

Weakly Supervised Dense Video Captioning

Zhiqiang Shen (UIUC)

<https://arxiv.org/abs/1704.01502>

Presented at CVPR 2017

Video Captioning Annotation Examples & Key Findings



Ground Truth

two men are in a wrestling match

a referee coaches a wrestling match

a wrestler does a victory dance

man wrestle professionally

two guys are wrestling in a competition

Key Findings:

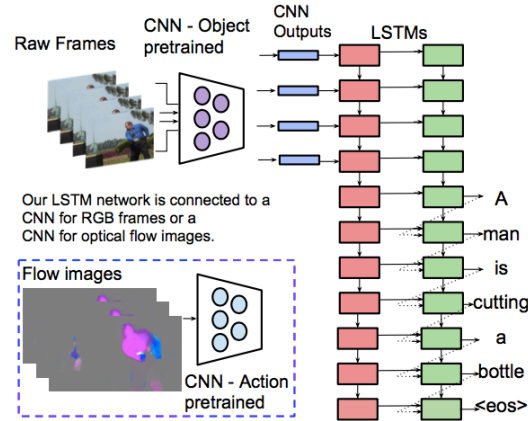
- 1) 1 clip vs multiple diverse sentences annotations for different regions/segments
- 2) Current video captioning methods only generate 1 description per clip.
- 3) Training with 1:N (visual feature : sentences) is Inaccurate even with soft-attention.

Motivations: weakly grounding/attend sentence to region-sequences for captioning?

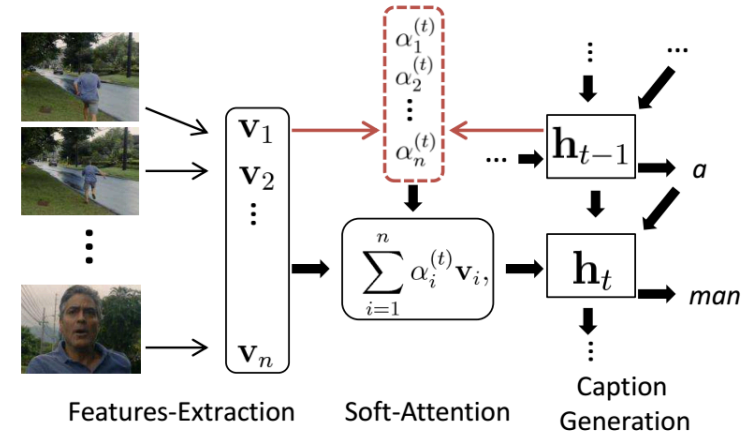
What is Traditional Video Captioning?

- Goal: Generate one sentence per input clip
 - With multiple video-level sentence annotations.
 - With only one global video clip representation.
- Weakness
 - One visual representation to N-sentences
- Typical Methods
 - Sequence-to-sequence video captioning
 - Describing Videos by Exploiting Temporal Structure (soft-attention)
 - Jointly Modeling Embedding and Translation to Bridge Video and Language

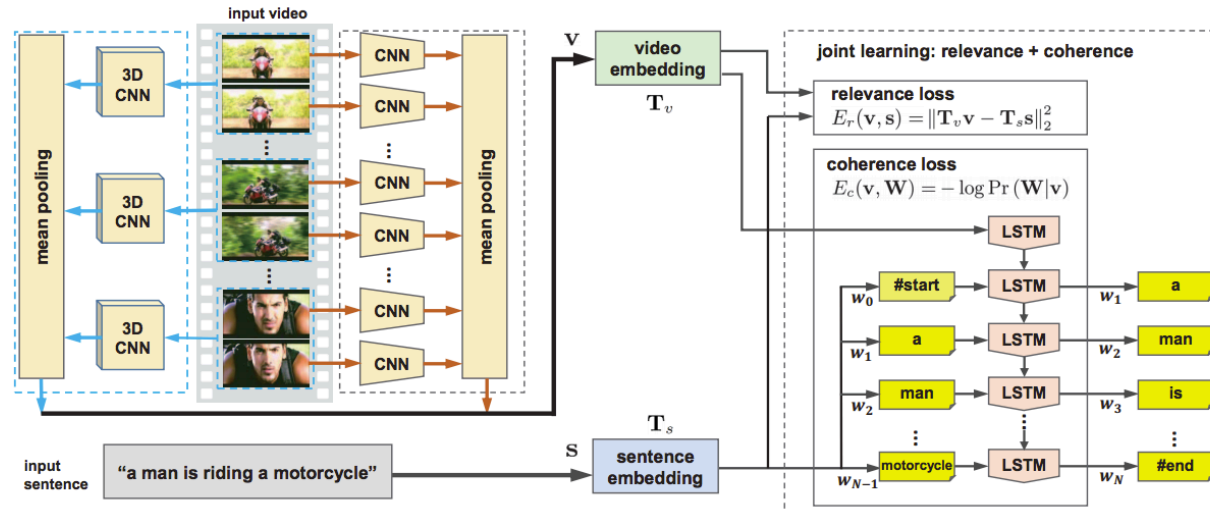
Methods for Single Sentence Captioning



Sequence-to-Sequence (ICCV2015)



Soft-attention (ICCV2015)

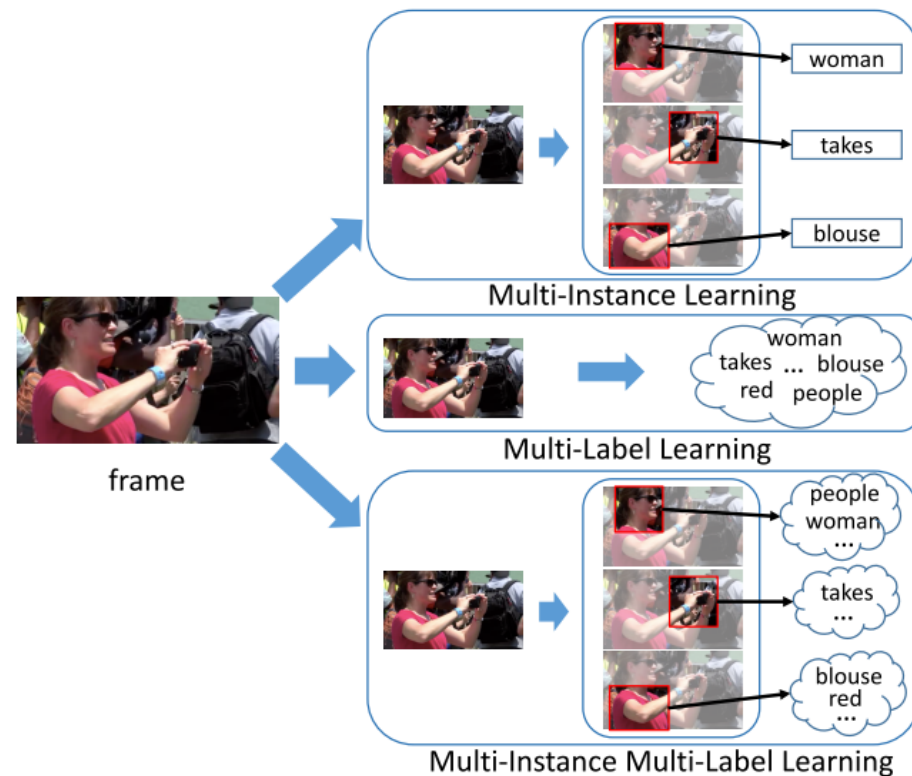
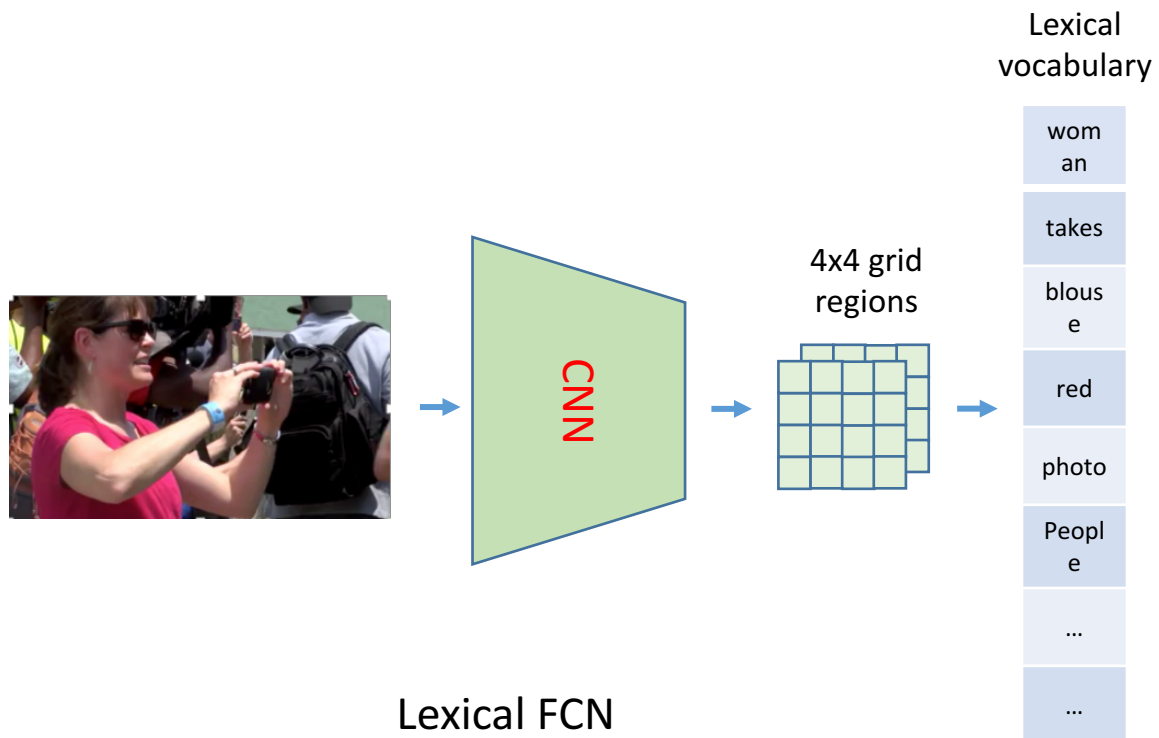


Jointly embedding (CVPR2016)

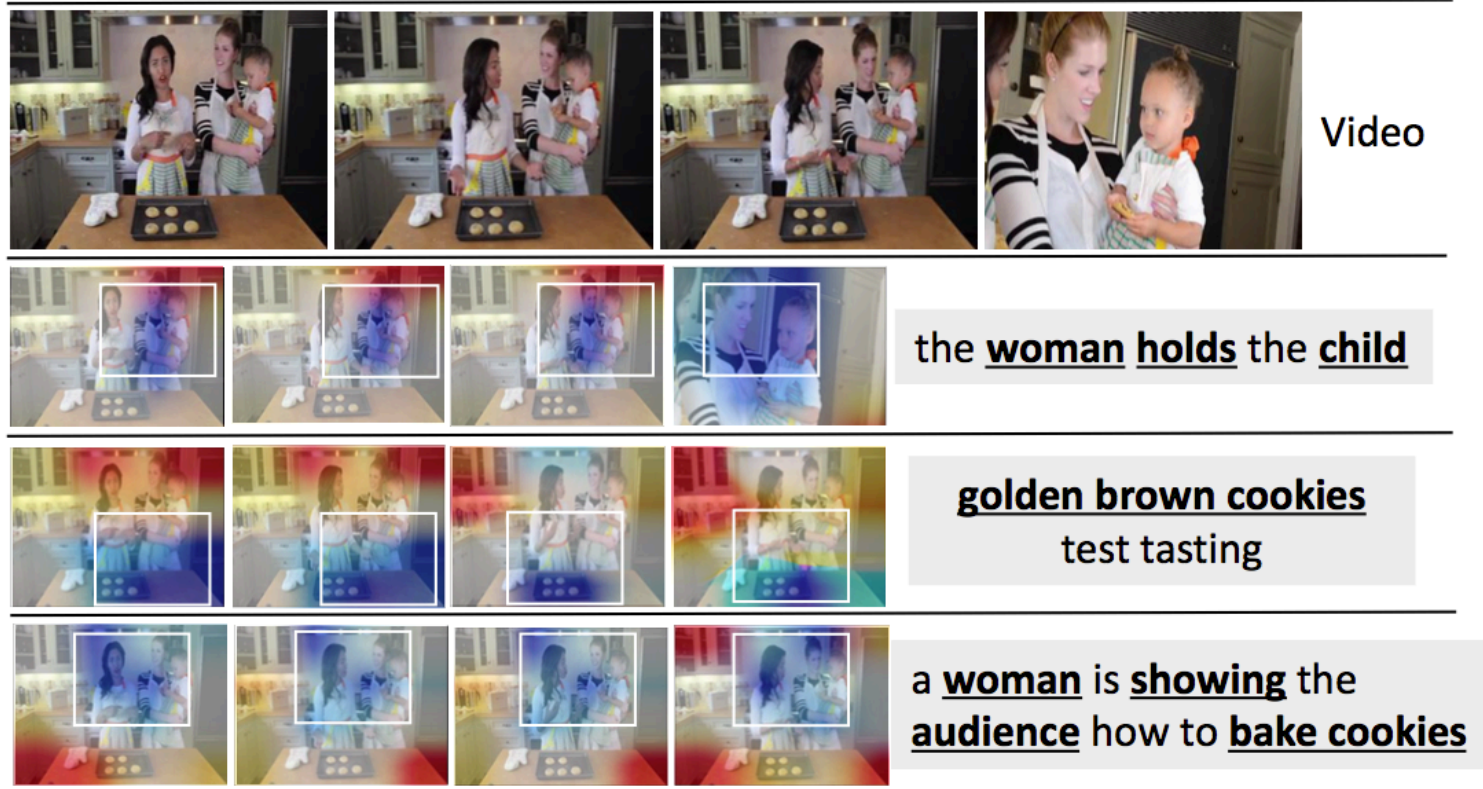
What is Dense Video Captioning?

- Goal: Generate multiple diverse/informative sentences per input clip
 - With only video level sentence annotations.
 - Through weakly attend sentences to region-sequences.
- Avoid
 - Mis-matching due to 1-visual : N-sentences
 - Tedious annotations for strong supervised learning.
- Three Components
 - Lexical FCN: for region-level encoding using lexical vocabulary
 - Sentence to region-sequence association
 - Sentence generation with sequence-to-sequence learning

Training Lexical-FCN

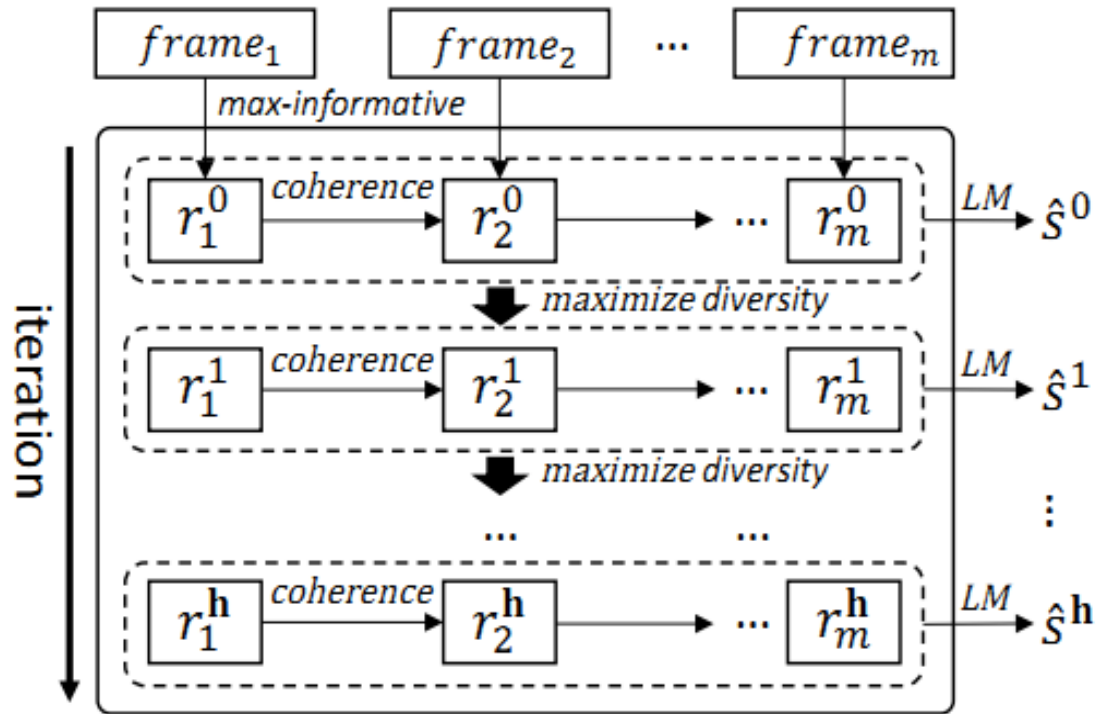


Learned Response Maps



From the last CNN conv layer

Sentence-to-region-sequence Association



- Informative

$$f_{\text{inf}}(\mathbf{x}_v, \mathcal{A}_t) = \sum_w p^w;$$

- Coherence

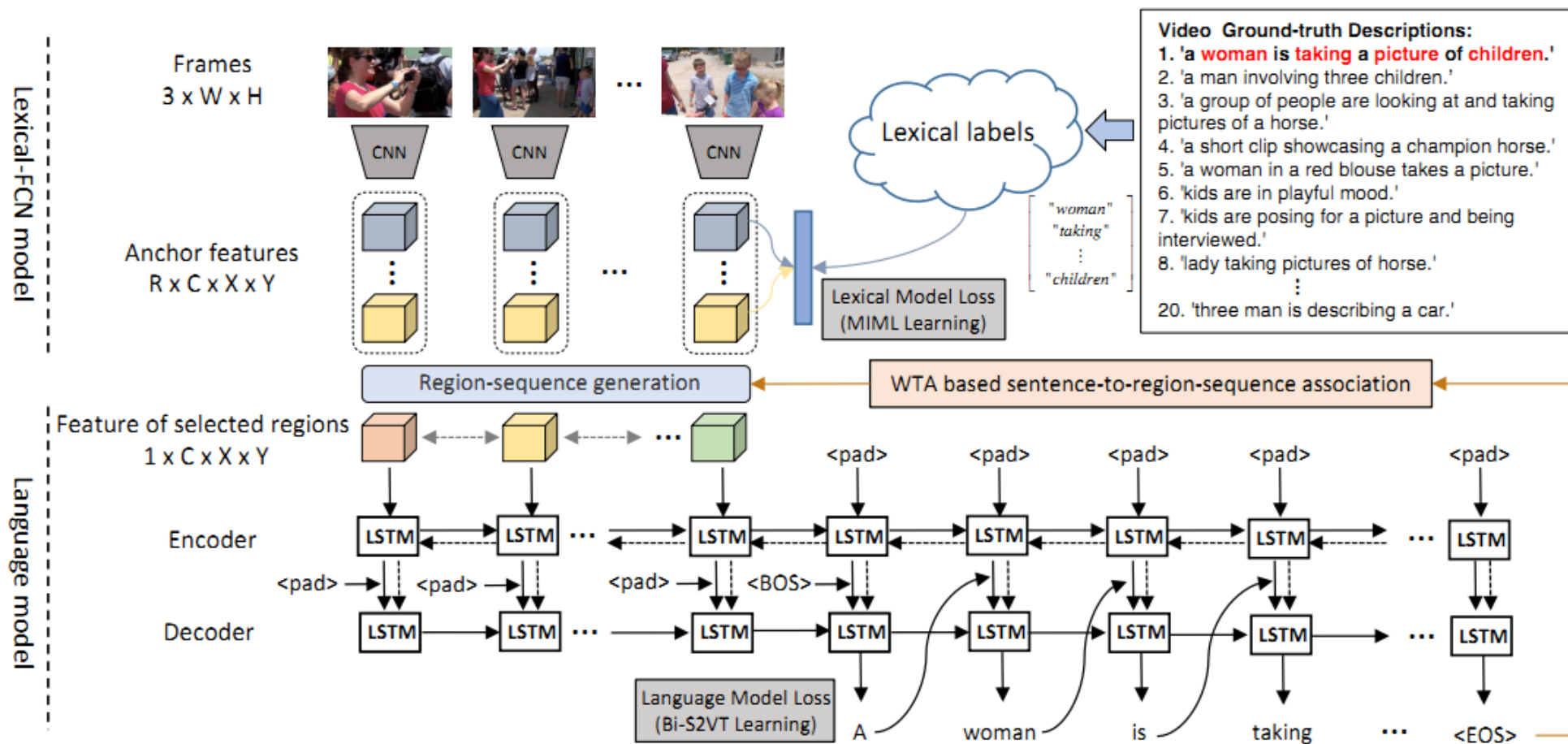
$$f_{\text{coh}} = \sum_{r_s \in \mathcal{A}_{t-1}} \langle \mathbf{x}_{r_t}, \mathbf{x}_{r_s} \rangle,$$

- Diversity (KL divergence)

$$f_{\text{div}} = \sum_{i=1}^N \int_w p_i^w \log \frac{p_i^w}{q^w} dw.$$

Greedy solution: submodular maximization with these 3 cues






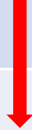
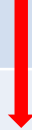





Framework Overview



Ablation Experiments on MSR-VTT Validation Set

Method	METEOR	BLEU-4	ROUGE-L	CIDEr
VideoLAB	27.7	39.5	61.0	44.2
Aalto	27.7	41.1	59.6	46.4
V2t_navigator	29.0	43.7	61.4	45.7
Ours w/o category	27.7	39.0	60.1	44.0
Ours category-wise	28.2	40.9	61.8	44.7
Ours+C3D+Audio	29.4	44.2	62.6	50.5

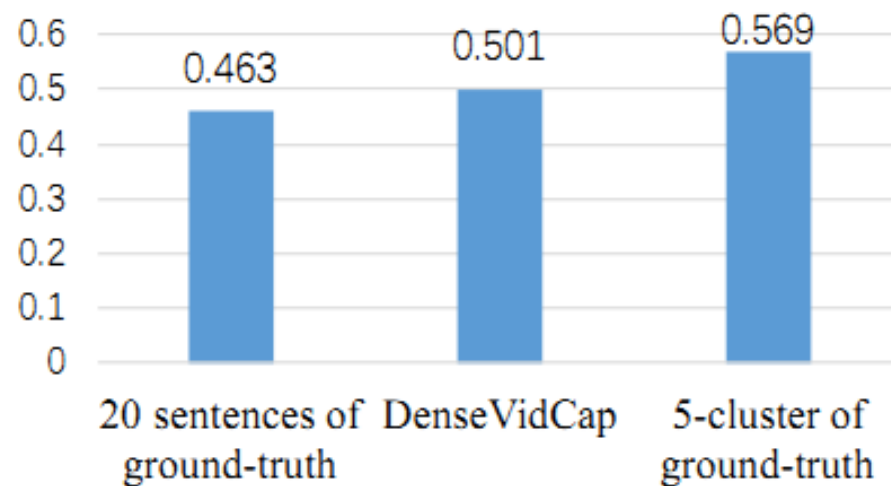
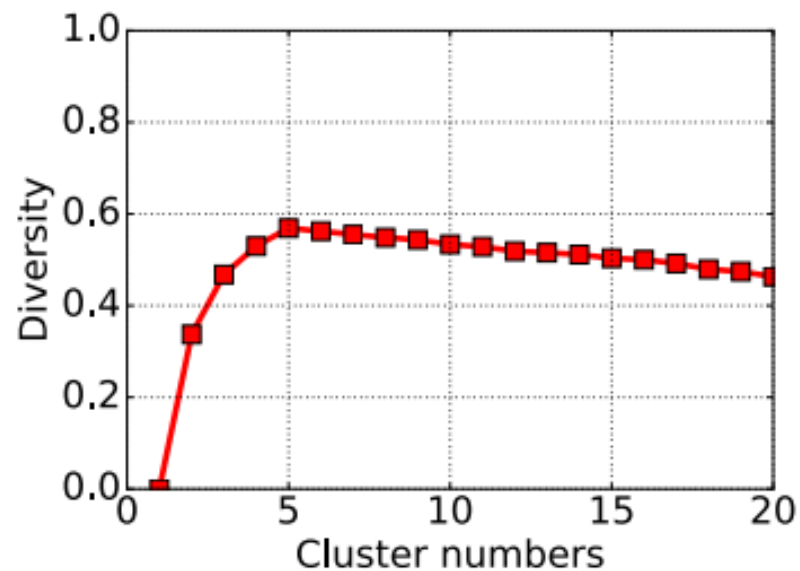
Ablation Experiments on MSR-VTT Validation Set

Method	METEOR	BLEU-4	ROUGE-L	CIDEr
VideoLAB	27.7	39.5	61.0	44.2
Aalto	27.7	41.1	59.6	46.4
V2t_navigator	29.0	43.7	61.4	45.7
Ours w/o category	27.7 	39.0 	60.1 	44.0 
Ours category-wise	28.2 	40.9 	61.8 	44.7 
Ours+C3D+Audio	29.4 	44.2 	62.6 	50.5 

Best Single Model Results on MSR-VTT

Team	Memo	METEOR	BLEU-4	ROUGE-L	CIDEr
Ruc-UVA	RUC + UVA + ZJU	26.9	38.7	58.7	45.9
VideoLab	UCB + Austin + ...	27.7	39.1	60.6	44.1
Aalto	Aalto Univ.	26.9	39.8	59.8	45.7
V2t-navigator	RUC + CMU	28.2	40.8	60.9	44.8
Ours	Fudan + ILC	28.3	41.4	61.1	48.9

Measure of Diversity on MSR-VTT



Diversity:

$$D_{div} = \frac{1}{n} \sum_{\mathbf{s}^i, \mathbf{s}^j \in \mathbf{S}; i \neq j} (1 - \langle \mathbf{s}^i, \mathbf{s}^j \rangle)$$

Examples: Dense Video Captioning

Region Sequences & DenseVidCap



a man is **drinking** from a **cup**



a man is **drinking** from a **bottle**



a man in a **suit** is **talking** to another man in a **suit**

Demo: Dense Video Captioning



Summary of DenseVidCap

- Lexical-FCN for weakly region modeling.
- Attend sentences to region-sequence with Lexical-FCN outputs.
- Dense video captioning with only video-level annotations.
- Avoid the problem of 1:N (feature to sentences) matching

Projects in ICCV'17

DSOD (Deeply Supervised Object Detectors from Scratch).

- RoI pooling is just like **max pooling**
- Forward / backward

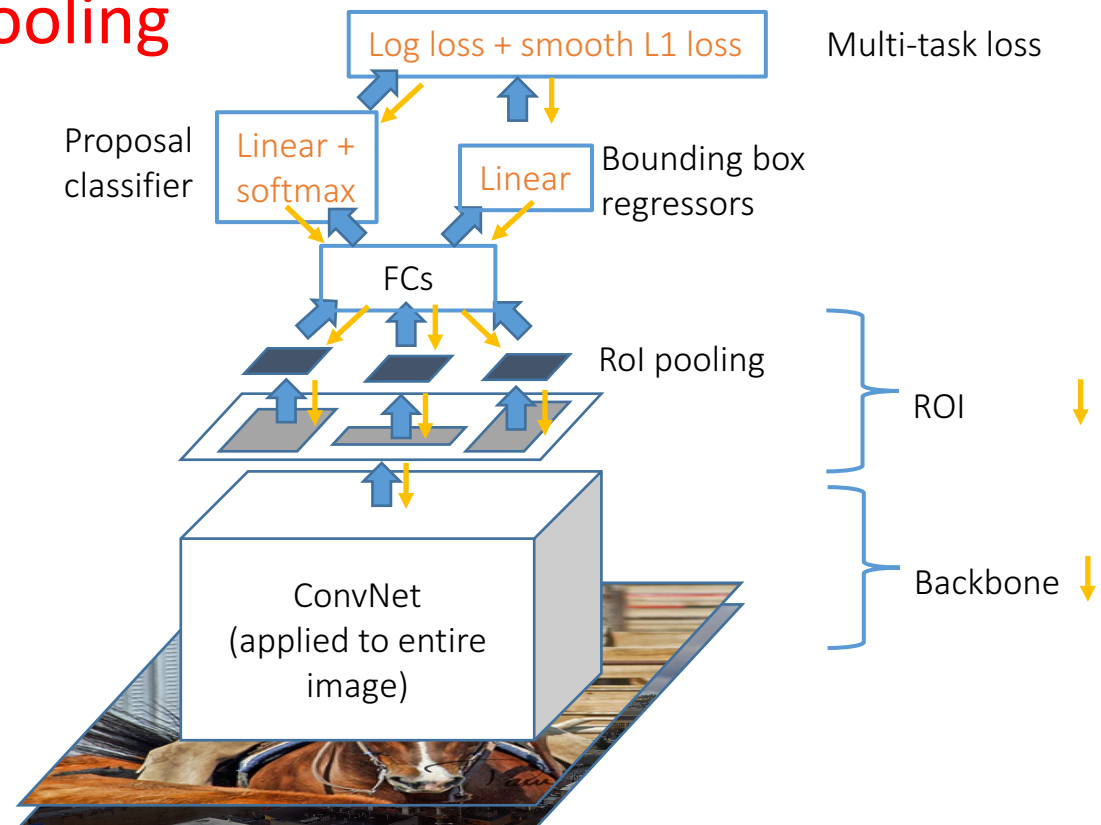
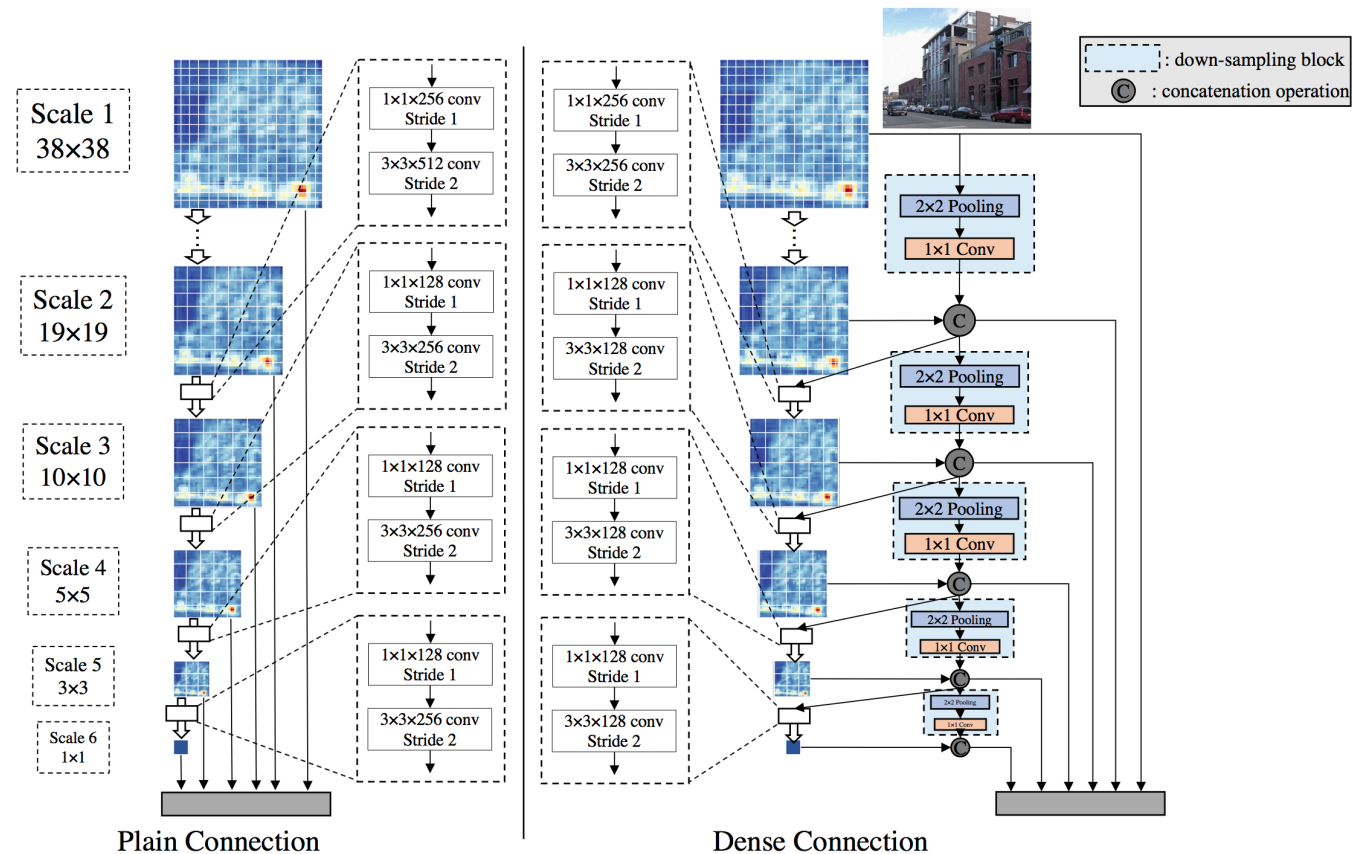


Figure from Ross Girshick

Principles

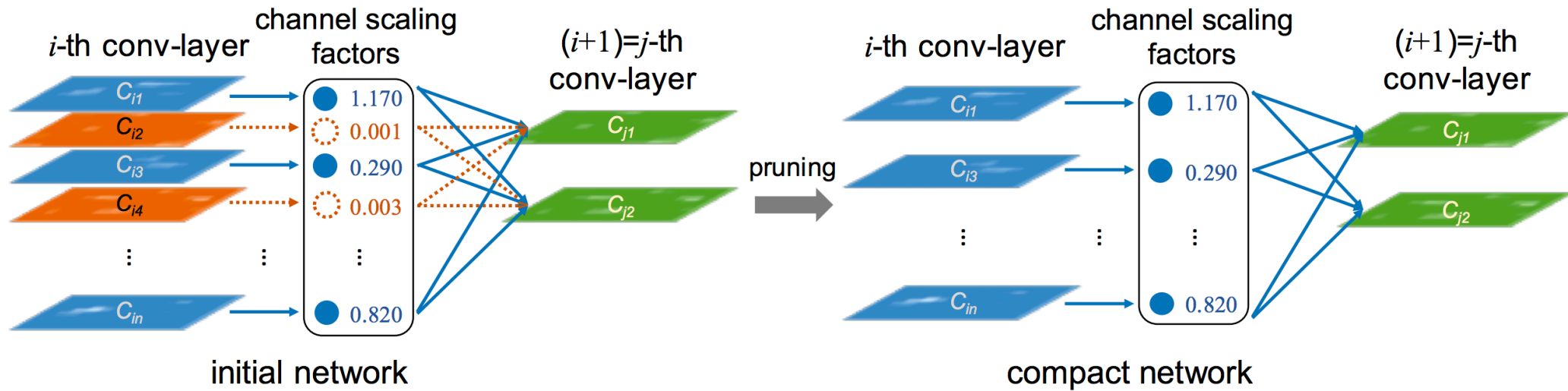
- Proposal-free.
- Deep Supervision.
- Dense Prediction Structure.
- Stem Block.



- Paper: <https://arxiv.org/abs/1708.01241>
- Code & models: <https://github.com/szq0214/DSOD>
- Network: <http://ethereon.github.io/netscope/#/gist/b17d01f3131e2a60f9057b5d3eb9e04d>

Projects in ICCV'17

➤ Network Slimming (ICCV'17)



“Learning Efficient Convolutional Networks through Network Slimming”. Zhuang Liu, Jianguo Li, **Zhiqiang Shen**, Gao Huang, Shoumeng Yan, Changshui Zhang. ICCV'17

Code: <https://github.com/liuzhuang13/slimming>

Thanks & Questions

My homepage: <http://www.zhiqiangshen.com/>

Email: zhiqiangshen0214@gmail.com

Any comments or suggestions are welcome!