End-to-End Dense Video Captioning with Masked Transformer

Luowei Zhou*
University of Michigan
luozhou@umich.edu

Yingbo Zhou*
Salesforce Research
yingbo.zhou@salesforce.com

Jason J. Corso
University of Michigan
jjcorso@eecs.umich.edu

Richard Socher
Salesforce Research
richard@socher.org

Caiming Xiong†
Salesforce Research
cxiong@salesforce.com

*in CVPR 2018

Motivation

- Dense Video Caption: event detection + event description.
- Previous models: build two models and train separately.
- Language information cannot have direct impacts on event proposal.
- This paper: end-to-end training, use language to help localization.
  - A masking network to convert discrete event proposals to differentiable mask, ensuring the consistency between proposal and caption.
  - Self-attention mechanism for capturing long-term dependencies.

Method

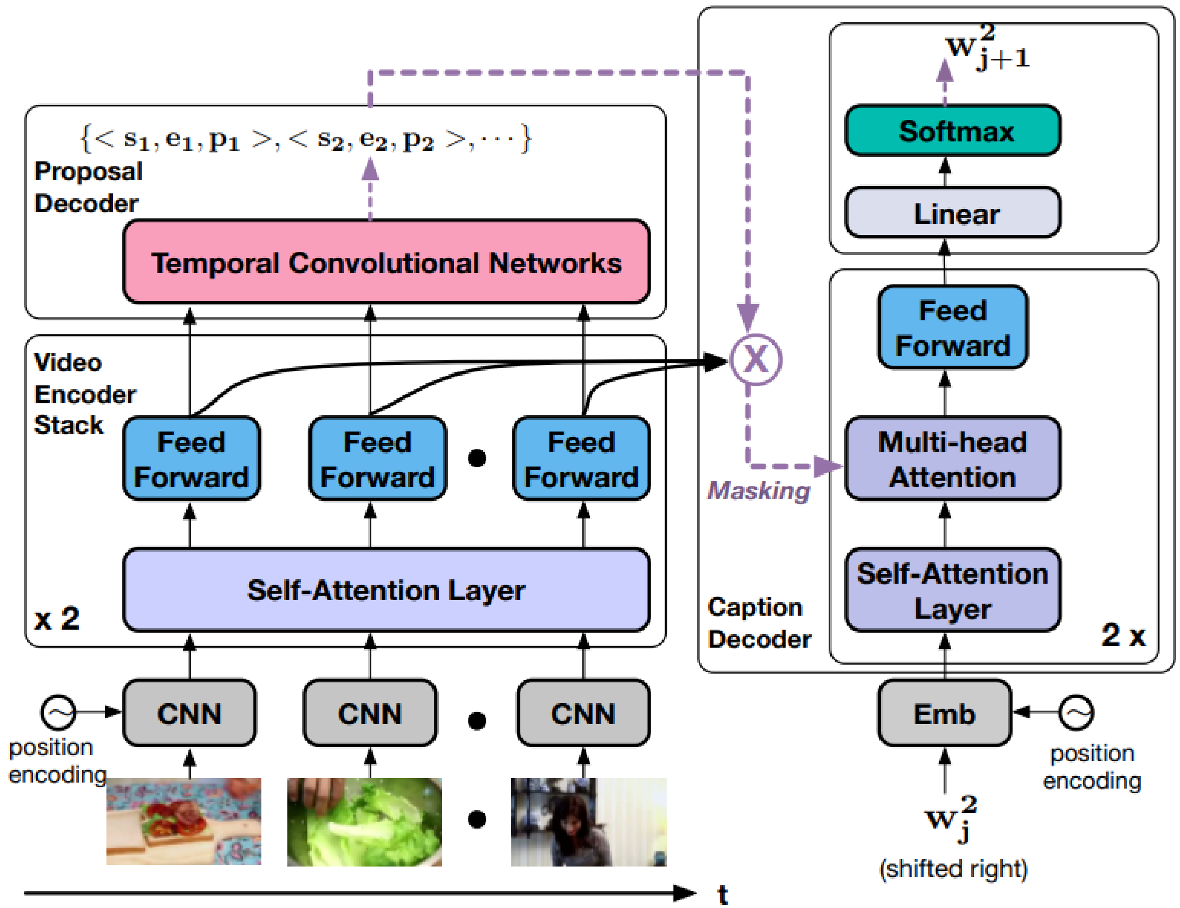


Figure 1. Dense video captioning is to localize (temporal) events from a video, which are then described with natural language sentences. We leverage temporal convolutional networks and self-attention mechanisms for precise event proposal generation and captioning.

Method

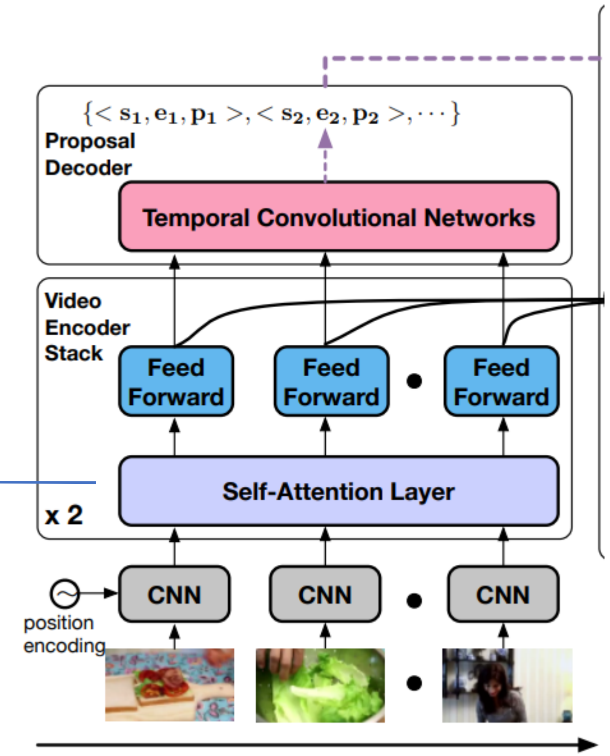
- Video Encoder: a video $X = \{x_1, \ldots, x_T\}$ -> *F*

$$\mathrm{V}(F^l) = \Psi(\mathrm{PF}(\Gamma(F^l)), \Gamma(F^l))$$

$$\Gamma(F^l) = \begin{pmatrix} \Psi(\mathrm{MA}(f_1^l, F^l, F^l), f_1^l)^\top \\ \cdots \\ \Psi(\mathrm{MA}(f_T^l, F^l, F^l), f_T^l)^\top \end{pmatrix}^\top$$

$$\Psi(\alpha, \beta) = \mathrm{LayerNorm}(\alpha + \beta)$$

$$\mathrm{PF}(\gamma) = M_2^l \max(0, M_1^l \gamma + b_1^l) + b_2^l$$

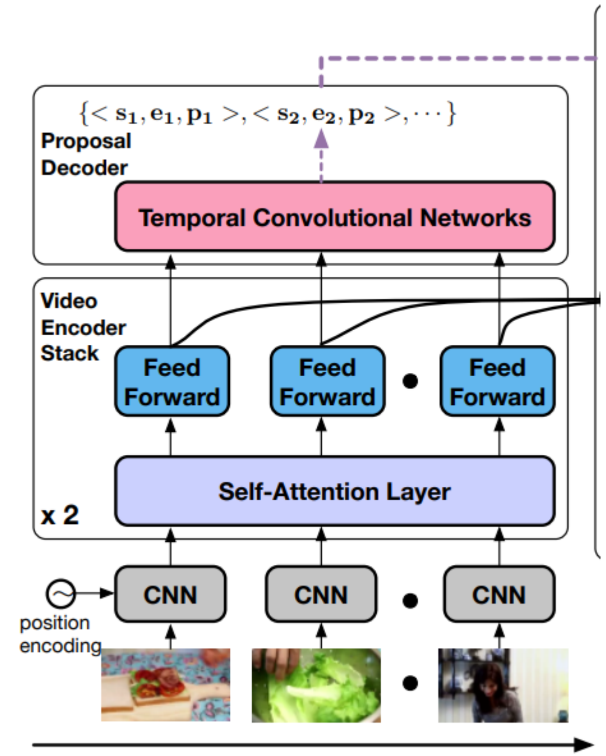


where $\Psi(\cdot)$ represents the function that performs layer normalization on the residual output, $\mathrm{PF}(\cdot)$ denotes the 2-layered feed-forward neural network with ReLU nonlinearity for the first layer, M_1^l, M_2^l are the weights for the feed-

Method

- Proposal Decoder: N explicit anchors.
  - Based on ProcNets*, mainly temporal conv nets.
  - Each proposal is composed of score P_e and boundaries (S_p, E_p)

$$c_p = c_a + \theta_c l_a \quad l_p = l_a \exp\{\theta_l\},$$
$$S_p = c_p - l_p/2 \quad E_p = c_p + l_p/2.$$



*: Towards Automatic Learning of Procedures from Web Instructional Videos, in AAAI 2018.

Method

- Captioning Decoder: masked transformer.

$$Y_{\leq t}^{l+1} = \mathrm{C}(Y_{\leq t}^{l}) = \Psi(\mathrm{PF}(\Phi(Y_{\leq t}^{l})), \Phi(Y_{\leq t}^{l}))$$

$$\Phi(Y_{\leq t}^{l}) = \begin{pmatrix} \Psi(\mathrm{MA}(\Omega(Y_{\leq t}^{l})_1, \hat{F}^l, \hat{F}^l), \Omega(Y_{\leq t}^{l})_1) \\ \cdots \\ \Psi(\mathrm{MA}(\Omega(Y_{\leq t}^{l})_t, \hat{F}^l, \hat{F}^l), \Omega(Y_{\leq t}^{l})_t) \end{pmatrix}$$

Enc-Dec Attention

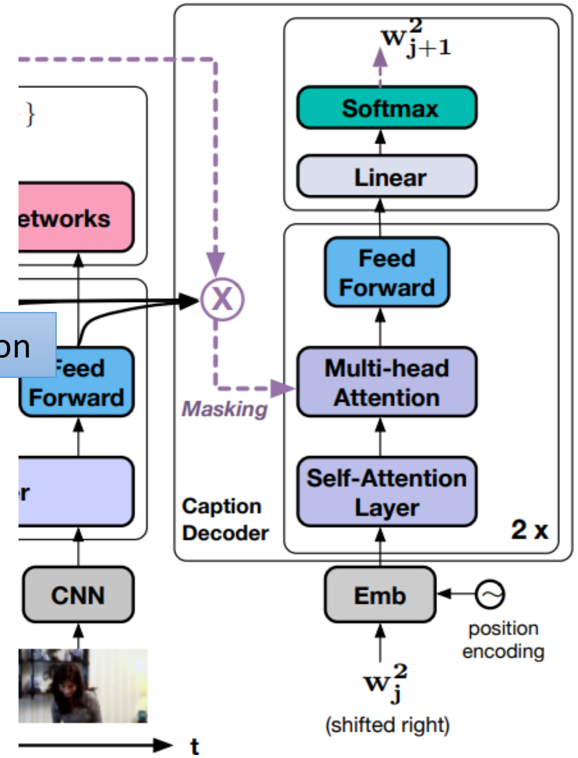
$$\Omega(Y_{\leq t}^{l}) = \begin{pmatrix} \Psi(\mathrm{MA}(y_1^l, Y^l, Y^l), y_1^l)^\top \\ \cdots \\ \Psi(\mathrm{MA}(y_t^l, Y^l, Y^l), y_t^l)^\top \end{pmatrix}$$

Dec self-attention

$$\hat{F}^l = f_M(S_p, E_p) \odot F^l$$

Masking function

$$p(w_{t+1} | X, Y_{\leq t}^{L}) = \mathrm{softmax}(W^V y_{t+1}^{L})$$



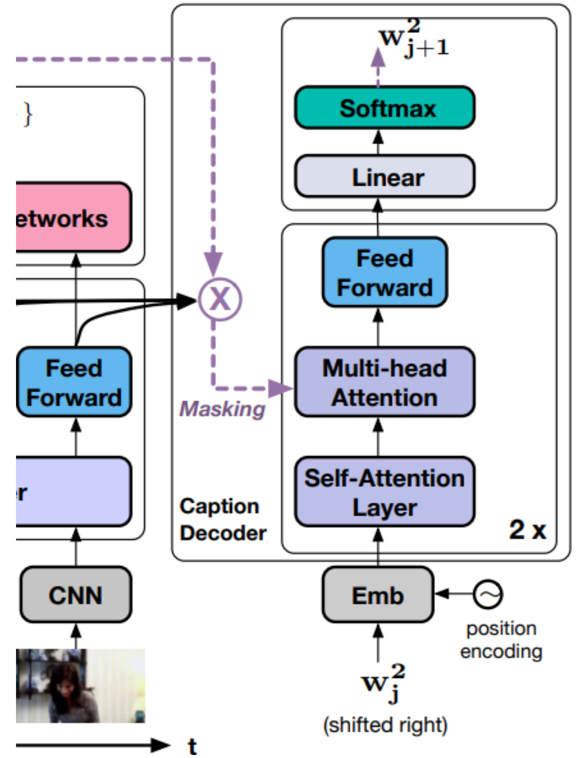
Method

- Differentiable Proposal Mask: proposal specific repre.

$$f_M(S_p, E_p, S_a, E_a, i) = \sigma(g(\\ [\rho(S_p,:), \rho(E_p,:), \rho(S_a,:), \rho(E_e,:), \mathrm{Bin}(S_a, E_a,:)]))$$

$$\rho(pos, i) = \begin{cases} \sin(pos/10000^{i/d}) & i \textit{ is even} \\ \cos(pos/10000^{(i-1)/d}) & \textit{otherwise} \end{cases}$$

$$\mathrm{Bin}(S_a, E_a, i) = \begin{cases} 1 & \textit{if } i \in [S_a, E_a] \\ 0 & \textit{otherwise} \end{cases}$$



- Gated masking:

$$f_{GM}(S_p, E_p, S_a, E_a, i) = \\ P_e \mathrm{Bin}(S_p, E_p, i) + (1 - P_e) f_M(S_p, E_p, S_a, E_a, i)$$

The continuous mask is used as a supplement for the proposal mask in case the confidence is low from the proposal module.

Method

- Model training:

$$\mathcal{L}_r = \mathrm{Smooth}_{\ell 1}(\hat{\theta}_c, \theta_c) + \mathrm{Smooth}_{\ell 1}(\hat{\theta}_l, \theta_l)$$

Regression loss for event boundary.

$$\mathcal{L}_m^i = \mathrm{BCE}(Bin(S_p, E_p, i), f_M(S_p, E_p, S_a, E_a, i))$$

Mask prediction loss.

$$\mathcal{L}_e = \mathrm{BCE}(\hat{P}_e, P_e)$$

Event classification loss.

$$\mathcal{L}_c^t = \mathrm{CE}(\hat{w}_t, p(w_t | X, Y_{\leq t-1}^L))$$

Captioning model loss.

$$\mathcal{L} = \lambda_1 \mathcal{L}_r + \lambda_2 \sum_i \mathcal{L}_m^i + \lambda_3 \mathcal{L}_e + \lambda_4 \sum_t \mathcal{L}_c^t$$

Experiment

- Results from ActivityNet Caption Dataset:

Table 4. Event proposal results from ActivityNet Captions dataset. We compare our proposed methods with our baseline method ProcNets-prop on the validation set.

Method	Average Recall (%)
ProcNets-prop [42]	47.01
Bi-LSTM (ours)	50.65
Self-Attn (our)	**52.95**

Dense video captioning challenge leader board results. ts from the same team, we keep the highest one.

Method	METEOR
DEM [19]	4.82
Wang et al.	9.12
Jin et al.	9.62
Guo et al.	9.87
Yao et al.[2](Ensemble)	12.84
Our Method	**10.12**

Table 1. Captioning results from ActivityNet Caption Dataset

Method	B@3	B@4	M
Bi-LSTM +TempoAttn	2.43	1.01	7.49
Masked Transformer	4.47	2.14	9.43
End-to-end Masked Transformer	**4.76**	**2.23**	**9.56**

Summarization

- End-to-end training.
- First to use self-attention in video captioning.