y_i is a sequence of descriptive words corresponding to each event e_i . The probability of the expected sentences \mathbf{Y} .

$$P(Y|e) = \prod_{i=1}^m P(y_i | e_i) \quad (1)$$

which is to predict y_i conditioned on the event e_i . The context-aware model considers local context $v_{\neq i}$ (the neighboring video clip) and global context $e_{\neq i}$ (the clips of past and future events) respectively. The context-aware probability can be approximated as

$$P(Y|e) = \prod_{i=1}^m P(y_i | e_i, v_{\neq i}, e_{\neq i}) \quad (2)$$

3 Methodology

3.1 Context-agnostic model

The context-agnostic model of captioning is to generate a descriptive sentence given the short-trimmed video clip of each event. The paradigm

for multi-modal video captioning is an encoder-decoder network as in (Hessel et al., 2019). First, we pre-process each event and extract features separately. For the event e_i , we extract both video feature v_i and transcript feature t_i if available. Next, both the video features and transcript features are concatenated together as the input to the transformer encoder. This encoder implements self-attention of each modality and cross attention between both modalities in one unified transformer. Finally, a transformer decoder generates the text tokens of the description with the enhanced features.

3.2 Context-aware model

We propose a context-aware video event captioning model with a hierarchical context-aware network (HCN) and the architecture is a general framework for either uni-modal or multi-modal inputs as explained in Figure 2.

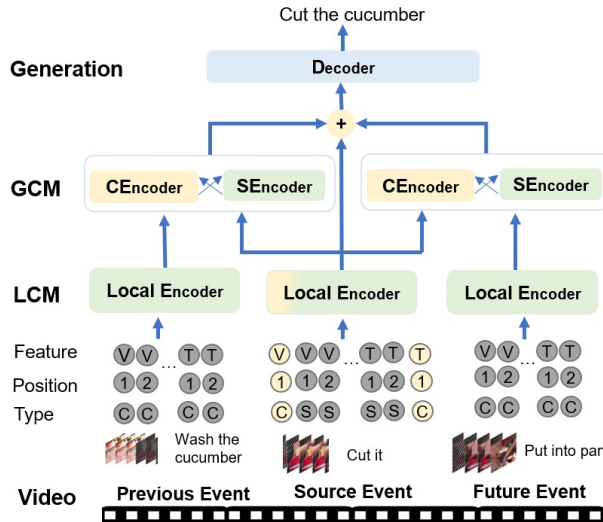


Figure 2: HCN provides a general framework with 4 modules: local context module (LCM), global context module (GCM), context gate, and decoder. LCM enhances the visual feature by local context and optionally fuse both visual and text features with multi-modal inputs. GCM employs a cross attention model to encode the source visual feature with other event features, which employs the SEncoder to encode source and context separately and adopts the CEncoder to selectively attend to context.

3.2.1 Multi-modal Feature Representation

For visual features, we adopt a pre-trained 3D feature extractor to extract k features as $v_i = \{v_j | j \in [1, k]\}$ of the i -th event. We further add a projection layer to map the raw feature to the input dimension through an embedding layer $f(v_i) =$

$\{e | e = Embedding(v_i)\}$. For transcript text, we tokenize the text into words and represent each word with 1-hot representation. The tokens within each event are represented as $t_i = \{t_j | j \in [1, l]\}$, where l is the length of the tokens corresponding to the number of the transcript text in the speech of the event. Moreover, we embed each token to continuous representation by an embedding layer $f(t_i) = \{e | e = Embedding(t_i)\}$. Similar to the work in (Hessel et al., 2019), we build the vocabulary using all tokens in the captioning sentence.

The input for each event comprises of three types of embedding: 1) visual feature $f(v_i)$ (and speech text feature $f(t_i)$ if available); 2) position embedding $p(v_i)$ and $p(t_i)$ as introduced in the transformer model (Vaswani et al., 2017); 3) type embedding $s(v_i)$ and $s(t_i)$ representing whether the current embedding is from context or source.

$$E(v_i) = [f(v_i) + p(v_i) + s(v_i)] \quad (3)$$

$$E(t_i) = [f(t_i) + p(t_i) + s(t_i)] \quad (4)$$

where $+$ is the add operator, $E(v_i)$ and $E(t_i)$ are the embeddings of video and text respectively. For multi-modal input, both visual and text features are concatenated for further processing.

We extract two types of contextual information: event-agnostic local context and event-aware global context. *Event-agnostic context* takes frames temporally close to the video event. Video is a continuous signal and neighboring video frames are likely to be semantically related to each other e.g. same objects. This is especially helpful for recognizing objects with state change or occluded in the current event. Moreover, objects are likely to be explicitly mentioned in the contextual transcript which can be used to deal with object co-reference and ellipsis typically for instructional videos. *Event-aware context* utilizes the video frames of both previous and future events, which attempts to model the relation between events. The global context provides overall features and prior knowledge of temporal dependency. Specifically for a particular domain like a recipe, the event “mix the flour and water” is often followed by “knead the dough”. This prior knowledge of event dependency learned from a global context is effective for understanding long videos.

3.2.2 Hierarchical Context-aware Network

The overall pipeline includes 4 modules: 1) the hierarchical model starts with a local context module (LCM) to encode the local context features,

the neighbor video clip temporally close to the event. Specifically, the LCM adopts a flat attention model similar to (Ma et al., 2020) to enhance the source video feature by local context. Besides, given multi-modal inputs, LCM is a general model to fuse both the visual features $f(v_i)$ and the text features $f(t_i)$ inside the event with one unified transformer as in (Hessel et al., 2019); 2) we further employ a global context module (GCM) to make the source event to interact with other event features flexibly. The GCM is a cross attention model, which contains one source encoder SEncoder and one cross encoder CEncoder. SEncoder is a self-attention module for encoding event features, and CEncoder is a cross attention module for interaction between source and context events; 3) the hierarchical context-aware model further combines both the neighbor-clip (around the event) or inter-event (other events) context from both previous and future using gating mechanism; 4) finally, an auto-regressive decoder is used to generate the sentence with a masked transformer model.

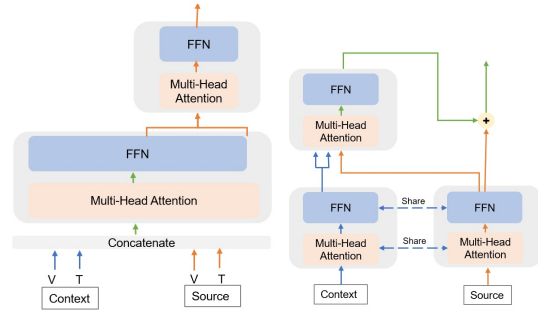
Local Context Module We first introduce the local context module to encode multi-modal source video features together with the event-agnostic context features (surrounding frames). The flat transformer in (Ma et al., 2020) is effective for encoding contextual information with full interaction between source and context features. In addition, when the speech text is available for multi-modal video captioning, this flat encoder can also perform the fusion of visual and text modalities, which is similar to (Hessel et al., 2019). To sum up, we employ one unified flat encoder to accomplish two actions: source-context interaction and multi-modal fusion as explained in Figure 3a.

$$E(e_i) = [E(v_i); E(t_i)] \quad (5)$$

$$H(m_i) = \text{FFN}(\text{MultiHead}([E(v_{i \pm k_l}); E(e_i)])) \quad (6)$$

$$H(e_i^l) = H(m_i)[i_1 : i_n] \quad (7)$$

where $[\cdot]$ is concatenation operation, FFN means the feed-forward network and MultiHead is the multi-head attention network in transformer (Vaswani et al., 2017). We apply residual connection for all components. We only perform equation 5 for multi-modal video event captioning, and $E(e_i)$ is the concatenation of the visual embedding and text embedding for the event i . We then feed the embedding $E(e_i)$ together with the embedding of neighbor frames $E(v_{i \pm k_l})$ into the



(a) Local context module (b) Global Context Module

Figure 3: The Local context module (LCM) is a flat attention model which adopts a unified attention model for interaction and fusion, and only selects the output of source embedding for further processing. Global Context Module (GCM) is a cross attention model, which adopts a cross attention model to selectively attend to context. Finally, a GRU gate \oplus is used to combine the context-enhanced feature with the source feature.

transformer blocks and get context-aware encoding $H(m_i)$, and k_l is the local context length. Finally, we only select the output of source encoding instead of using all embedding for further processing. Intuitively, the source is more important than the context. In equation 7, $H(e_i^l)$ is the hidden state of the source input, which requires the model to focus on the current source event, i_1 is the start of the event i and i_n is the end of the event i . LCM outputs the enhanced event representation by local context and multi-modal inputs.

Global Context Module We then illustrate the global context module to encode the output of LCM together with event-aware context (previous or future events). GCM is a cross attention model, which selectively attends to previous or future events to enhance the source video representation. Different from LCM, which applies a unified transformer to encode a short context, GCM exploits a cross attention model similar to (Maruf et al., 2019) to encode long global context efficiently. The unified transformer model is hard to deal with long input sequences due to complexity. The cross attention model facilitates the source to interact with each context event and can easily be scaled out for long videos. Figure 3b illustrates the GCM model structure.

We exploit the GCM for each contextual event and then combine all the encoding through a context gating mechanism similar to (Maruf et al., 2019). First, the self-attention module encodes each source or context event separately. Then, the

cross attention module empowers the source to attend to context.

$$H(\hat{e}_i) = \text{FFN}(\text{MultiHead}([H(e_i^l)])) \quad (8)$$

$$H(e_j) = \text{FFN}(\text{MultiHead}([E(e_j)])) \quad (9)$$

$$H(e_j^c) = \text{FFN}(\text{MultiHead}([H(\hat{e}_i), H(e_j)])) \quad (10)$$

where $H(\hat{e}_i)$ is the encoding of source event i , $H(e_j)$ is encoding of the j -th context event, and $H(e_j^c)$ is the source attended to the j -th event.

Next, we adopt a gated recurrent unit (GRU) (Cho et al., 2014) to selectively update the source feature with context enhanced feature which is shown to be effective in our ablation study.

$$z_j = \sigma(w_z H(\hat{e}_i) + u_z H(e_j^c) + b_z) \quad (11)$$

$$r_j = \sigma(w_r H(\hat{e}_i) + u_r H(e_j^c) + b_r) \quad (12)$$

$$\hat{h}_j = \phi(w_h H(\hat{e}_i) + u_h (r_j \odot H(e_j^c)) + b_h) \quad (13)$$

$$h_j = (1 - z_j) \odot H(e_j^c) + z_j \hat{h}_j \quad (14)$$

where σ is a logistic sigmoid operation, ϕ is the activation function tanh, \mathbf{w} and \mathbf{u} are learnable weight matrices, and h_j is the encoded representation after the source event i attended to the context event j .

Context Gating We adopt the gate in (Tu et al., 2018) to regulate the source $H(e_i^l)$ and context information h_j . Then we get the context-enhanced source embedding for further decoding.

$$\gamma = \sigma(w_j h_p + w_k h_f) \quad (15)$$

$$h_c = \gamma h_p + (1 - \gamma) h_f \quad (16)$$

$$\lambda = \sigma(w_c h_c + w_s H(e_i^l)) \quad (17)$$

$$H = \lambda h_c + (1 - \lambda) H(e_i^l) \quad (18)$$

where h_c is the integration of all previous context h_p and future context h_f . The w_j , w_k , w_c and w_s are learnable parameter matrices, and H is the final representation.

3.2.3 Decoding and Loss

The decoder is an auto-regressive transformer model to generate tokens one by one. We adopt the cross-entropy loss to minimize the negative log-likelihood over ground-truth words and apply the label smoothing strategy.

$$\mathcal{L} = - \sum_{i=1}^m \log P(y_i | e_i, v_{\neq i}, e_{\neq i}) \quad (19)$$

4 Experiment

4.1 Dataset and evaluation metrics

We run our experiments on both Youcook2 dataset (Zhou et al., 2018a) and ActivityNet Caption dataset (Krishna et al., 2017). YouCook2 is the task-oriented instructional video dataset for video procedural captioning on the recipe domain. We follow the data partition in VideoBERT (Sun et al., 2019b) which uses 457 videos in the YouCook2 validation set as the testing set and the rest for development. In all, we use 1,278 videos for training and validation. We extract the visual feature by S3D model pre-trained on Howto100M (Miech et al., 2019) dataset through MIL-NCE (Miech et al., 2020) model. This visual representation is a better representation of Howto videos. The ASR transcript is automatically extracted from the off-the-shelf recognition tool¹.

Different from the Youcook2 dataset, Activitynet captions are open-domain videos with overlapping proposals, while Youcook2 has non-overlapping event proposals. We apply the same data partition in (Iashin and Rahtu, 2020b) with the ground truth labels. We directly download the copy of the dataset in (Iashin and Rahtu, 2020b) which contains 9,167 (out of 10,009) training and 4,483 (out of 4,917) validation videos. The dataset only contains partially available videos (91%) due to no longer available Youtube links. To make a fair comparison, we only list the experimental results on the same dataset. This open-source code and data portal contains the speech content extracted from the closed captions (CC) from the YouTube ASR system.

We employ the metrics BLEU3, BLEU4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin and Och, 2004) and CIDEr (Vedantam et al., 2015) to evaluate the performance. We follow the work in (Iashin and Rahtu, 2020a) on ActivityNet caption dataset which reported BLEU3, BLEU4 and METEOR. We directly apply the open-source tool² to evaluate our results as in (Krishna et al., 2017).

¹<https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/>

²https://github.com/ranjaykrishna/densevid_eval/tree/9d4045aced3d827834a5d2da3c9f0692e3f33c1c

Methods	V/T	B-3	B-4	M	R-L	CIDEr
Bi-LSTM + TempoAttn (Shou et al., 2016)	V	-	0.87	8.15	-	-
EMT(Zhou et al., 2018a)	V	-	4.38	11.55	27.44	38
VideoBERT(Sun et al., 2019b)	V	6.80	4.04	11.01	27.50	49
VideoBERT (+S3D feature)(Sun et al., 2019b)	V	7.59	4.33	11.94	28.80	55
CBT(Sun et al., 2019a)	V	-	5.12	12.97	30.44	64
DPC	VT	7.60	2.76	18.08	-	-
AT+video(Hessel et al., 2019)	VT	-	9.01	17.77	36.65	112
Transformer (w/o context)	V	12.79	6.35	16.56	37.17	113
HCN	V	13.74	7.26	17.11	38.35	121
Transformer (w/o context)	VT	15.00	7.10	18.07	38.31	123
HCN	VT	15.72	9.01	19.51	41.03	141

Table 1: The dense video event captioning results on the Youcook2 dataset, and these results are based on the validation set. The column "V/T" means whether the results come from uni-modal or multi-modal features. Transformer(w/o context) is the base method similar to (Hessel et al., 2019).

Methods	V/T	B-3	B-4	M
WLT (Rahman et al., 2019)	V	3.04	1.46	7.23
MDVC (Iashin and Rahtu, 2020b)	VT	4.52	1.81	10.09
BMT (Iashin and Rahtu, 2020a)	VT	4.63	1.99	10.90
Transformer (w/o context)	V	4.44	1.83	9.93
HCN	V	5.54	2.48	10.90
Transformer (w/o context)	VT	4.43	1.86	10.05
HCN	VT	5.82	2.62	10.64

Table 2: The dense video event captioning results on the Activitynet Captions dataset and these results are based on the ground truth proposals of the validation set. The column "V/T" means whether the results come from uni-modal or multi-modal features.

4.2 Implementation details

We develop our model based on the open-source code ³ of MDVC(Iashin and Rahtu, 2020b), and will release our code later. The embedding size of video, hidden size of the multi-head, and feed-forward layer are 1024, 512, and 128 respectively. The number of the head is 8 and the dropout rate is 0.4. We set the local context length k_l as 10, that is, the 10 previous and 10 future frames as a local event-agnostic context, and one previous event and one next event as a global event-aware context for a trade-off between performance and efficiency. We adopt the Adam optimizer (Kingma and Ba, 2015) with learning rate of $1e-4$, and set two momentum parameters $\beta_1=0.9$ and $\beta_2=0.98$. For label smoothing, and the smoothing rate is 0.4. We set the batch size to 128. For model complexity, the HCN model introduces only 3% more parameters to the base model. All models are trained on 1 Tesla P100 GPUs for 4 hours for Youcook2 and 30 hours for Activitynet Captions.

Video features We sampled frames at 16 fps and took the feature activations before the final

³<https://github.com/v-iashin/MDVC>

linear classifier of the S3D backbone and applied 3D average pooling to obtain a 1024-dimension feature vector. We got 1 feature per second and set k to 80.

4.3 Compare with State-of-the-art results

We demonstrate the results of our context-aware model on the Youcook2 dataset in Table 3. There are several existing baseline models: (1) Bi-LSTM with Temporal Attention (Bi-LSTM + TempoAttn) (Shou et al., 2016), which adopts Bi-LSTM language encoder; (2) End-to-End Masked Transformer (EMT) (Zhou et al., 2018b), an transformer based model; (3,4) VideoBERT (Sun et al., 2019b) and Contrastive Bidirectional Transformer (CBT) (Sun et al., 2019a), the pre-training based methods; (5) AT+Video (Hessel et al., 2019), the multi-modal transformer method. Besides the work (Shou et al., 2016) using a recurrent network, other baseline methods adopted the transformer model. Our context-aware model achieves the best results for uni-modal video event captioning and outperforms the context-agnostic base model by a large margin. Furthermore, our HCN model with multi-modal inputs can achieve comparable results with state-of-the-art results.

We list experimental results on a partial dataset of ActivityNet Captions as (Iashin and Rahtu, 2020b) and ignore others on the full dataset as (Krishna et al., 2017) to make a fair comparison. Table 2 presents the results of baseline methods and HCN. There are several baseline methods: (1) WLT (Rahman et al., 2019), a weakly supervised method with multi-modal input; (2) multi-modal video event captioning (MDVC) (Iashin and Rahtu, 2020b), a transformer-based model with multi-modal inputs; (3) BMT (Iashin and Rahtu, 2020a), a better use of

visual-audio information. Among these methods, WLT encoded the context using a recurrent network, while others are transformer models. HCN outperforms the base context-agnostic methods to a large extent and achieves state-of-the-art results.

From both experimental results, we can see that our methods with context-aware information can improve the base context-agnostic model by a large margin for both unimodal or multi-modal input.

4.4 Ablation Study

Methods	B-4	M	R-L	CIDEr
HCN	7.26	17.11	38.35	121.41
- type embedding	6.95	17.02	38.02	122.12
- future event	6.82	16.69	37.23	118.71
- past event	6.43	16.65	37.25	116.97
- GRU gate	6.50	16.71	37.86	119.16
- global context	6.94	17.10	37.68	121.06
- local context	7.17	17.03	37.93	119.87
Base (w/o context)	6.35	16.56	37.17	113.34

Table 3: Ablation study on the Youcook2 dataset. ‘-’ means to remove the setting from the full HCN model.

We introduce the ablation study of the HCN model on the Youcook2 dataset. In our experiment, we use uni-modal input and illustrate the ablation results in Table 3. We remove one component at a time from the full HCN model to compare the performance. *Type embedding*: we remove the type embedding which is used to distinguish whether the input is source or context event. From the results, we can observe the performance drop by removing the type embedding. *Past/Future context*: we investigate the model with the only past context or future context and found that both past and future contexts are effective and complementary with each other. The model with the context in both directions achieves the best result. *Cross attention gate*: The GRU gate in the cross attention model is more effective than the simple combination, which shows that the GRU gate is better for modeling a sequential context. *Local/global context*: From the results in Table 3, we can see that the global context is more effective than the local context. The HCN model with both contexts outperforms all the models. *Context length*. 1) With regards to the local context, the results of 10 or 20 context frames are similar with CIDEr as 141.1 and 141.3 correspondingly, while the performance with 40 frames drops with CIDEr as 138. 2) For the global context, we have increased the number of previous and next events as the global context, but there is no further

improvement. We found that irrelevant events even bring noise or duplicated information to learn.

4.5 Qualitative Analysis

We analyzed several cases and found two interesting videos shown in Figure 4 and 5. We depict the visual thumbnail, ground-truth caption, predicted results of our baseline and HCN methods.





Video				
ASR	I have like a whole pork	give it a nice whack	The pork, turn it around turn around	ok I'll put it in
GT		add a piece of pork in a ziplock bag and pound it	coat the pork in flour egg and breadcrumbs	place the pork onto a hot pan of oil
Base		coat the chicken in the bag	coat the pork in the flour mixture	fry the chicken in the pan
HCE		marinate the pork with flour	coat the pork in flour egg and breadcrumbs	fry the meat in the pan

Figure 4: In this case, it is hard to distinguish the fine-grained object ‘‘chicken’’ or ‘‘pork’’ from both visual and the transcript (co-reference ‘‘it’’). The baseline method would like to predict ‘‘chicken’’ with a prior bias for the ambiguous object leading to inconsistent captions between events. Modeling event dependency can make coherent captions. Besides, as shown in event 1, our HCN model can leverage local context to learn the entity ‘‘pork’’ from previous frames.


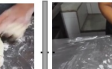
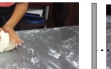
Video			
ASR	never put to it later	this and like this for about 5 minutes	and i'll pass two three four
GT	add salt to the dough	knead the dough	cut the dough into pieces
Base	knead the dough	knead the dough	knead the dough
HCE	add some grated cheese to the dough	knead the dough	cut the dough into pieces and roll into pieces

Figure 5: In this case, the baseline context-agnostic model would like to generate the same captions given similar visual inputs. With event-aware context, our HCN model can sequentially generate reasonable sentence sequences and reduce redundancy.

From the case in Figure 4, we can see that the baseline context-agnostic model generates the caption of each event solely leading to inconsistent captions. The baseline model predicts the ambiguous object as ‘‘chicken’’ for event 1 with prior bias, but output the object as ‘‘pork’’ for event 2. Our HCN model can tackle this issue and is prone to predict captions with a consistent object in the procedure. Besides, as shown in event 1, the entity ‘‘pork’’ can also be learned from previous frames. The context-aware model is effective in resolving entity ambiguity and generating coherent captions.

The case in Figure 5 presents another challenge. Since the visual cue of the three events is very similar, the base context-agnostic model inevitably

predicts the same caption as "knead the dough". The HCN model can learn the prior dependency between events, and hinder generating redundant sentences for similar events in the video. Therefore, the HCN model can generate the correct sentence for event 3. However, although the model tries to predict different captions for event 1, it is still hard to recognize the fine-grained entity "salt" from the video, and all models predict the object by mistake. Fine-grained entity recognition from a video is still a challenging problem.

To sum up, from these cases we can see that, 1) the neighboring context can provide extra information to make an accurate and coherent prediction. 2) the HCN model can capture the temporal dependency between events as prior knowledge, and generate consistent and less duplicate captions between events. 3) fine-grained object recognition from a video is still a challenging problem. Visual coreference resolution (Kottur et al., 2018) can be the future work to tackle this problem.

5 Related Work

Video Captioning The tasks mainly contain three types of captioning: single-sentence captioning (Xu et al., 2016; Wang et al., 2018b; Zhang et al., 2018), paragraph-level captioning (Yu et al., 2016; Lei et al., 2020; Ging et al., 2020) and event-level captioning (Krishna et al., 2017; Li et al., 2018; Wang et al., 2018a; Mun et al., 2019; Chen et al., 2019; Zhou et al., 2018b). The difference between these tasks is whether to generate one or multiple sentences for the whole video or each separate event of the video. In this paper, we focus on the more challenging dense event-level video captioning task to generate descriptions for each event. Previous works (Krishna et al., 2017; Li et al., 2018; Wang et al., 2018a) mainly exploited recurrent neural models such as long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) or recurrent unit (GRU) (Cho et al., 2014) to encode context. However, the recurrent model suffers from modeling long dependency effectively. Zhou et al. (Zhang et al., 2018; Sun et al., 2019b,a) introduced a self-attention model (Vaswani et al., 2017) which generates the caption based on the clip of each event solely. Compared with these works, we are the first to implement a novel video-level hierarchical context-aware network for dense video event captioning.

Multi-modal Video Captioning Video natu-

rally has multi-modal inputs including visual, speech text, and audio. Previous works explore visual RGB, motion, optical flow features, audio features (Hori et al., 2017; Wang et al., 2018b; Rahman et al., 2019) as well as speech text features (Shi et al., 2019; Hessel et al., 2019; Iashin and Rahtu, 2020b) for captioning. According to the work in (Shi et al., 2019; Hessel et al., 2019; Iashin and Rahtu, 2020b), although the speech text is noisy and informal, it can still capture better semantic features and improve performance especially for instructional videos. Later on, Lashin et al. (Iashin and Rahtu, 2020b) proposed to embed all visual, audio, and speech text for dense video event captioning. However, context-aware models are rarely investigated in multi-modal video event captioning. Therefore, we propose a novel attention model for effectively encoding the local and global context to tackle ambiguous object recognition and transcript co-reference through jointly modeling multi-modal inputs.

Context-aware Language Generation Our work is inspired by context-aware language generation e.g. document-level neural machine translation (NMT) (Miculicich et al., 2018; Maruf et al., 2019; Ma et al., 2020). Miculicich et al. (Miculicich et al., 2018) adopted a hierarchical context-aware network in a structured and dynamic manner. Maruf et al. (Maruf et al., 2019) and Ma (Ma et al., 2020) further explored a scalable and effective attention mechanism. For the local neighbor-clip and global inter-event context, we further design a hierarchical context-aware network with a hybrid mechanism of multi-modal video captioning to dynamically leverage various video-level information through a gating scalar.

6 Conclusion and Discussion

Dense video event captioning is a typical video understanding task to learn procedural events in a long untrimmed video. It is essential to model holistic video information for event understanding. In this paper, we propose a novel hierarchical context-aware network to encode both the local and global context of long videos. Our HCN model is effective in modeling context and outperforms the context-agnostic model by a large margin.

In future work, we tend to extend our hierarchical network to further investigate how to effectively attend to the long context to filter ambiguous and irrelevant information.

References

- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Shizhe Chen, Yuqing Song, Yida Zhao, Qin Jin, Zhaoyang Zeng, Bei Liu, Jianlong Fu, and Alexander Hauptmann. 2019. Activitynet 2019 task 3: Exploring contexts for dense captioning events in videos. *arXiv preprint arXiv:1907.05092*.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *arXiv preprint arXiv:2011.00597*.
- Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. 2019. A case study on combining asr and visual features for generating instructional video captions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202.
- Vladimir Iashin and Esa Rahtu. 2020a. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*.
- Vladimir Iashin and Esa Rahtu. 2020b. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614.
- Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.
- Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511.
- Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. *arXiv preprint arXiv:1903.08788*.
- Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.
- Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV*.
- Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6588–6597.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

- Tanzila Rahman, Bicheng Xu, and Leonid Sigal. 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8908–8917.
- Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. 2019. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6382–6391, Florence, Italy. Association for Computational Linguistics.
- Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058.
- Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. *Proceedings of the IEEE international conference on computer vision*.
- Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018a. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198.
- Xin Wang, Yuan-Fang Wang, and William Yang Wang. 2018b. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.
- Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 374–390.
- Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.