CPTR : Full Transformer Network for Image Captioning

백인진 22.09.01

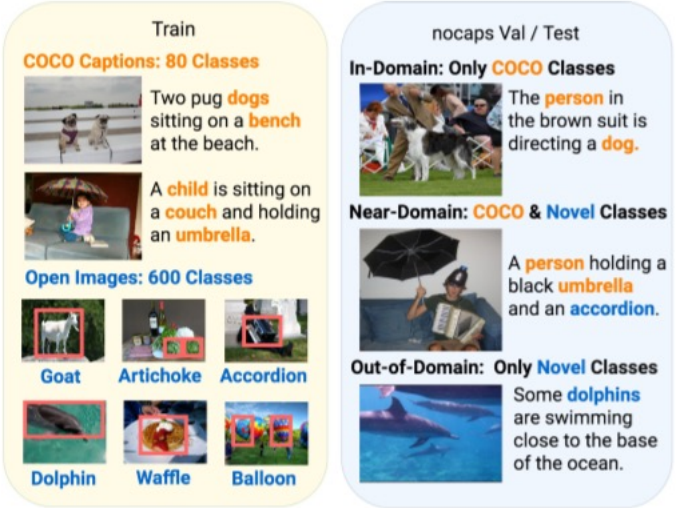
Introduction

Image Captioning

Edit

377 papers with code • 27 benchmarks • 49 datasets

Image Captioning is the task of describing the content of an image in words. This task lies at the intersection of computer vision and natural language processing. Most image captioning systems use an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence. The most popular benchmarks are nocaps and COCO, and models are typically evaluated according to a BLEU or CIDER metric.



CNN&RNN / CNN&Transformer

Decoder part has made a lot of progress
But, **Encoder** part always remains unchanged

Motivation

- Past CNN encoder
  - Global context modeling which can only be fulfilled by enlarging receptive field gradually as the convolution layers go deeper
- CPTR Encoder ; transformer
  - Encoder can utilize long-range dependencies among the sequentialized patches from the very beginning via self-attention mechanism
- CPTR Decoder ; transformer
  - ‘words-to-patches’ attention in the cross attention layer of decoder which is proved to be effective

Framework

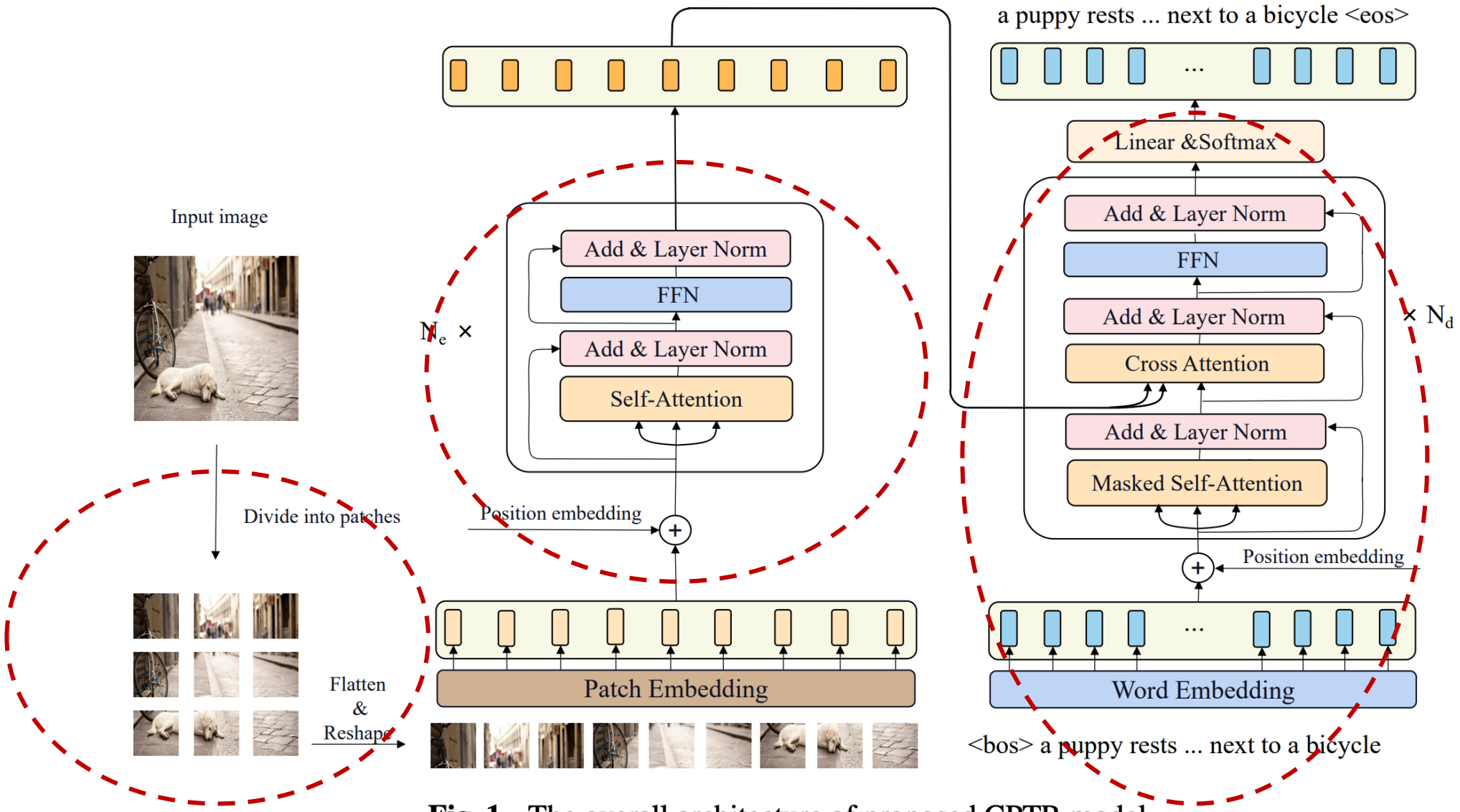
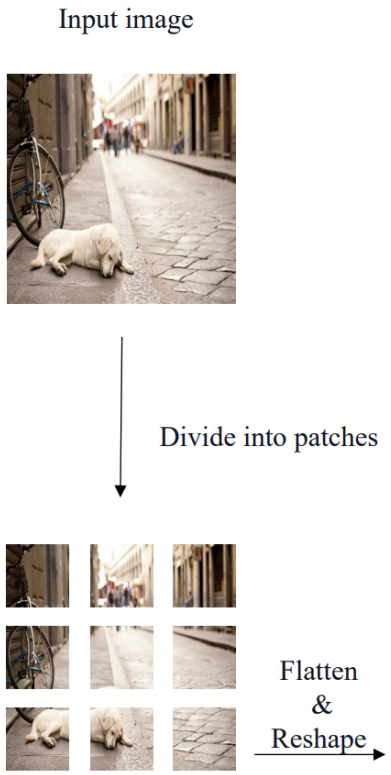


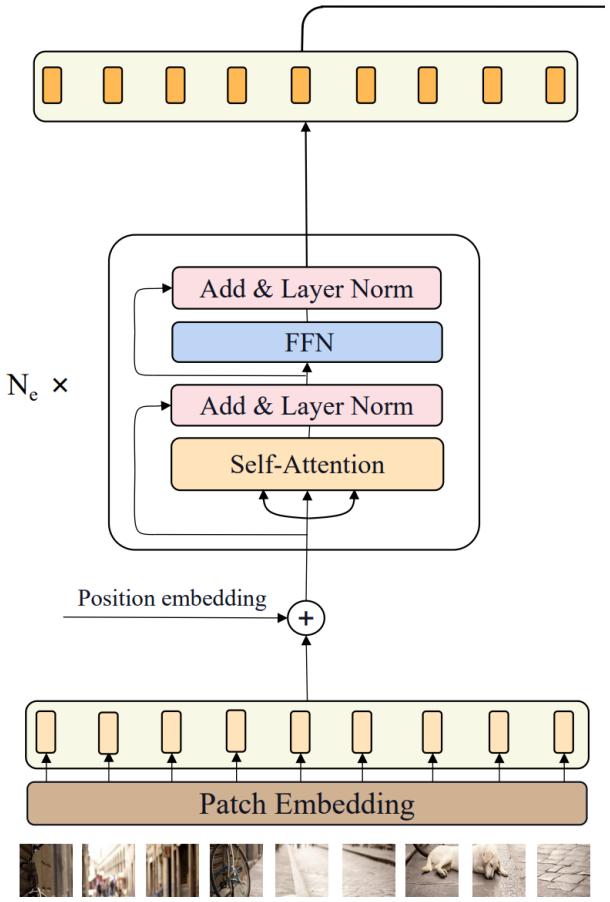
Fig. 1. The overall architecture of proposed CPTR model.

Framework



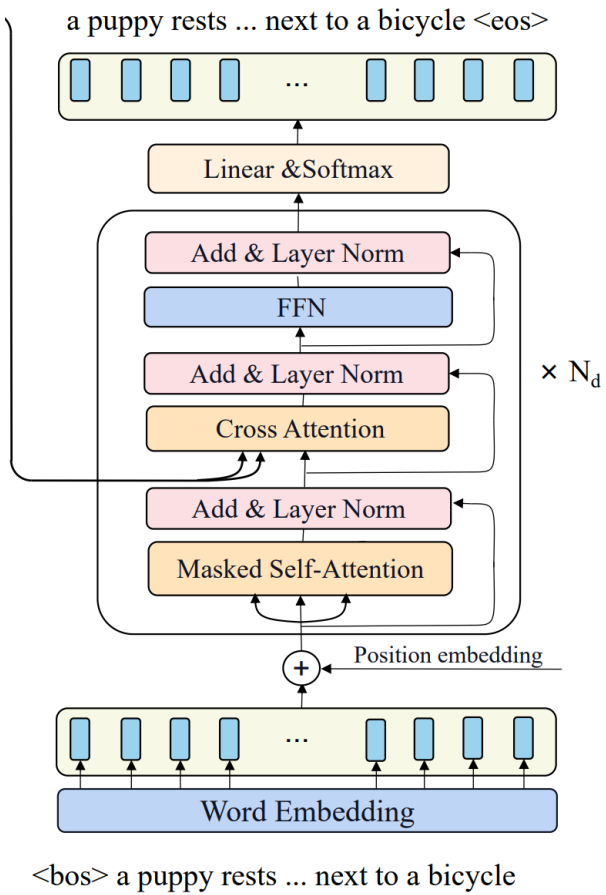
- **Input**
  - Resize the input image into a fixed resolution,
  - Divide the resized image into N patches
  - Reshape them into a 1D patch sequence
  - Add a learnable 1D position embedding to the patch features

Framework



- **Encoder**
  - Consists of N_e stacked layers, each of which consists of a multi-head self-attention sublayer followed by a positional feed-forward sub-layer
    - $MHA(Q,K,V) = Concat(h_1, \dots, h_H)W^O$
    - $Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
    - $FFN(x) = FC_2\left(Dropout\left(GELU\left(FC_1(x)\right)\right)\right)$
    - $x_{out} = LayerNorm(x^{in} + Sublayer(x^{in}))$

Framework



- **Decoder**
  - Add sinusoid positional embedding
  - Consists of N_d stacked identical layers with each layer containing a masked multi-head self-attention sublayer followed by a multi-head cross attention sublayer and a positionally feed-forward sublayer sequentially
  - The output feature of the last decoder layer is utilized to predict next word via a linear layer whose output dimension equals to the vocab size

Training

- Loss
  - To minimize cross entropy loss
  - $L_{XE}(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta(y_t^* | y_{1:t-1}^*)\right)$
- Finetuning
  - Use self-critical training
  - Reinforcement learning method by giving the difference in CIDEr-D score between the actual caption and the generated caption as a reward

Result

- MSCOCO online test server

Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
CNN+RNN														
SCST [9]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.0
LSTM-A [10]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [4]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RF-Net [11]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [12]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [13]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
CNN+Transformer														
ETA [14]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
CPTR	**81.8**	95.0	**66.5**	89.4	**51.8**	**80.9**	**39.5**	**70.8**	**29.1**	**38.3**	**59.2**	**74.4**	**125.4**	**127.3**

- COCO Karpathy

Method	B-1	B-2	B-3	B-4	M	R	C
CNN+RNN							
LSTM [18]	-	-	-	31.9	25.5	54.3	106.3
SCST [9]	-	-	-	34.2	26.7	55.7	114.0
LSTM-A [10]	78.6	-	-	35.5	27.3	56.8	118.3
RFNet [11]	79.1	63.1	48.4	36.5	27.7	57.3	121.9
Up-Down [4]	79.8	-	-	36.3	27.7	56.9	120.1
GCN-LSTM [12]	80.5	-	-	38.2	28.5	58.3	127.6
LBPF [19]	80.5	-	-	38.3	28.5	58.4	127.6
SGAE [13]	80.8	-	-	38.4	28.4	58.6	127.8
CNN+Transformer							
ORT [20]	80.5	-	-	38.6	28.7	58.4	128.3
ETA [14]	81.5	-	-	39.3	28.8	58.9	126.6
CPTR	**81.7**	**66.6**	**52.2**	**40.0**	**29.1**	**59.4**	**129.4**

- The superiority of CPTR model over conventional CNN+ architecture to the capacity of modeling global context at all encoder layers

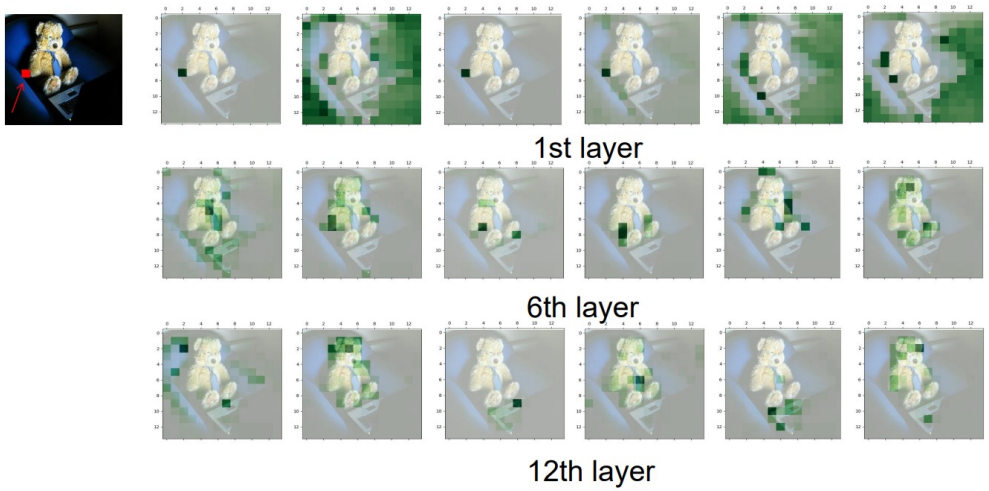
Ablation study

Pretrained Model	Res	#Layer	Dim	B-4	M	R	C
From scratch	224	4	768	16.5	17.3	42.1	45.5
ViT21K	224	4	768	33.5	27.4	55.8	110.6
ViT21K+2012	224	4	768	33.8	27.4	55.8	111.6
ViT21K+2012	224	1	768	33.8	27.5	56.1	111.2
ViT21K+2012	224	2	768	33.4	27.5	56.0	110.9
ViT21K+2012	224	6	768	33.7	27.5	55.9	110.9
ViT21K+2012	224	4	512	34.0	27.4	56.0	111.0
ViT21K+2012	384	4	768	**34.9**	**28.2**	**56.9**	**116.5**

- **Pretraining** vitals for CPTR model
- CPTR is little sensitive to the decoder hyperparameter including the number of layers and feature dimension
- **Increasing input image resolution** from 224*224 to 384*384 while maintaining the patch size equals to 16 can bring huge performance gains
  - the length of patch sequence increases from 196 to 576 due to the increasing input resolution
  - can divide image more specifically and provide more features to interact with each other via the encoder self-attention layer.

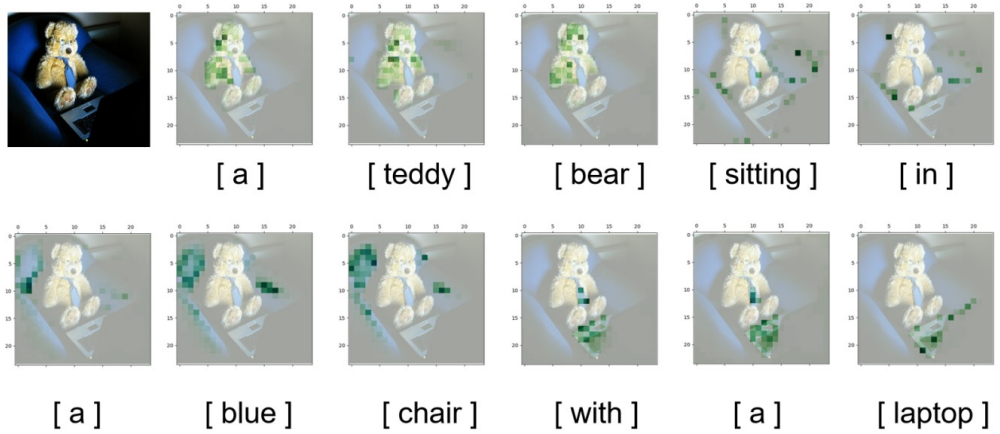
Attention Visualization

Predicted encoder self-attention



- Shallow layer
  - both the local and global contexts are exploited by different attention heads thanks to the full Transformer design
- Middle layer
  - tends to pay attention to the primary object
- Last layer
  - fully utilizes global context and pays attention to all objects in the image

'words-to-patches' cross attention



- Last layer
  - Weights in the decoder during the caption generation process
  - Correctly attend to appropriate image patches when predicting every word

Conclusion

- **Contributions**
  - CPTR is totally convolution-free and possessed the capacity of modeling global context information
  - CPTR can exploit long range dependencies from the beginning
  - The decoder ‘words-to-patches’ attention can precisely attend the corresponding visual patches to predict words

References

- CPTR : Full Transformer Network for Image Captioning

https://arxiv.org/abs/2101.10804