Day-10 (Speaker 2)

Speaker: Prof Mohit Bansal, University of North Carolina.

Title: Knowledgeable & Spatio Temporal Vision + Language

Video-based Dynamic Spatial-Temporal Knowledge

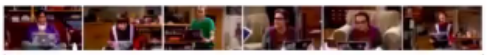
- Joint video+language understanding tasks, where models need to perform complex cross-modal reasoning on dynamic spatio-temporal information:
  - Video+Language Question Answering (TVQA)
  - Video+Language Question Answering with Spatial Grounding (TVQA+)
  - Video+Language Moment Retrieval (TVR)
  - Multilingual Video+Language Moment Retrieval (mTVR)
  - Video+Language Next-Event Prediction (VLEP)
  - VALUE Benchmark

TVQA (videos with audio and subtitles)

- Largest video-QA dataset with 6 video categories/genres, videos+subtitles QA, **compositional**, spatio-temporal **localization** (timestamps + bounding boxes)

[Lei et al., EMNLP 2018]

TVQA Compositionality (Localization + VQA)



0s 62s

Write a question:

[What/Why/...] ___ [when/before/after] ___

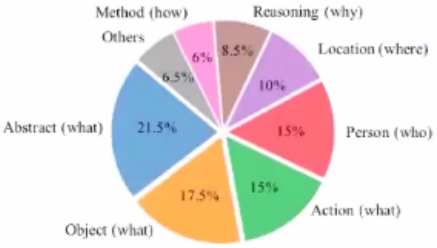
Question + Localization

What is Sheldon holding when
he is talking to Howard about swords?

[Lei et al., EMNLP 2018]

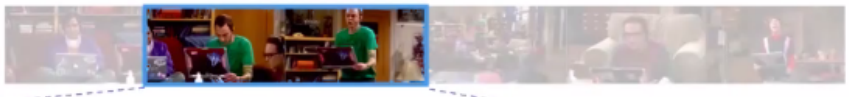
TVQA Data Statistics

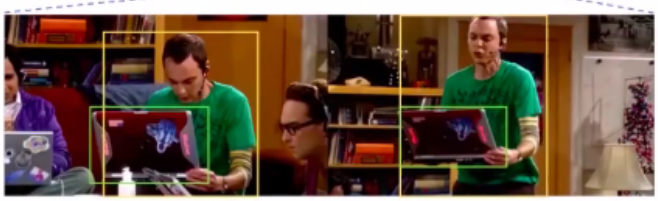
Show	Genre	#Sea.	#Epi.	#Clip	#QA
BBT	sitcom	10	220	4,198	29,384
Friends	sitcom	10	226	5,337	37,357
HIMYM	sitcom	5	72	1,512	10,584
Grey	medical	3	58	1,427	9,989
House	medical	8	176	4,621	32,345
Castle	crime	8	173	4,698	32,886
Total	—	44	925	21,793	152,545



Dataset	V. Src.	QType	#Clips / #QAs	Avg. Len.(s)	Total Len.(h)	Q. Src. text	Q. Src. video	Timestamp annotation
MovieFIB	Movie	OE	118.5k / 349k	4.1	135	✓	-	-
Movie-QA	Movie	MC	6.8k / 6.5k	202.7	381	✓	-	✓
TGIF-QA	Tumblr	OE&MC	71.7					-
Pororo-QA	Cartoon	MC	16.1					-
TVQA (our)	TV show	MC	21.8k / 152.5k	76.2	461.2	✓	✓	✓

TVQA+ (spatial localization+explainability)





00:02.314 → 00:06.732
Howard: Sheldon, he's got Raj. Use your sleep spell. Sheldon! Sheldon!

00:06.902 → 00:10.992
Sheldon: I've got the Sword of Azeroth.

Question: What is Sheldon holding when he is talking to Howard about the sword?
Correct Answer: A computer.

https://tvqa.cs.unc.edu/

[Lei et al., ACL 2020]

TVR: Text+Video Moment Retrieval

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.
Query Type: video + subtitle

https://tvr.cs.unc.edu/

[Lei et al., ECCV 2020]

TVR: Text+Video Moment Retrieval

Video Corpus Moment Retrieval (VCMR)

- A query + A video corpus ➡ Retrieve the matched moment from the corpus.
  - Retrieve the GT video. (Video Retrieval)
  - Localize the moment from the retrieved video. (Single Video Moment Retrieval)

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.
Query Type: video + subtitle

[Lei et al., ECCV 2020]

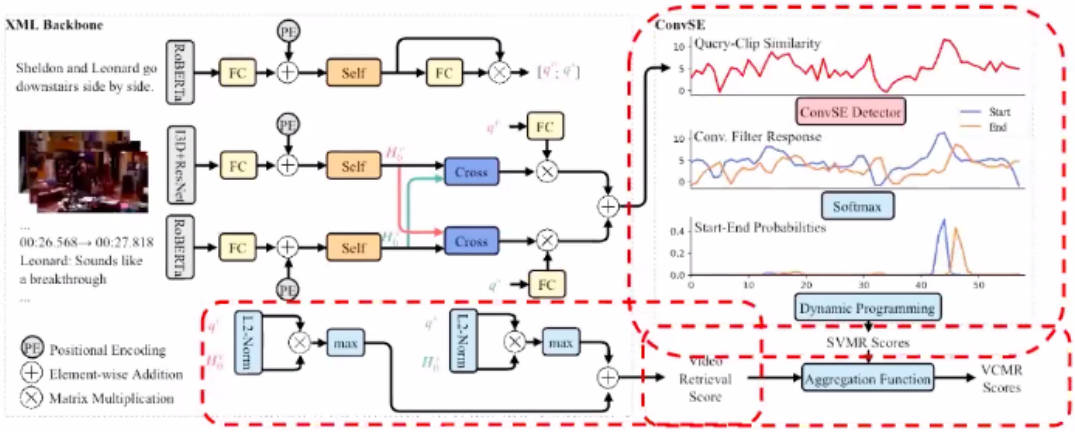
TVR: Text+Video Moment Retrieval

Dataset	Domain	#Q/#videos	Vocab. size	Avg. Q len.	Avg. len. (s) moment/video	Q context video	Q context text	Free st-ed	Q type anno.	Individual Q
TACoS [28]	Cooking	16.2K / 0.1K	2K	10.5	5.9 / 287	✓	-	✓	-	-
DiDeMo [13]	Flickr	41.2K / 10.6K	7.6K	8.0	6.5 / 29.3	✓	-	-	-	✓
ActivityNet Captions [21]	Activity	72K / 15K	12.5K	14.8	36.2 / 117.6	✓	-	✓	-	-
CharadesSTA [8]	Activity	16.1K / 6.7K	1.3K	7.2	8.1 / 30.6	✓	-	✓	-	-
TVR	TV show	109K / 21.8K	57.1K	13.4	9.1 / 76.2	✓	✓	✓	✓	✓

- TVR is the largest moment retrieval dataset, containing 109K human annotated query-moment pairs on 22K videos.
- It is the only moment-retrieval dataset requiring both video and subtitle context and with query type annotations.
- It also has a much larger vocab size, making textual understanding challenging.

[Lei et al., ECCV 2020]

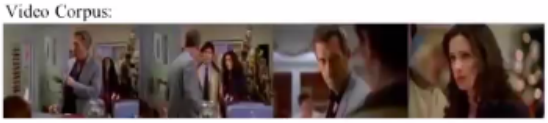
TVR: Text+Video Moment Retrieval



[Lei et al., ECCV 2020]

mTVR: Multilingual Video-Subtitle Moment Retrieval

Video Corpus:



00:00,327 → 00:04,320
Whitney: This is my fiancé…
惠特尼：这是我的未婚夫…

00:32,192 → 00:34,626
House: Nine months later, a miracle…
豪斯：9个月之后，一个奇迹…

00:03,897 → 00:07,731
Ross: Somebody seems to be…
罗斯：有人在…

00:36,497 → 00:38,761
Rachel: Call me when you get this.
瑞秋： 听到留言请回电。

00:07,786 → 00:13,156
Monica: Who wasn't invited…
莫妮卡：还没有被邀请到…

00:44,223 → 00:52,929
Rachel: Daddy, I can't marry him…
瑞秋：爸爸，我不能嫁给他…

Query:
Rachel explains to her dad on the phone why she can't marry her fiancé.
瑞秋在电话里向她父亲解释了她不能和其未婚夫结婚的原因。

Query Type: video + subtitle

[Lei et al., ACL 2021]

VLEP: Video-and-Language Next-Event Prediction

- Given a video (with dialogue) as premise, predict what is most likely to happen next by selecting from two provided future events. This task requires using commonsense knowledge, which is quite challenging for modern AI systems.

Premise Event

Premise Summary : A woman with a white shirt with black buttons grinds fruit slush in a blender.

Future Events

(Which event is more likely to happen right after the premise?)

A. The woman in the white shirt pours the slush into a cup.

Rationale: Slushy drinks are more commonly served in a cup, but there are hollowed out watermelon rinds sitting around the blender.

B. The woman in the white shirt pours the slush into a watermelon rind and passes it to Mark.

Rationale: There are hollowed out watermelon rinds sitting around the blender.

https://github.com/jayleicn/VideoLanguageFuturePred

[Lei et al., EMNLP 2020]

VLEP: Video-and-Language Next-Event Prediction

- Given a video (with dialogue) as premise, predict what is most likely to happen next by selecting from two provided future events. This task requires using commonsense knowledge, which is quite challenging for modern AI systems.

Premise Event

Premise Summary: The man being questioned refe finding the cell phone in the evidence bag and there being a text on it. Detective Beckett reaches toward the evidence bag.

Future Events

(Which event is more likely to happen right after the premise?)

A. Beckett takes the phone and reads the text.

Rationale: Dean mentioned a text on the phone, Beckett has reached toward the evidence bag.

B. Beckett picks up the phone and hands it to Dean.

Rationale: The detective probably wouldn't hand a piece of evidence to a suspect.

[Lei et al., EMNLP 2020]

VLEP: Video-and-Language Next-Event Prediction

- Human Annotation over 2 rounds:
  - Round 1: Standard Collection
  - Round 2: Human-and-Model-in-the-Loop Adversarial Data Collection.

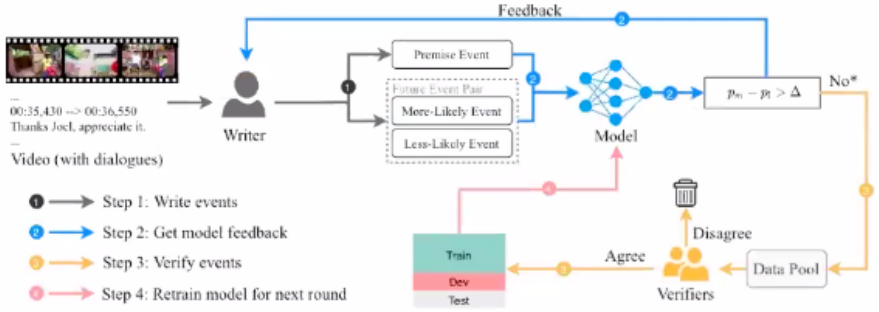


Illustration of our adversarial data collection procedure. p_m and p_l are the probabilities of the more-likely and the less-likely event being happening, respectively. Δ is a hyperparameter that controls how hard we want the collected example to be, it also helps to reduce prediction noises from imperfect models.

[Lei et al., EMNLP 2020]

VLEP: Video-and-Language Next-Event Prediction

- We collected 28.7K examples with 10K TV show and YouTube Vlog video clips from different genres. We also show top unique verbs in each genre.

Domain	Genre	#Shows (#Channels)	#Videos	#Examples
TV show	Sitcom	3	4,117	12,248
	Medical	2	1,558	5,198
	Crime	1	1,072	4,306
YouTube Vlogs	Travel, Food	6	2,406	4,812
	Family, Daily	3	1,081	2,162
Total	-	15	10,234	28,726

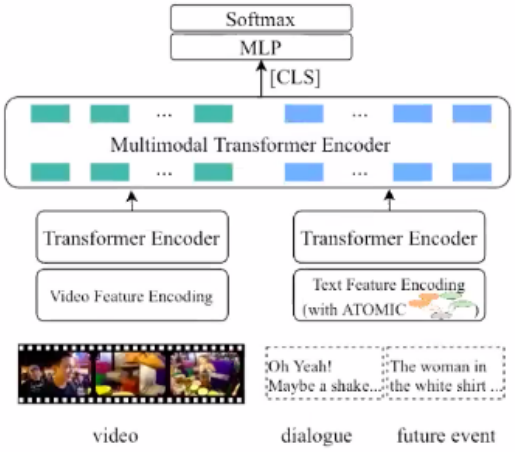
Data statistics by genre.

Genre	Top Unique Verbs
Sitcom	change, offer, hear, should, accept, yell, hang, join, apologize, shut, shout, realize
Medical	die, treat, cry, yell, smile, proceed, examine, approach, argue, save, admit, rush
Crime	kill, shoot, point, question, toss, hang, remove, catch, lie, deny, investigate,
Travel, Food	taste, add, pour, dip, cook, describe, cut, order, serve, stir, prepare, enjoy, buy
Family, Daily	drive, jump, wear, point, smile, touch, climb, dress, set, swim, hide, lay, blow

Top unique verbs in each genre.

[Lei et al., EMNLP 2020]

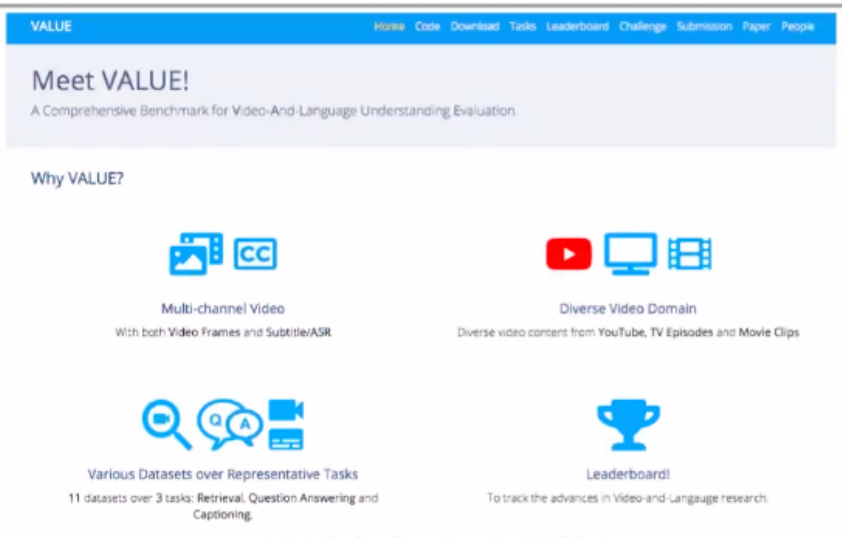
VLEP: Video-and-Language Next-Event Prediction



Model	Accuracy (%)
chance	50.00
future only	58.09
video + future	59.03
dialogue + future	66.63
video + dialogue + future	67.46
human (dialogue + future)	76.25
human (video + dialogue + future)	90.50

[Lei et al., EMNLP 2020]

VALUE Benchmark



https://value-benchmark.github.io/

[Li + Lei et al., ArXiv 2021]

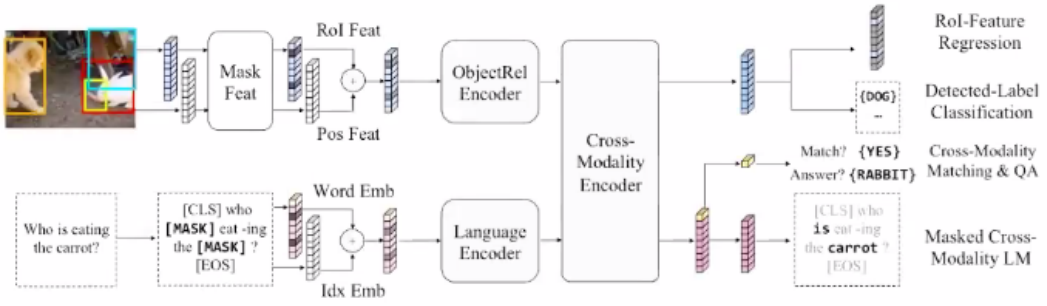
Language Models & Vision-Language Pretraining Knowledge

Language Models & V&L Pretraining Knowledge

- Enhancing large-scale pretrained language models with different methods/types of visual grounding for improved NLU/NLG/Multimodal/Robotic tasks:
  - Large-Scale Cross-Modal V&L Pre-Training (LXMERT)
  - Video+Language Pre-Training from Noisy Videos+ASR (DeCEMBERT)
  - Video+Language Efficient Pre-Training via Sparse Sampling (ClipBERT)
  - Unifying V&L Pre-Training via Text Generation (VL-T5)
  - Improving NLU with Visually-Grounded Supervision (Vokenization)
  - Language Model Commonsense for Incomplete Robotic Instructions

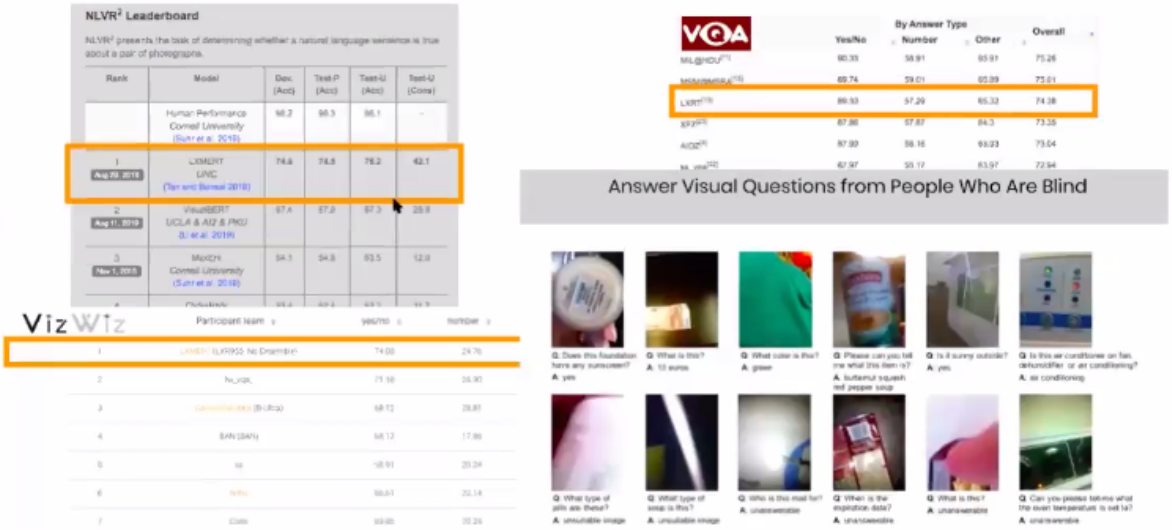
Large-Scale Cross-Modal Pretraining Knowledge: LXMERT

- LXMERT brings in knowledge from text, vision and cross-modal matching sides: vision-lang transformers with 3 encoders (object relations, language, cross-modal) & 5 pretraining tasks: masked-LM, masked-Object-Prediction (feature regression+label classification), cross-modality matching, image-QA.



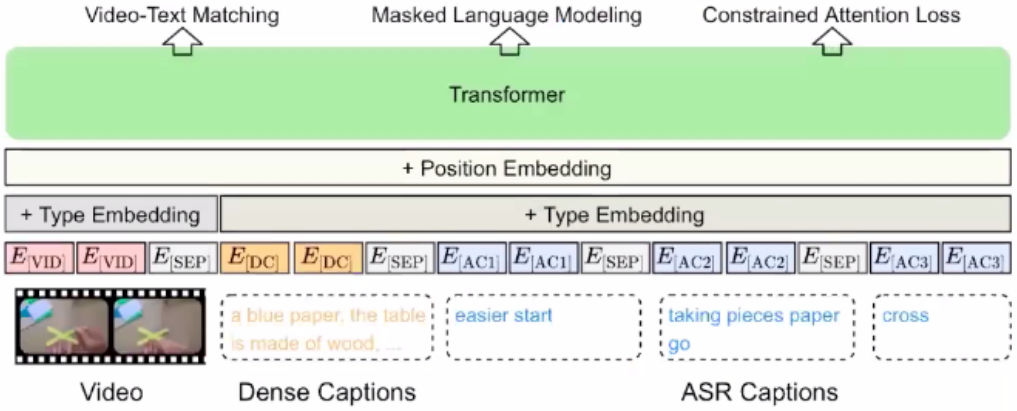
[Tan and Bansal, EMNLP 2019]

Large-Scale Cross-Modal Pretraining Knowledge: LXMERT



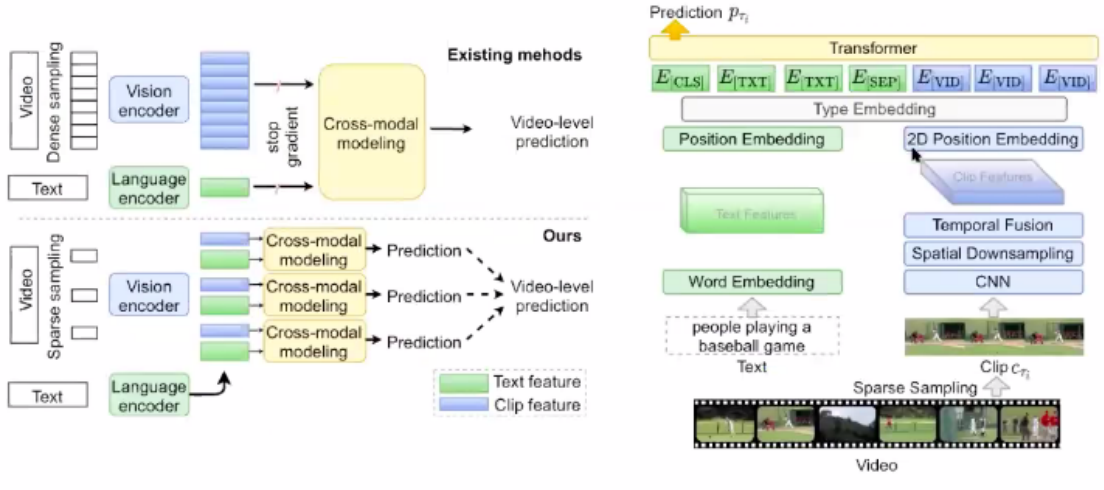
[Gurari et al., 2018; Bigham et al., 2010; Suhr et al., 2019; Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019]

DeCEMBERT



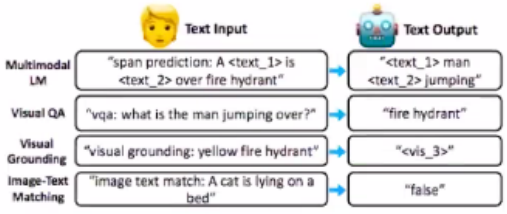
[Tang + Lei et al., NAACL 2021]

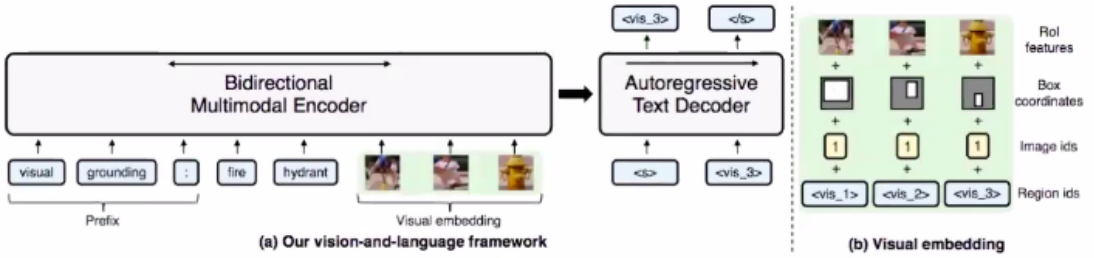
ClipBERT



[Lei + Li et. al, CVPR 2021]

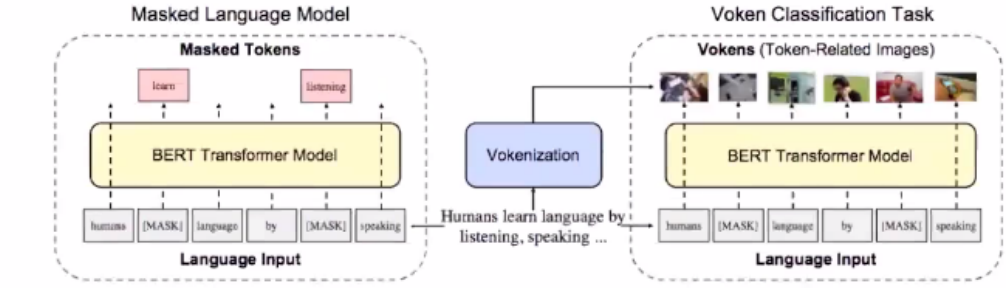
VL-T5: Unifying Vision-Language Tasks via Text Generation

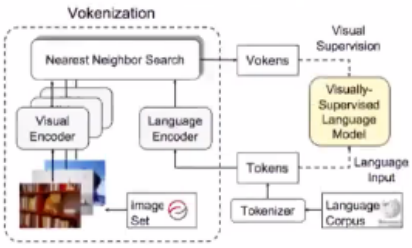




[Cho et al., ICML 2021]

Vokenization: Contextualized Visually-Grounded Distant Supervision





Method	SST-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
$\text{BERT}_{6L/512H}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
$\text{BERT}_{6L/512H}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
$\text{BERT}_{12L/768H}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
$\text{BERT}_{12L/768H}$ + Voken-cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
$\text{RoBERTa}_{6L/512H}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
$\text{RoBERTa}_{6L/512H}$ + Voken-cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
$\text{RoBERTa}_{12L/768H}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
$\text{RoBERTa}_{12L/768H}$ + Voken-cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

[Tan and Bansal, EMNLP 2020]

Some Concluding Thoughts and Next Steps

- Long-distance understanding in videos?
- Continual learning of new/unseen video information coming in?
- Strengths vs limitations of large-scale BERT/LXMERT pretraining?
- Structured/modular knowledge vs large-scale pretraining?
- Bringing in other (non-verbal) modalities?
- Longer ambiguities and interaction/clarifications for robotic tasks?