INF entrance to TRECVID2018 video to text task

Jia Chen[1], Shizhe Chen[2], Qin Jin[2], Alexander Hauptmann[1]

[1]Carnegie Mellon University

[2]Renmin University of China

Content

- Recap and what's new
- Network architecture
- Limitation of cross-entrpy loss
- Bridging the exposure bias
- Two losses
- Experiments

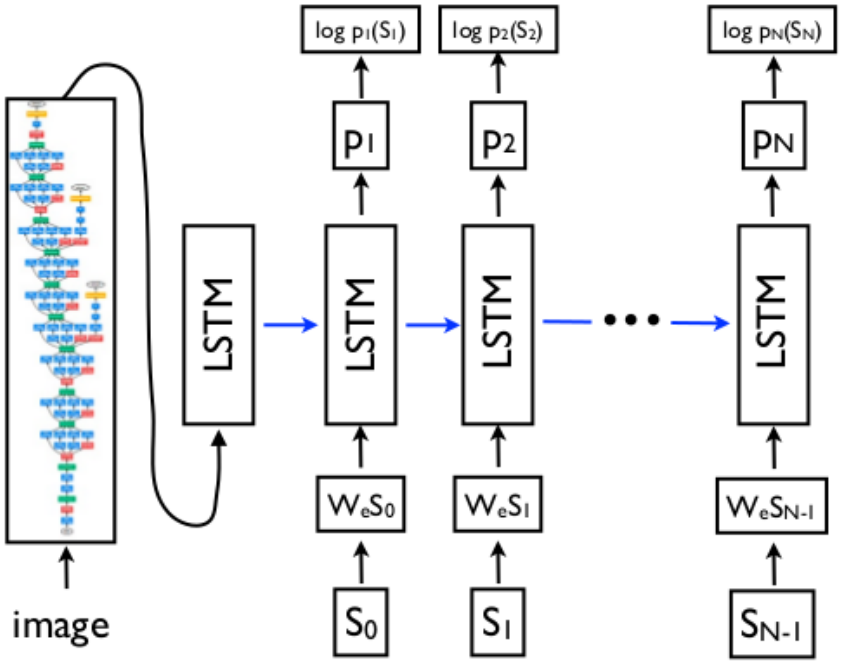
Recap and What's New

- Last year
  - Dataset vs. Network Architecture
    - dataset: low hanging fruit
    - network architecture: not too much improvement* (performance plateu)
- What's new in this year
  - Change the loss used in the caption task
    - brings large gain

*Knowing yourself: Improving video caption via in-depth recap. ACM MM 2017

Network Architecture

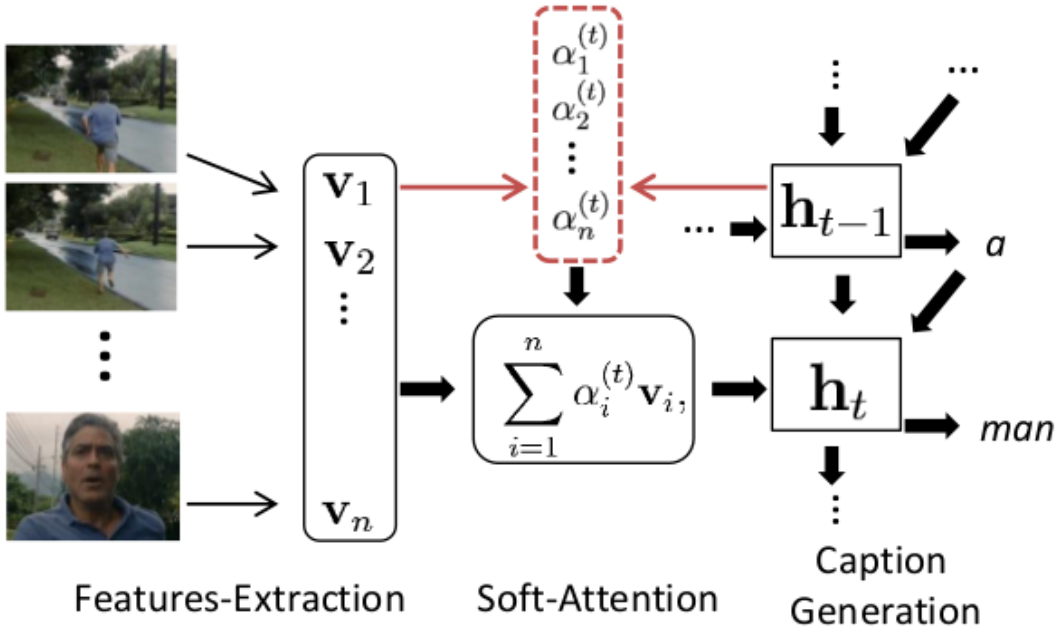
- Vanilla encoder-decoder architecture[2]



[2] Show and tell: A neural image caption generator. O Vinyal etc al. CVPR 2015

Network Architecture (cont'd)

- temporal attention[2]

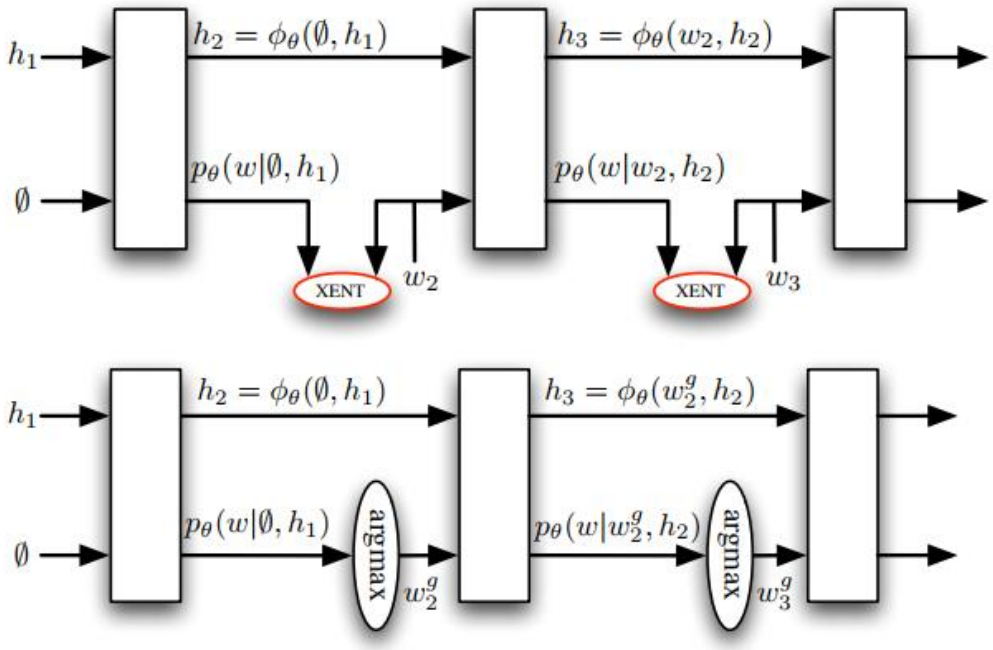


[2] Describing videos by exploiting temporal structure. Yao Li etc al. ICCV 2015

Limitation of cross-entrpy loss

train stage:

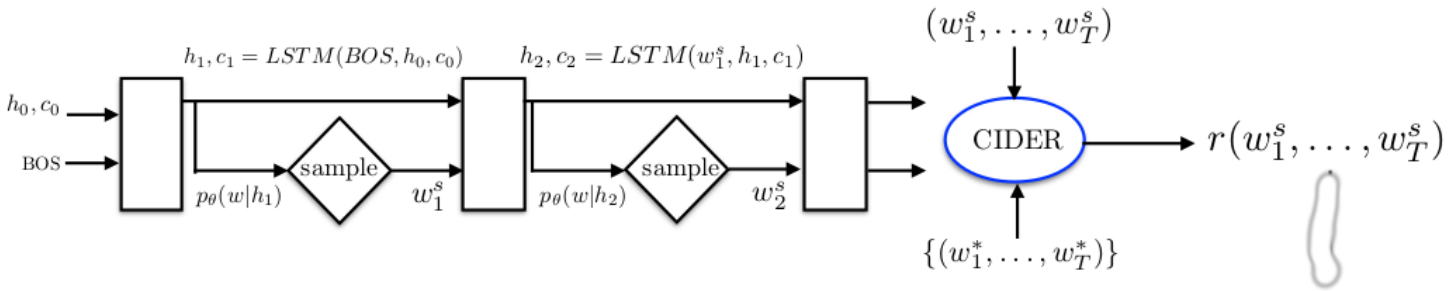
test stage:



[3] Sequence level training with recurrent neural networks. Ranzato, Marc'Aurelio, et al. ICLR 2015

Bridging the exposure gap

- Solution
  - feed step t-1's output to step t's input through sampling
  - use evaluation metric as reward*
  - use REINFORCE to train model (an algorithm of policy gradient in reinforcement learning)

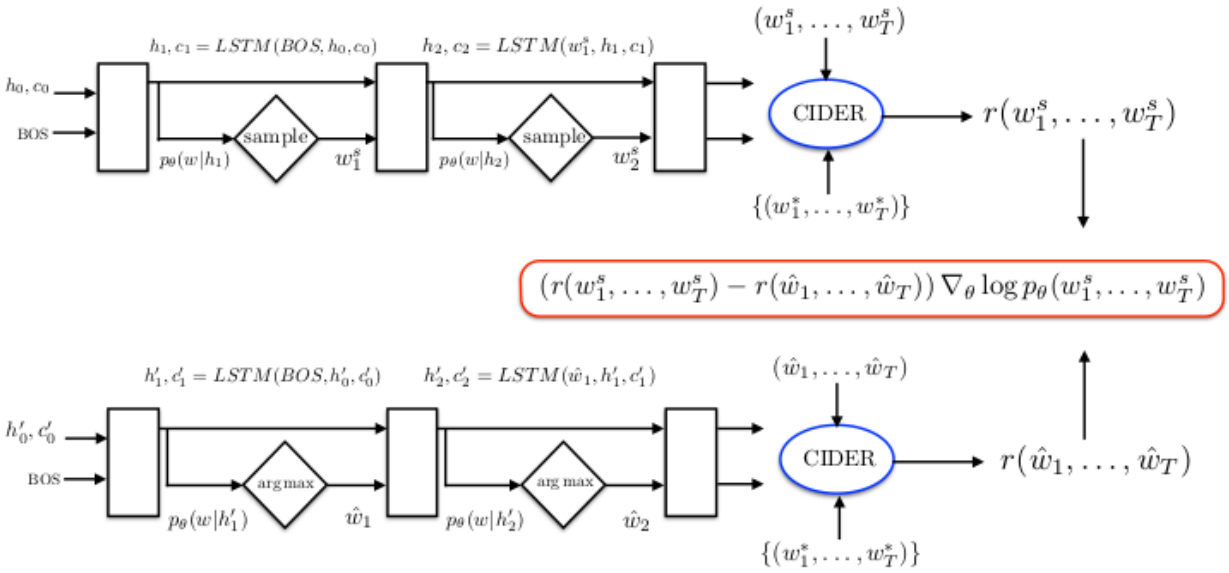


Bridging the exposure gap

- Caveat
  - sometimes the algorithm may exploit the loopholes in the reward
- Design a robust reward
  - CIDEr (closer to human evluation compared to BLEU and METEOR)
  - BCMR
    - weighted average of BLEU, CIDEr, METEOR, ROUGE

Two losses

- self-critique loss
  - greedy decoding as baseline to reduce variance



[4] Self-critical sequence training for image captioning. SJ Rennie, et al. CVPR 2017

Two losses

- PROS (partially observable set) loss*
  - $d(s_i, s_j)$ distance of two captions s_i and s_j

$$L_{accuracy} = -\mathbb{E}_{s \sim \Pr(\theta)}[r]$$

$$L_{diversity} = -\mathbb{E}_{s_i, s_j \sim \Pr(\theta)}[d(s_i, s_j)]$$

*work under progress

Experiments

- Training set
  - TGIF (all)
  - TRECVID16 (optional)
- Validation set
  - TRECVID17
- Feature
  - Resnet200 (pretrained on ImageNet)
  - I3D (pretrained on Kinetics-400)

Experiments

- performance on validation set

model	loss	BLEU4	METEOR	CIDEr
vanilla	cross entropy	7.1	12.4	27.6
	self critique	7.7	13.2	31.3
	PROS	8.1	13.9	32.5
temporal attention	cross entropy	7.6	12.5	28.9
	self critique	7.4	13.0	32.1

Experiments

- performance on TRECVID18

model	loss	BLEU4	METEOR	CIDEr
vanilla	PROS	2.4	23.1	41.6
attention	self critique	1.8	22.1	40.8

Conclusion

- Reformulate the problem (e.g. by loss) from scratch brings improvement over the current performance plateu