Span-based Localizing Network for Natural Language Video Localization

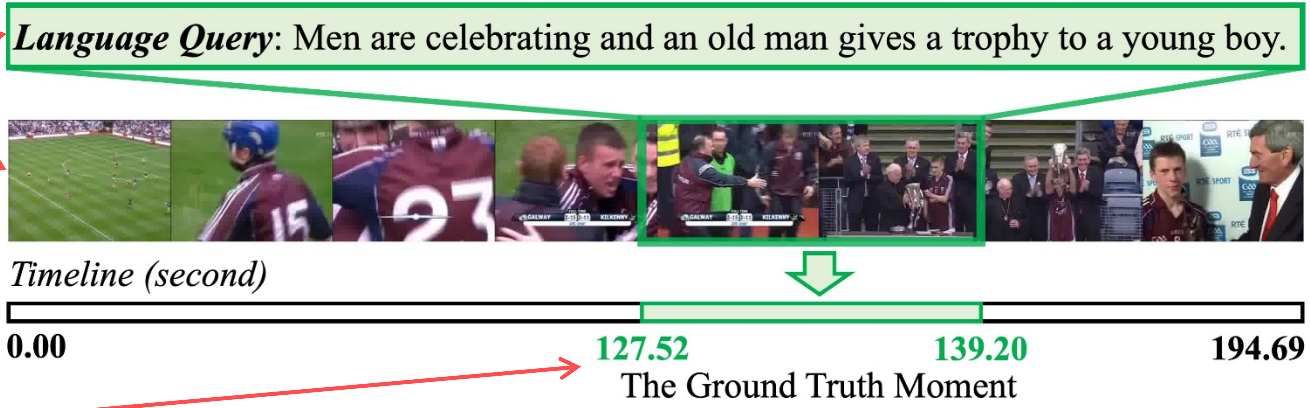
Hao Zhang[1,2], Aixin Sun[1], Wei Jing[3], Joey Tianyi Zhou[2]

[1] *School of Computer Science and Engineering, Nanyang Technological University, Singapore*
[2] *Institute of High Performance Computing, A*STAR, Singapore*
[3] *Institute of Infocomm Research, A*STAR, Singapore*

ACL 2020

What is Natural Language Video Localization (NLVL)

Input:
- ➢ A language query
- ➢ An untrimmed video



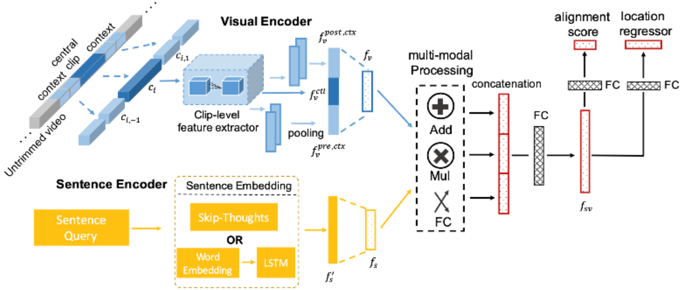
Output:
- ➢ A temporal moment

Agency for Science, Technology and Research SINGAPORE

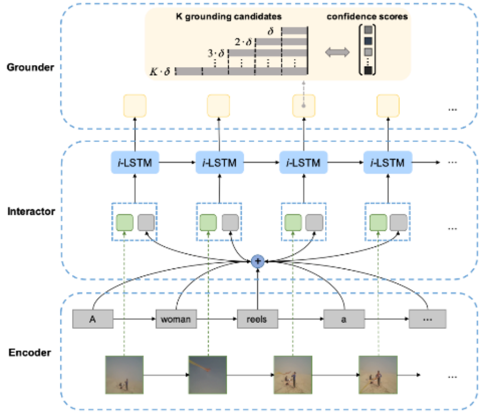
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Existing Works for NLVL

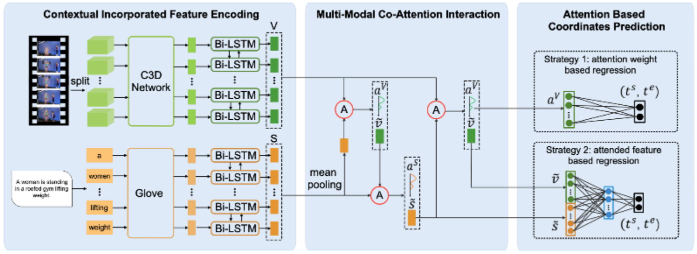
1. **Ranking based methods**, *e.g.*, CTRL, *Gao et al., 2017, ICCV*.



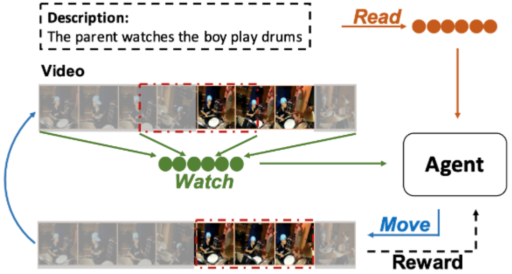
2. **Anchor based methods**, *e.g.*, TGN, *Chen et al., 2018 EMNLP*.



3. **Regression based methods**, *e.g.*, ABLR, *Yuan et al., 2019, AAAI*.



4. **Reinforcement learning based methods**, *e.g.*, RWM-RL, *He et al., 2019, AAAI*.



Agency for Science, Technology and Research
SINGAPORE

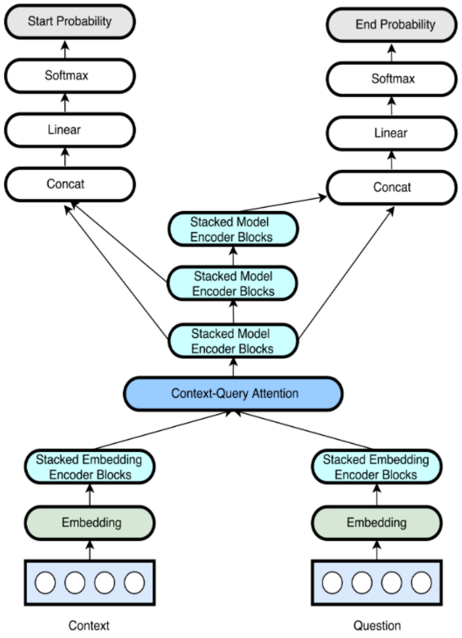
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

A Typical Span-based QA Framework

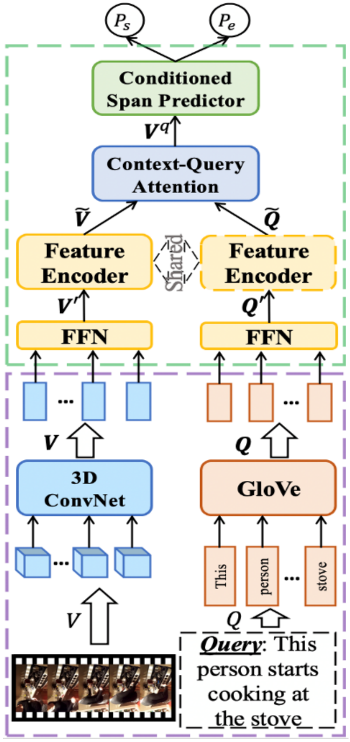
Span-based QA
- *Input*: **text passage** and language query.
- *Output*: **word phrase** as answer span.

NLVL
- *Input*: **untrimmed video** and language query.
- *Output*: **temporal moment** as answer span.

A different perspective:
- **NLVL ⟶ Span-based QA**

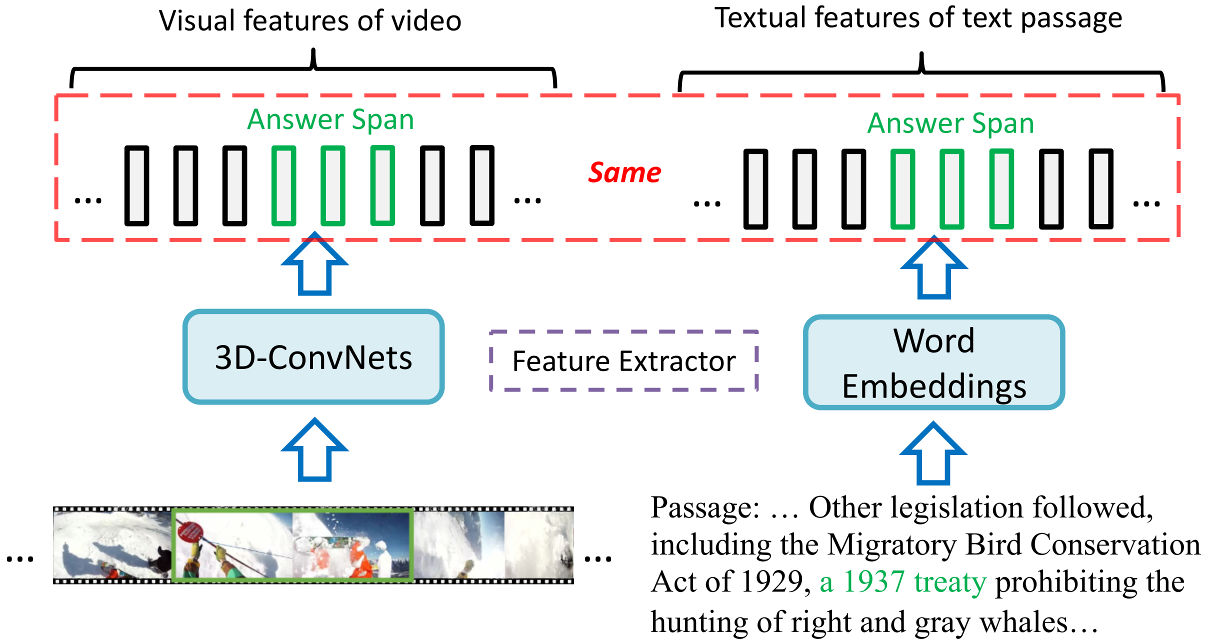


QANet for span-based QA, Yu et al., 2018, ICLR.



VSLBase for NLVL.

Similarities between NLVL and Span-based QA



NLVL shares significant similarities with span-based QA by treating:

Target moment ←→ Answer span

Video ←→ Text passage

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Differences between NLVL and Span-based QA

- Video is continuous and causal relations between video events are usually adjacent.
  - Many events in a video are directly correlated and can even cause one another.
- Natural language is inconsecutive and words in a sentence demonstrate syntactic structure
  - Causalities between word spans or sentences are usually indirect and can be far apart.
- Changes between adjacent video frames are usually very small, while adjacent word tokens may carry distinctive meanings.
- Compared to word spans in text, human is insensitive to small shifting between video frames.
  - Small offsets between video frames do not affect the understanding of video content.
  - The differences of a few words or even one word could change the meaning of a sentence.

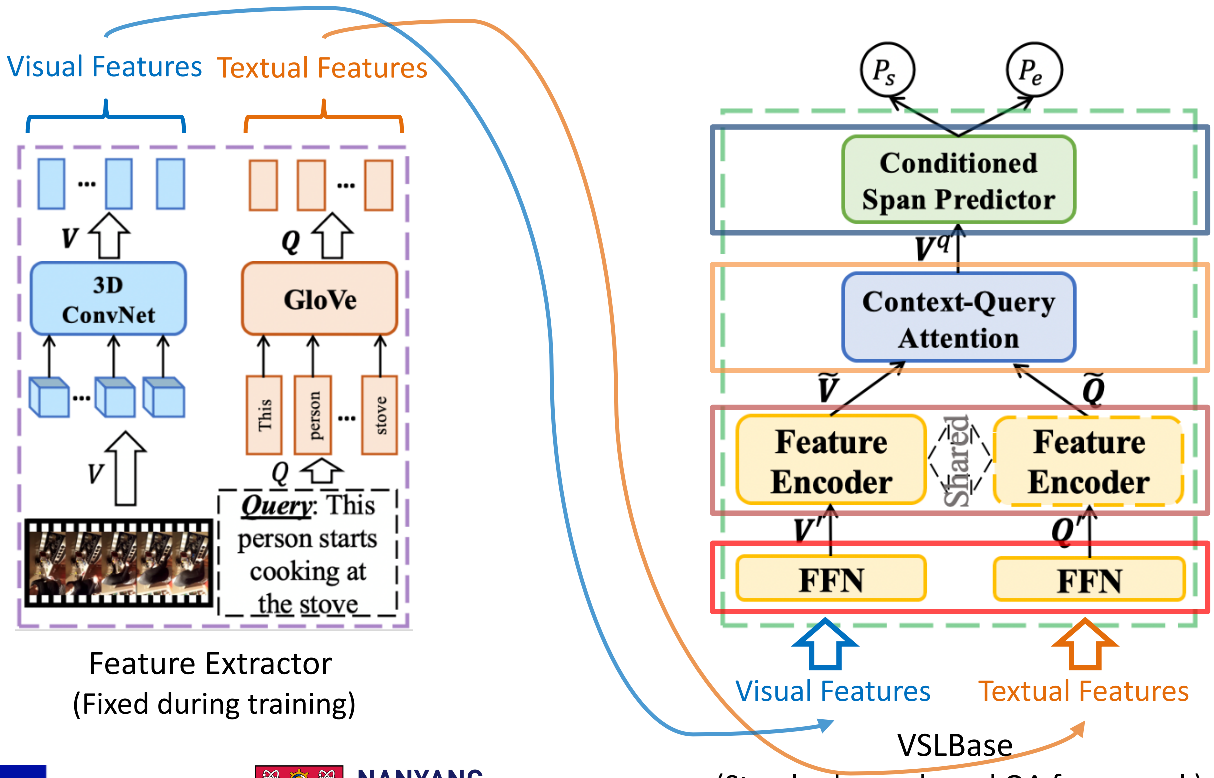
Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Span-based QA Framework for NLVL

$$\text{span}(\hat{a}^s, \hat{a}^e) = \arg\max_{\hat{a}^s, \hat{a}^e} P_s(\hat{a}^s) P_e(\hat{a}^e)$$

$$\text{s.t. } 0 \leq \hat{a}^s \leq \hat{a}^e \leq n$$



Feature Extractor
(Fixed during training)

VSLBase
(Standard span-based QA framework)

Predict the spans of start and end boundaries of target moment.

Capture the cross-modal interactions between visual and textual features.

A single transformer block to encode contextual information.

Project visual and textual features into same dimension.

Video Span-based Localizing Network (VSLNet)

- Query-Guided Highlighting (QGH) extends the boundaries of foreground to cover its antecedent and consequent contents.
- The **target moment** and **its adjacent contexts** are regarded as foreground; the **rest** as background.
- With QGH, VSLNet is guided to search for the target moment within *a highlighted region*.

Query-Guided Highlighting is introduced to address the two differences between NLVL and span-based QA.

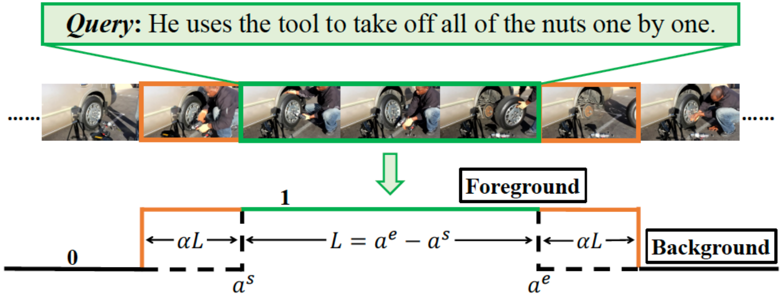
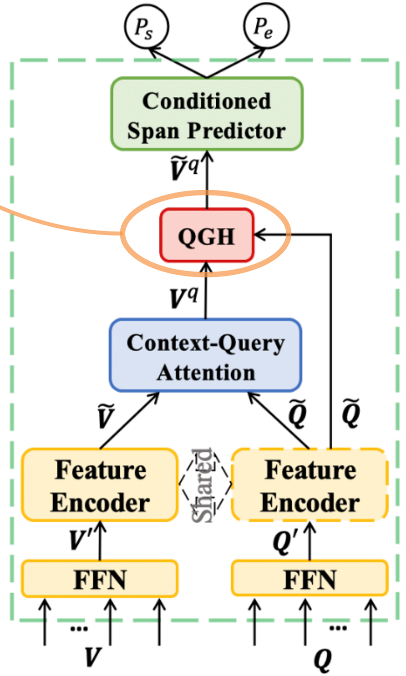


Illustration of foreground and background of visual features.
α is the ratio of foreground extension.



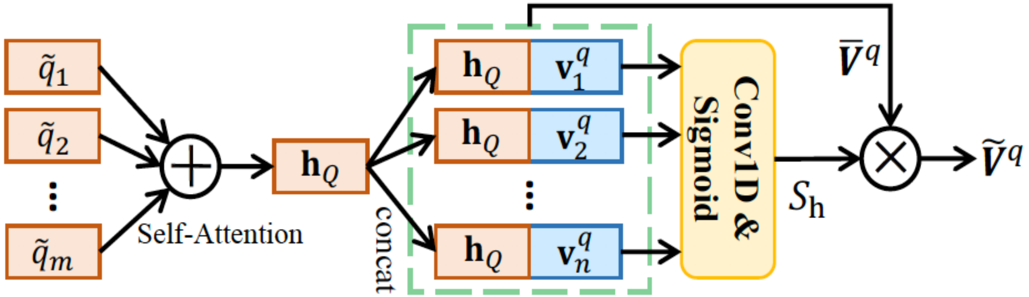
VSLNet

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Bridging the Gap between NLVL and Span-based QA

- Foreground ⟶ 1, background ⟶ 0.
- QGH is a **binary classification** module.



The structure of Query-Guided Highlighting

- The longer region provides additional contexts for locating answer span.
- The highlighted region helps the network to focus on subtle differences between video frames.

Evaluation Metrics

s_1: ground truth moment corresponding to text query $\boldsymbol{q_1}$,
"*clip c*": predicted moment.

- **Union**: the total length of both s_1 and "*clip c*"
- **Intersection**: the overlap between s_1 and "*clip c*"
- **Intersection over Union**: $\text{IoU} = \frac{\text{Intersection}}{\text{Union}}$

Evaluation Metrics:

- $\mathbf{Rank@}\boldsymbol{n}\mathbf{, IoU} = \boldsymbol{\mu}$
- $\mathbf{mIoU}$ (mean IoU)

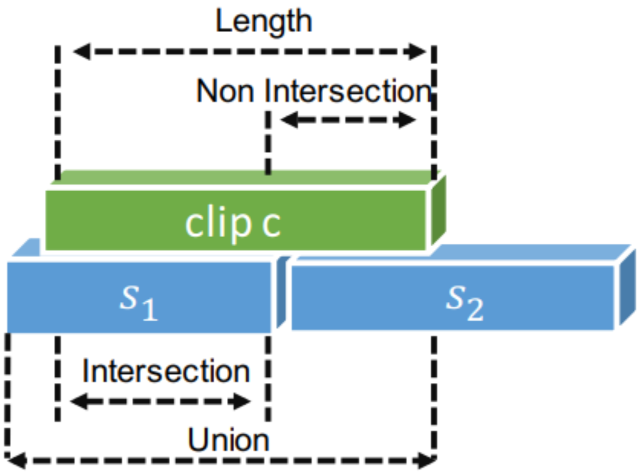


Figure from Gao et al. 2017, *ICCV*.

Agency for Science, Technology and Research SINGAPORE
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Benchmark Datasets

- **Charades-STA** is obtained from Charades dataset; the videos are about ***daily indoor activities***.
- **ActivityNet Captions** contains about 20k ***open-domain*** videos taken from ActivityNet dataset.
- **TACoS** is selected from MPII ***Cooking Composite Activities*** dataset.

Dataset	Domain	# Videos (train/val/test)	# Annotations	N_{vocab}	$\bar{L}_{video}$	$\bar{L}_{query}$	$\bar{L}_{moment}$	Δ_{moment}
Charades-STA	Indoors	$5,338/-/1,334$	$12,408/-/3,720$	$1,303$	$30.59s$	7.22	$8.22s$	$3.59s$
ActivityNet Cap	Open	$10,009/-/4,917$	$37,421/-/17,505$	$12,460$	$117.61s$	14.78	$36.18s$	$40.18s$
TACoS	Cooking	$75/27/25$	$10,146/4,589/4,083$	$2,033$	$287.14s$	10.05	$5.45s$	$7.56s$

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Compared Methods

- **Ranking based (multimodal matching) methods:** *CTRL* (Gao et al., 2017), *ACRN* (Liu et al., 2018), *ACL* (Ge et al., 2019), *QSPN* (Xu et al., 2019), *SAP* (Chen et al., 2019)
- **Anchor based methods:** *TGN* (Chen et al., 2018), *MAN* (Zhang et al., 2019)
- **Reinforcement learning based methods:** *SM-RL* (Wang et al., 2019), *RWM-RL* (He et al., 2019)
- **Regression based methods:** *ABLR* (Yuan et al., 2019), *DEBUG* (Lu et al., 2019)
- **Span based methods:** *L-Net* (Chen et al., 2019), *ExCL* (Ghosh et al., 2019)

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Comparison with State-of-the-Arts

- VSLNet significantly outperforms all baselines by a large margin over all evaluation metrics.
- The improvements of VSLNet are more significant under more strict metrics.
- VSLBase outperforms all compared baselines over $\mathrm{IoU} = 0.7$.

Model	$\mathrm{IoU} = 0.3$	$\mathrm{IoU} = 0.5$	$\mathrm{IoU} = 0.7$	mIoU
C3D model without fine-tuning as visual feature extractor				
CTRL	-	23.63	8.89	-
ACL-K	-	30.48	12.20	-
QSPN	54.70	35.60	15.80	-
SAP	-	27.42	13.36	-
SM-RL	-	24.36	11.17	-
RWM-RL	-	36.70	-	-
MAN	-	<u>46.53</u>	22.72	-
DEBUG	54.95	37.39	17.69	36.34
VSLBase	<u>61.72</u>	40.97	<u>24.14</u>	<u>42.11</u>
VSLNet	**64.30**	**47.31**	**30.19**	**45.15**
C3D model with fine-tuning on Charades dataset				
ExCL	65.10	44.10	23.30	-
VSLBase	<u>68.06</u>	<u>50.23</u>	<u>30.16</u>	<u>47.15</u>
VSLNet	**70.46**	**54.19**	**35.22**	**50.02**

Results (%) of "$\mathrm{R@1;\ IoU} = \mu$" and "mIoU" compared with SOTA on Charades-STA. Best results are in **bold** and second best <u>underlined</u>.

Comparison with State-of-the-Arts

Similar observations hold on ActivityNet Captions and TACoS datasets.

- VSLNet **outperforms** all baseline methods.
- VSLBase shows **comparable performance** with baseline methods.

- **Adopting span-based QA framework for NLVL is promising.**

Model	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
TGN	45.51	28.47	-	-
ABLR	55.67	36.79	-	36.99
RWM-RL	-	36.90	-	-
QSPN	45.30	27.70	13.60	-
ExCL*	63.00	**43.60**	24.10	-
DEBUG	55.91	39.72	-	39.51
VSLBase	58.18	39.52	23.21	40.56
VSLNet	**63.16**	43.22	**26.16**	**43.19**

Results (%) of "R@1; IoU $= \mu$" and "mIoU" compared with SOTA on ActivityNet Captions.

Model	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
CTRL	18.32	13.30	-	-
TGN	21.77	18.90	-	-
ACRN	19.52	14.62	-	-
ABLR	19.50	9.40	-	13.40
ACL-K	24.17	20.01	-	-
L-Net	-	-	-	13.41
SAP	-	18.24	-	-
SM-RL	20.25	15.95	-	-
DEBUG	23.45	11.72	-	16.03
VSLBase	23.59	20.40	16.65	20.10
VSLNet	**29.61**	**24.27**	**20.03**	**24.11**

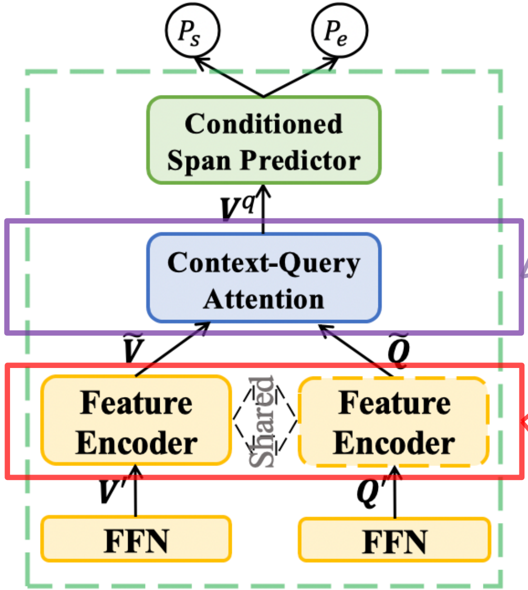
Results (%) of "R@1; IoU $= \mu$" and "mIoU" compared with SOTA on TACoS.

Why we Select Transformer Block and Context-Query Attention?

Module	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
BiLSTM + CAT	61.18	43.04	26.42	42.83
CMF + CAT	63.49	44.87	27.07	44.01
BiLSTM + CQA	65.08	46.94	28.55	45.18
CMF + CQA	68.06	50.23	30.16	47.15

Comparison between models with alternative modules in VSLBase on Charades-STA.

- **CMF shows stable superiority over BiLSTM regardless of other modules.**
- **CQA surpasses CAT whichever encoder is used.**



CQA: Context-Query Attention?

CAT: Direct concatenation of visual and textual features?

Single transformer block? (**CMF**: Conv + Multihead + FFN)

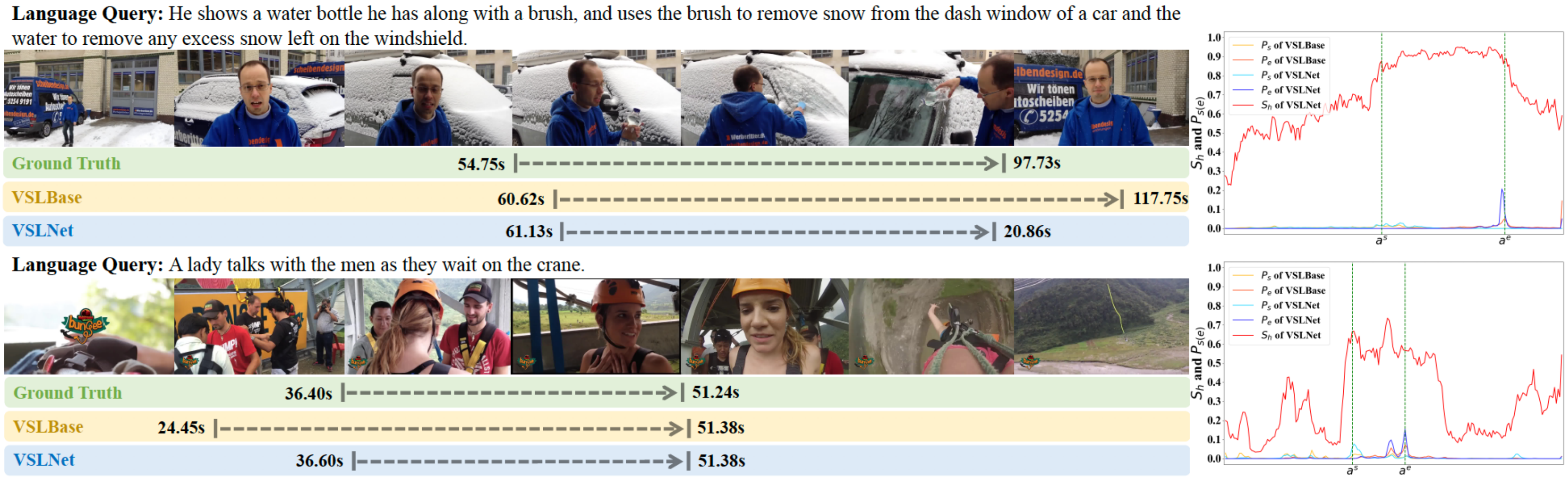
BiLSTM: Bidirectional LSTM?

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Qualitative Analysis

- ➢ The localized moments by VSLNet are closer to ground truth than that by VSLBase.
- ➢ The start and end boundaries predicted by VSLNet are softly constrained in the highlighted regions computed by QGH.



Visualization of predictions by VSLBase and VSLNet on ActivityNet Captions dataset.

Conclusion

- Span-based QA framework works well on NLVL task and is able to achieve state-of-the-art performances.
- With QGH, VSLNet effectively addresses the two major differences between video and text and improve the performance.
- Explore span-based QA framework for NLVL is a promising direction.

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Agency for Science, Technology and Research
SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Thank You!

Code at: https://github.com/IsaacChanghau/VSLNet