ICDT: Incremental Context Guided Deliberation Transformer for Image Captioning

No Author Given

No Institute Given

Abstract. Image Captioning is a task to generate descriptions for given images. Most encoder-decoder methods suffer from lacking the ability to correct the mistakes in predicted word. Though current deliberation motivated models can refine the generated text, they use single level image features throughout two stages. Due to the insufficient image information provided for the second-pass, deliberation action is ineffective in some cases. In this paper, we propose **I**ncremental **C**ontext Guided **D**eliberation **T**ransformer, namely **ICDT**, which consists of three modules, including: 1) Incremental Context Encoder, 2) Raw Caption Decoder and 3) Deliberation Decoder. Motivated by human writing habits in daily life, we treat the process of generating a caption as a deliberation procedure. The Raw Caption Decoder in first-pass constructs a draft sentence and then the Deliberation Decoder in second-pass polishes it to a better high-quality caption. In particular, for image encoding process, we design an Incremental Context Encoder that can provide cumulative encoded context based on different levels of image features for the deliberation procedure. Our encoder makes image features at different levels play specific roles in each decoding pass, instead of being simply fused and fed into the model for training. To validate the performance of the ICDT model, we evaluate it on the MSCOCO dataset. Compared with both Transformer-based models and deliberation-motivated models, our ICDT improves the state-of-the-art results and reaches 81.7% BLEU-1, 40.6% BLEU-4, 29.6% METEOR, 59.7% ROUGE and 134.6% CIDEr.

Keywords: Image Captioning · Deliberation Networks · Transformer.

1 Introduction

Image captioning task aims to generate a descriptive sentence for a given image, and its challenges lie not only in comprehensive image understanding but also in generating a sentence that matches the visual semantics of the image. The majority of proposed image captioning models following the encoder-decoder framework[25, 4, 13, 2, 32, 31, 10] has achieved promising progress on public datasets.

Despite the great success, such single-pass decoding process suffer from lacking the ability to correct the mistakes in predicted words. To overcome this limitation, deliberation motivated models are introduced to image captioning[30, 28, 7, 6, 8, 14] for better decoding. Motivated by human writing behaviours, deliberation models should firstly generate a rough caption of the image from a

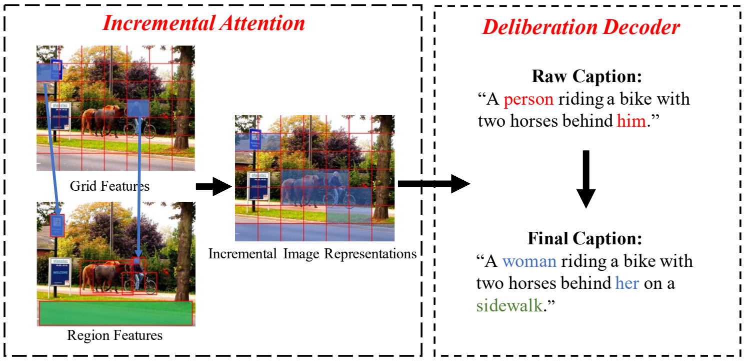

Incremental Attention
Grid Features
Region Features
Incremental Image Representations
Deliberation Decoder
Raw Caption:
“A person riding a bike with two horses behind him.”
Final Caption:
“A woman riding a bike with two horses behind her on a sidewalk.”

Fig. 1. An illustration of incremental context guided deliberation process. The left part shows how attention works to obtain incremental representation, and the right part summarizes the deliberation results. The blue arrow indicates attention operation and the blue mask denotes the attended area in original image.

global perspective, and then use the details to modify the rough caption. However, most of deliberated-based models are especially focus on text refinement, but use single level image features throughout two stages. These methods suffer from a drawback: the visual features from the first stage is insufficient for fixing the wrong words in the deliberation process. To utilize more diverse image features, Some works are proposed to fusion or interact of grid and region features to complement each other's advantages by using attention modules [10, 5, 27, 19, 18]. However, the direct use of two sources of features is prone to produce semantic noise. e.g. A grid containing a horse's leg may interact with the incorrect region containing a branch just because they have similar appearances. Therefore, merits of the two features should be leveraged separately with different functions instead of being used equally, and that can be well applied through two processes of deliberation.

To tackle the above problems and effectively combine two different stages, we try to design a method to utilize grid and region features in an incremental way to guide the deliberation procedure. As shown in Fig. 1, the grid level features attend to semantically related regions to get the incremental image representation. With the condition of grid attentive regions, the missing details of objects in the image can be captured, which guides the deliberation decoder to modify the word *person* to the correct detailed word *women*. Besides, unpredicted words in the raw caption like *sidewalk* can also be decoded by the incremental context. To this end, the introduction of the deliberation decoder and the rational use of the two features are organically combined, which inspires us to design a comprehensive end-to-end model.

In this paper, we propose Incremental Context Guided Deliberation Transformer, namely ICDT. As shown in Fig 2, ICDT is a two-pass decoding model, consisting of three modules:1) Incremental Context Encoder: encode grid level features as global first-pass context while adding local region level features to it as

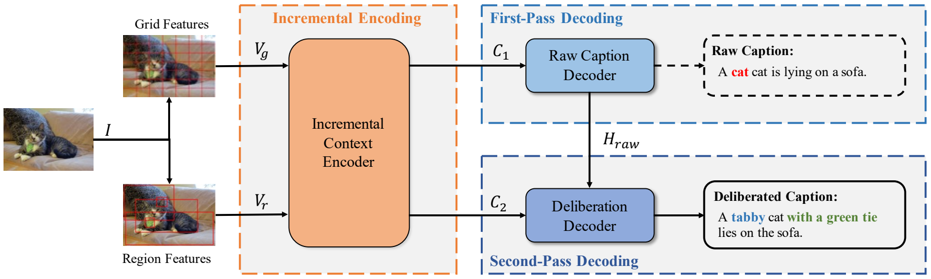


Fig. 2. Overview of our Incremental Context Guide Deliberation Transformer model.

incremental second-pass context, 2) Raw Caption Decoder: a non-autoregressive Transformer decoder use the global information provided by first-pass context to generate a raw caption, 3) Deliberation Decoder: polish the raw caption to a fine caption under the guidance of incremental second-pass context.

The major contributions of our paper can be summarized as follows:

- We propose a novel two-pass decoding model ICDT to achieve polishing generated sentence guided by two different level image features in an incremental way.
- We design an Incremental Context Encoder to obtain both global image features and incremental image features. With the Incremental Context Encoder, the second decoder of ICDT can be guided correctly to modify and detail the raw caption in the deliberation procedure.
- Experiments on MS-COCO dataset demonstrate that our model achieve new state-of-art performance for image captioning, *i.e.*, 134.6% CIDEr scores on *Karpathy*[12] test set.

2 Related Work

2.1 Image Encoding over Different Features

With the advantage of covering the information of the entire image without over-compressing the information, grid features were used in many image captioning models [21, 17, 29]. Compared with grid features, region features can provide object-level information of the image. By introduing region features[2, 9, 10, 5], the quantitative performance of image captioning was significantly improved. Nevertheless, the above works predicted sentences by using only one kind of features and lack full utilization of image information.

In order to integrate the advantages of both grid and region features, Wang et al.[26] proposed a hierarchical attention network to combine text, grid, and region features and explore the intrinsic relationship between different features. Luo et al.[18] proposed a cross-attention module with a graph to exploit complementary advantages of region and grid features.

2.2 Deliberation-motivated Methods

Motivated by human behaviour in the process of describing an image, deliberation aims to polish the existing caption results for further improvement. Wang et al. [30] proposed Review Net as a rudiment of the deliberation network for image captioning, which outputs a thought vector after each review step to capture the global properties in a compact vector representation. Sammani et al. [22] introduced a Modification Network to modify existing captions from a given framework by modeling the residual information. Latterly, [23] proposed a caption-editing model to perform iterative adaptive refinement of an existing caption. Related to ruminant decoder [8], [14] introduced a two-pass decoding framework, where a Cross Modification Attention is used to enhance the semantic expression of the image features and filter out error information from the draft caption to get better image captions. Although the above methods involve the deliberation process, they still focus only on the relationship between original image features and the draft caption, ignoring the effect of using different granularity image features throughout the process of generation.

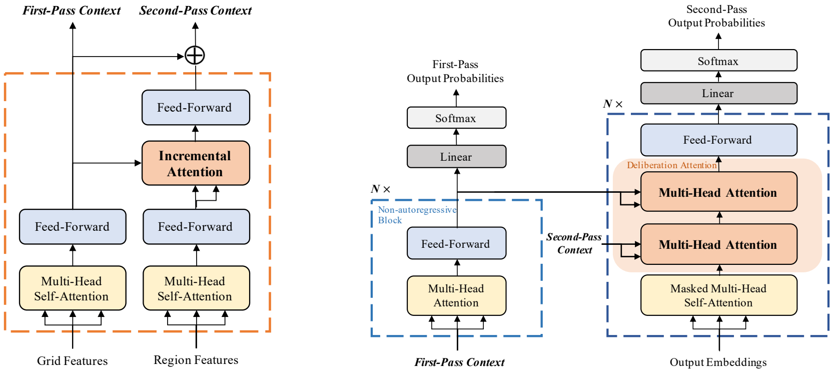
3 Methodology

3.1 Problem Statement

In order to obtain a precise caption, we define the image captioning task as generating a refined sentence based on a raw caption. Formally, give an image I, we first need to generate a sequence $Y^{*'} = \{y_1^{*'}, y_2^{*'}, ..., y_T^{*'}\}$, where $y_T^{*'} \in \mathcal{D}$ is the predicted word in the raw caption, $\mathcal{D}$ is the dictionary, and T denotes the sequence length. In the deliberation procedure, we polish the raw caption guided by the extra image information, and finally get a fine caption $Y^* = \{y_1^*, y_2^*, ..., y_T^*\}, y_T^* \in \mathcal{D}$.

3.2 Incremental Context Encoder

Efficient encoding of visual features of images is a prerequisite for generating high-quality captions. The deliberation-motivated methods usually encode single features like grids or regions, and then use the same encoded context when generating raw captions and final captions. This makes the deliberation process can not acquire additional information to modify the generated text and only focus on optimizing the language model. In this paper, we try to design an Incremental Context Encoder (ICE) that can provide incremental context for the two-pass decoder, so that it can provide extra information to guide the generation of final captions when deliberating raw captions. As shown in Fig. 3a, the grid features are encoded as first-pass context while integrating the region features through incremental attention as the second-pass context.



(a) Incremental Encoder

(b) Raw and Deliberation Decoder

Fig. 3. The architecture of Encoder and Decoder modules.

Basic Encoding The ICE employs a vanilla Transformer encoder module for basic encoding. Since Grid features can cover the full content of a given image to describe the global scenes, we utilize it as input for basic encoding to obtain a first-pass context for generating a raw caption. The input grid features are directly extracted from the RCNN model. Considering the positional information of grids, we introduce a learnable embedding layer and combine them:

$$V_g = E_g + E_{pos} \tag{1}$$

where E_{pos} indicates the positional embedding and E_g stands for the extracted encoding grid features.

After that, we feed the combined feature V_g to the Transformer encoder module. Each encoder layer contains two sub-layers, including a multi-head self-attention (MHA) layer and a feed-forward network (FFN) layer:

$$H_g^{'(l)} = MHA(H_g^{(l)}, H_g^{(l)}, H_g^{(l)}) \tag{2}$$

$$H_g^{(l+1)} = FFN(H_g^{'(l)}) \tag{3}$$

where $H_g^{(0)} = V_g$ and l is the number of encoder layer. The hidden states of grids $H_g^{(l)}$ are fed into the $(l{+}1)$-th MHA layer. Specifically, the FFN is a position-wise fully connected layer consisting two linear projections with a ReLU activation in between:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{4}$$

Through the Transformer encoder, we get the encoded hidden state from N-th layer H_g^N as the first-pass context C_1.

Incremental Attention Encoding Once the first-pass context was obtained through basic encoding, we need to add extra information to it to guide the deliberation pass. Although grids can provide information covering the entire image, it still lacks attention to the salient objects. So we take region features as object-level information to improve the capability of understanding the objects. In order to achieve the purpose of integrating region features on the basis of the grid, we design an Incremental Attention Encoding mechanism.

Same as basic encoding, we first feed the extracted region features to the MHA and FFN component to get the middle encoded context $C_r^{(l)}$:

$$H_r^{'(l)} = MHA(H_r^{(l)}, H_r^{(l)}, H_r^{(l)}) \tag{5}$$

$$C_r^{(l)} = FFN(H_r^{'(l)}) \tag{6}$$

where $H_r^{(0)} = V_r$, which denotes region feature vector extracted from object detection model.

Then we use the encoded first-pass context C_1 as the query matrix and $C_r^{(l)}$ as the key and value matrix. The scale-dot product between grids and regions stands for attentive relationships, which can be leveraged as a weight matrix applying to region features. For each grid, the weighted context contains region information at the corresponding location. The incremental attention can be stated recursively as follows:

$$H_{inc}^{'(l)} = \text{softmax}\left(\frac{(C_1)\left(C_r^{(l)}\right)^T}{\sqrt{d_{C_r^{(l)}}}}\right)C_r^{(l)} \tag{7}$$

After incremental attention, an FFN layer is also applied to $H_{inc}^{'}$:

$$H_{inc}^{(l)} = FFN(H_{inc}^{'(l)}) \tag{8}$$

Notice that the incremental attention encoding and basic encoding are computed in the same layer. Finally, we directly add N-th incremental encoded context H_{inc}^N to C_1 and get the second-pass context C_2:

$$C_2 = C_1 + H_{inc}^N \tag{9}$$

With the first-pass context C_1, the Raw Caption Decoder (RCD) generates a sequence of raw caption $Y^{*'} = \{y_1^{*'}, y_2^{*'}, ..., y_T^{*'}\}$, where T is the length of the raw caption. Different from other deliberation-motivated models, we use a non-autoregressive decoder as RCD. The non-autoregressive decoder enables parallel prediction during inference decoding. As shown in Fig 3b, the RCD removes the softmax layer during prediction, and directly use the vector output by the linear layer as the raw caption embedding feeding to the deliberation decoder.

Therefore, we remove the Masked Multi-head Self-attention layer of the vanilla Transformer decoder, and use the first-pass context from ICE directly to the MHA layer:

$$H^{'(l)}_{raw} = MHA(H_1^{(l)}, H_1^{(l)}, H_1^{(l)}) \tag{10}$$

where $H_1^{(0)} = C_1$. After that, A Feed-forward layer also is introduced:

$$H^{(l)}_{raw} = FFN(H^{'(l)}_{raw}) \tag{11}$$

And then we add a projection linear layer and a softmax layer for the training stage.

$$Y^{*'} = \text{softmax}(\text{proj}(H^{N}_{raw})) \tag{12}$$

Due to the non-autoregressive design, the RCD executes in parallel for both training and inference stages. However, this makes RCD unable to directly generate coherent sentences.

3.3 Deliberation Decoder

The deliberation decoder (DD) aims to polish the preliminary caption guided by the second-pass context. To achieve the deliberation procedure, we design a Transformer-like autoregressive decoder. Fig 3b illustrates the detailed structure of DD. As the same as vanilla Transformer, the output embedding of target sentence E_s is fed into a Masked Multi-head self-attention (MMHA) layer during the training state:

$$H^{'(l)}_{s} = MMHA(H_s^{(l)}, H_s^{(l)}, H_s^{(l)}) \tag{13}$$

where $H_s^{(0)} = E_s$. After that, DD incorporates the second-pass context which contains the attentive region features by grids. Since the ICE can leverage the extra region features to enhance the detailed information of objects, we use the Multi-head attention layer to stress the relationship between caption and attentive regions:

$$H^{'(l)}_{deli} = \text{softmax}\left(\frac{\left(H^{'(l)}_{s}\right)(C_2)^T}{\sqrt{d_{C_2}}}\right) C_2 \tag{14}$$

Then we add an additional multi-head attention layer and take the embedding of the raw caption as input:

$$H^{(l)}_{deli} = \text{softmax}\left(\frac{\left(H^{'(l)}_{deli}\right)(H_{raw})^T}{\sqrt{d_{H_{raw}}}}\right) H_{raw} \tag{15}$$

where $H^{(l)}_{deli}$ and H_{raw} are treated as query and key matrix respectively, which contributes to learning a weight for refining the raw caption. Notice that the projection layer of RCD and output embedding layer of DD shares the same parameter weights which are used to embed the vocabulary of the caption. And

H_{raw} is extracted from the projection layer of RCD to avoid the extra embedding layer breaking the end-to-end structure of ICDT.

Finally, the DD also uses a FNN layer before the projection linear layer:

$$Y^{*} = \operatorname{softmax}\left(FFN(H_{deli}^{N})\right) \tag{16}$$

3.4 Training details

Following standard practice of image captioning, we first calculate the cross-entropy loss for each decoder:

$$L_{\mathrm{XE}}^{i}(\theta) = -\sum_{t=0}^{T-1} \log\left(p_{\theta}\left(Y_{t}^{*} \mid Y_{0:t-1}^{*}, I\right)\right) \tag{17}$$

where Y_t^* is the ground-truth word, and θ is the parameter of i-th decoder. We obtain the overall learning objective by adding the losses of Raw Caption Decoder and Deliberation Decoder:

$$L_{\mathrm{XE}} = L_{\mathrm{XE}}^{raw} + L_{\mathrm{XE}}^{delib} \tag{18}$$

Following Cornia et al.[5], we also introduce reinforcement learning for further finetune to make up the difference between cross-entropy loss and evaluation metrics between cross-entropy loss and evaluation metrics. When training with reinforcement learning, we use the CIDEr-D score as a reward through Self-Critical Sequence Training (SCST)[21]. At prediction time, we simplify the Raw Caption Decoder as an inner decoder layer instead of generating sentences. After that, the Deliberation Decoder can generate the final caption directly through beam search, and the highest scored sequence has been kept as the best caption.

4 Experiments

4.1 Experimental Settings

Dataset and Evaluation Metrics. Microsoft COCO 2014 dataset[16] is the widely used benchmark for image captioning. Each image is annotated with 5 caption sentences. We follow the setting of Karpathy and Fei-Fei[12] for the offline evaluation, where 113,287 images are used for training, 5,000 images for validation and 5,000 images for testing. To evaluate the quality of generated captions, we use COCO caption evaluation tool[1] to calculate the standard evaluation metrics, including BLEU-1/4[20], METEOR[3], ROUGE[15], CIDEr[24] and SPICE[1]. All metrics can reflect the quality of the generated caption text from different aspects.

[1] https://github.com/tylin/coco-caption

Table 1. Comparison results on Transformer-based models. For fair comparison, all 'Grid Only' models takes the result based on features extracted from ResNext-101 backbone, and all 'Region Only' models use ResNet-101 backbone.

Feature	Model	BLEU-1	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Grid Only	AoA	80.7	39.0	28.9	58.7	129.5	22.6
	M^2	80.8	38.9	29.1	58.5	131.7	22.6
	X-Transformer	81.0	39.7	29.4	58.9	132.5	23.1
	RSTNet	81.1	39.3	29.4	58.8	133.3	23.0
Region only	ETA	81.5	39.9	28.9	59.0	127.6	22.6
	ORT	80.5	38.6	28.7	58.4	128.3	22.6
	CPTR	**81.7**	40.0	29.1	59.4	129.4	-
	AoA	80.2	38.9	29.2	58.8	129.8	22.4
	M^2 Transformer	80.8	39.1	29.2	58.6	131.2	22.6
	X-Transformer	80.9	39.7	29.5	59.1	132.8	**23.4**
	DRT	**81.7**	40.4	29.5	59.3	133.2	23.3
Grid and Region	I^2RT	80.9	39.2	29.3	58.9	130.9	22.9
	DLCT	81.4	39.8	29.5	59.1	133.8	23.0
	ICDT(Our model)	**81.7**	**40.6**	**29.6**	**59.7**	**134.6**	23.2

Implementation Details. Since our ICDT model needs to use both grid and region features, we adopt the same data preprocessing method as Luo et al.[18]. The pre-trained Faster R-CNN provided by Jiang et al[11] was used to extract features from both levels simultaneously. For extracting features, it removes the delation and uses a normal C5 layer to extract grid features. For grid features, an additional average-pool was applied to get 7×7 grid size vectors. Meanwhile, the 2048-d output vector from the first FC-layer of the detection head was used as region features.

In our implementation, we set the dimension of each layer in encoder and decoders to 512, the number of heads to 8. The number of layers for both encoder and decoder is set to 4. We set the dimension d_f of FFN to 2048. We employ dropout with keep probability 0.9 after each attention and feed-forward layer. In the XE pre-training stage, we warm up our model for 4 epochs with the learning rate linearly increased to $1e^{-4}$, and then decays by rate 0.8 every 3 epochs. When training with SCST, the learning rate starts from $5e^{-5}$ and decays by rate 0.1 every 50 epochs. We train all models using the Adam optimizer with momentum of 0.9 and 0.999, a batch size of 128. We use beam search with a beam size of 5 to generate captions when validating and testing.

4.2 Quantitative Analysis

Compared with Transformer-based Methods As shown in Table 1, we compare ICDT with other Transformer-based model for image captioning. Since ICDT considers both image grid and region granularity features, the models selected for comparison are divided into three groups, including:

Table 2. Comparison results on Deliberation-motivated models

Model	BLEU-1	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Review Net	-	29.0	23.7	-	88.6	-
Skeleton Key	74.2	33.6	26.8	55.2	107.3	19.6
Stack-Captioning	78.6	36.1	27.4	56.9	120.4	20.9
Deliberate Attention	79.9	37.5	28.5	58.2	125.6	22.3
Ruminant Decoding	80.5	38.6	28.7	58.7	128.3	22.3
ETN	80.6	39.2	-	58.9	128.9	22.6
CMA-DM	80.6	39.2	29.0	58.9	129.0	22.6
ICDT(Our model)	**81.7**	**40.6**	**29.6**	**59.7**	**134.6**	**23.2**

- Grid Only: The model only takes the grid features to generate the image caption, where RSTNet[33] is the original model, and AoA, M^2 and X-Transformer are the experimental models used to compare with region features in the original paper.
- Region Only: Models that only use region features, Because R-CNN model is the mainstream way of the region feature extraction in image captioning, all the baselines in this group take the official results of the original model.
- Grid and Region: Models that utilize both grid and region features at the same time,

From the results of the model on different evaluation metrics, our method fully surpasses the previous Transformer-based methods in terms of BLEU-1, BLEU-4, METOR, ROUGE and CIDEr. The CIDEr score of our DLCT reaches 134.6%, wich advances DLCT 0.8%. The boost of performance demonstrate the advantages of our ICDT which use incremental context encoder instead of cross fusion of region and grid features. In addition, according to the evaluation results, the model using two features achieves higher scores on CIDEr and SPICE metrics than the model using only single feature. In particular, compared with the Transformer-based SOTA model DLCT, our method achieves better performance in all indicators, reflecting the advantages of introducing a deliberation decoder. Next we will compare ICDT with all deliberation-motivated models.

Compared with Deliberation-motivated Methods. Table 2 summaries all models designed with deliberation actions. As shown in Table 2, our ICDT model consistently exhibits better performance than the others. Since all of the deliberation-motivated models use LSTM instead of Transformer, their feature encoding and sequence generation capabilities are not as good as our proposed Transformer-based Model. However, the deliberation idea still shows its capability on image captioning task and deserves to be generalized more widely.

4.3 Ablation Study

In order to verify the effectiveness of each module in ICDT, we design ablation experiments based on the vanilla Transformer. As shown in Table 3, we sepa-

Table 3. Performance comparison of Incremental Context Encoder (ICE) and Raw Caption Decoder & Deliberation Decoder (R&D) for grids (G) and regions (R). E+D denotes traditional encoder-decoder framework which is based on vanilla Transformer.

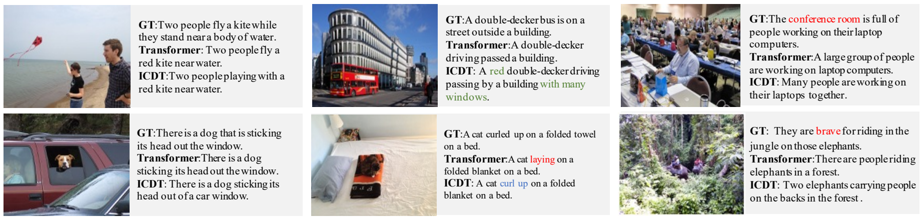
Feature	BLEU-1	BLEU-4	METOR	ROUGE	CIDEr	SPICE
E+D(G)	81.2	39.0	29.0	58.6	131.2	22.4
E+D(R)	80.1	39.0	28.9	58.6	130.1	22.4
E+D(G + R)	80.9	38.9	29.2	58.6	131.6	22.7
ICE+D(G + R)	81.1	39.3	29.5	58.9	132.8	22.9
E+R&D(G)	81.2	39.2	29.0	59.0	131.5	22.4
E+R&D(R)	80.8	39.1	29.1	58.9	130.1	22.3
E+R&D(G + R)	81.3	39.4	**29.8**	58.9	132.4	22.8
ICE+R&D(G + R)	**81.7**	**40.6**	29.6	**59.7**	**134.6**	**23.2**

rately use different visual features to validate the impact of Incremental Context Encoder. Further, all models are extended to two-pass decoders that we can evaluate the influence of Deliberation Decoder.

Impact of Incremental Context Encoder. To better understand the effect of our Incremental Context Encoder, we conduct four experiments on different features. The ICE+D model surpasses both single feature and fusion feature encoded models, which illustrates the effectiveness of Incremental Context Encoder. By integrating the attentive region feature and adding it to grid feature, the captioning model can better understand the corresponding region information and enrich the final encoded context. In sum, ICE+D outperforms E+D in most of the metrics and performs slightly worse in BLUE-1. We believe this is due to the fact that the Grid feature tends to highlight individual words rather than object entities in the raw image after self-attention.

Impact of Deliberation Decoder As shown in the lower part of Table 3, we also conduct several experiments to demonstrate the effectiveness of our Deliberation Decoder. After adding the deliberation decoder, the performance of experimental models can be further improved whether using the ordinary Transformer encoder or our proposed ICE. Specifically, the BLEU-4, ROUGE and CIDEr scores have the most significant improvements. The results also show that after the introduction of Deliberation Decoder, the fluency and correctness of the final generated caption can be significantly improved through the polishing.

In addition, we analyzed the experimental results of E+R&D(G+R) and ICE+R&D(G+R). ICE+R&D surpassed E+R&D by nearly 2% on the CIDEr metric. Owing to the ICE we designed can guide two decoding passes, although ordinary E+R&D can perform second-pass polishing, same encoded context from single encoder leads the difficulty to perform effective refinement on the raw caption generated in the first-pass. However, ICE+R&D adds incremental im-



GT:Two people fly a kite while they stand near a body of water.
Transformer: Two people fly a red kite near water.
ICDT:Two people playing with a red kite near water.

GT:A double-decker bus is on a street outside a building.
Transformer:A double-decker driving passed a building.
ICDT: A red double-decker driving passing by a building with many windows.

GT:The conference room is full of people working on their laptop computers.
Transformer:A large group of people are working on laptop computers.
ICDT: Many people are working on their laptops together.

GT:There is a dog that is sticking its head out the window.
Transformer:There is a dog sticking its head out the window.
ICDT: There is a dog sticking its head out of a car window.

GT:A cat curled up on a folded towel on a bed.
Transformer:A cat laying on a folded blanket on a bed.
ICDT: A cat curl up on a folded blanket on a bed.

GT: They are brave for riding in the jungle on those elephants.
Transformer:There are people riding elephants in a forest.
ICDT: Two elephants carrying people on the backs in the forest .

Fig. 4. Examples of image captioning results by vanilla Transformer and our ICDT with ground truth sentences.

age representations to deliberation decoder, which allows to obtain additional information to correct and polish the raw caption.

4.4 Qualitative Analysis

We show several example image captions generated by vanilla Transformer and ICDT in Fig. 4. In genegral, our ICDT can generate more detailed and correct captions. For two examples in the first column, both Transformer and ICDT can provide accurate descriptions. For examples in the middle column, we can see that our ICDT is able to capture more contextual information from the image to generate richer and more correct descriptions in some cases. The third column shows that both Transformer and ICDT fail to provide a high-quality caption which contains some specific information in the ground truth sentences. One possible reason is that human can get the information such as "*conference room*" and "*brave*" based on their background knowledge or associations about this scene, while Transformer and ICDT do not currently have such capabilities. This can propose a valuable direction for future research in image captioning.

5 Conclusion

In this paper, we propose a novel comprehensive two-pass decoding based model, Incremental Context Guided Deliberation Transformer (ICDT) for image captioning. In the first-pass a Raw Caption Decoder uses grid features alone to obtain a raw description for the image, and in the second-pass a Deliberation Decoder guided by rich image feature representations to polish the raw description to a high-quality caption . In order to cooperate with deliberation decoding procedure, we propose an Incremental Context Encoder to encode more accurate and detailed image information incrementally. As far as we know, ICDT is the only model that comprehensively considers different level features to guide the deliberation process. Results show that our approach outperforms the state-of-the-art methods.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic Propositional Image Caption Evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 382–398. Springer International Publishing, Cham (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6077–6086. IEEE, Salt Lake City, UT (Jun 2018)
3. Banerjee, S., Lavie, A.: 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. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (2005)
4. Cho, K., Courville, A., Bengio, Y.: Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. IEEE Transactions on Multimedia **17**(11), 1875–1886 (2015), conference Name: IEEE Transactions on Multimedia
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. pp. 10578–10587 (2020)
6. Gao, L., Fan, K., Song, J., Liu, X., Xu, X., Shen, H.T.: Deliberate Attention Networks for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 8320–8327 (Jul 2019), number: 01
7. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-Captioning: Coarse-to-Fine Learning for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018), number: 1
8. Guo, L., Liu, J., Lu, S., Lu, H.: Show, Tell, and Polish: Ruminant Decoding for Image Captioning. IEEE Transactions on Multimedia **22**(8), 2149–2162 (2020), conference Name: IEEE Transactions on Multimedia
9. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image Captioning: Transforming Objects into Words. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
10. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on Attention for Image Captioning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4633–4642. IEEE, Seoul, Korea (South) (Oct 2019)
11. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In Defense of Grid Features for Visual Question Answering. pp. 10267–10276 (2020)
12. Karpathy, A., Fei-Fei, L.: Deep Visual-Semantic Alignments for Generating Image Descriptions. pp. 3128–3137 (2015)
13. Li, L., Tang, S., Zhang, Y., Deng, L., Tian, Q.: GLA: Global–Local Attention for Image Description. IEEE Transactions on Multimedia **20**(3), 726–737 (Mar 2018), conference Name: IEEE Transactions on Multimedia
14. Lian, Z., Zhang, Y., Li, H., Wang, R., Hu, X.: Cross Modification Attention Based Deliberation Model for Image Captioning. arXiv:2109.08411 [cs] (Sep 2021), arXiv: 2109.08411
15. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla,

T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)

17. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3242–3250. IEEE, Honolulu, HI (Jul 2017)
18. Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.W., Ji, R.: Dual-level Collaborative Transformer for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence **35**(3), 2286–2293 (May 2021), number: 3
19. Pan, Y., Yao, T., Li, Y., Mei, T.: X-Linear Attention Networks for Image Captioning. pp. 10971–10980 (2020)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002)
21. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-Critical Sequence Training for Image Captioning. pp. 7008–7024 (2017)
22. Sammani, F., Elsayed, M.: Look and Modify: Modification Networks for Image Captioning. arXiv:1909.03169 [cs] (Mar 2020), arXiv: 1909.03169
23. Sammani, F., Melas-Kyriazi, L.: Show, Edit and Tell: A Framework for Editing Image Captions. pp. 4808–4816 (2020)
24. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-Based Image Description Evaluation. pp. 4566–4575 (2015)
25. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164. IEEE, Boston, MA, USA (Jun 2015)
26. Wang, W., Chen, Z., Hu, H.: Hierarchical Attention Network for Image Captioning. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 8957–8964 (Jul 2019), number: 01
27. Wang, Y., Zhang, W., Liu, Q., Zhang, Z., Gao, X., Sun, X.: Improving Intra- and Inter-Modality Visual Relation for Image Captioning. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4190–4198. Association for Computing Machinery, New York, NY, USA (2020)
28. Wang, Y., Lin, Z., Shen, X., Cohen, S., Cottrell, G.W.: Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. pp. 7272–7281 (2017)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 2048–2057. PMLR (Jun 2015), iSSN: 1938-7228
30. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R.: Review Networks for Caption Generation. In: Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)
31. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring Visual Relationship for Image Captioning. pp. 684–699 (2018)
32. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting Image Captioning With Attributes. pp. 4894–4902 (2017)
33. Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words. pp. 15465–15474 (2021)