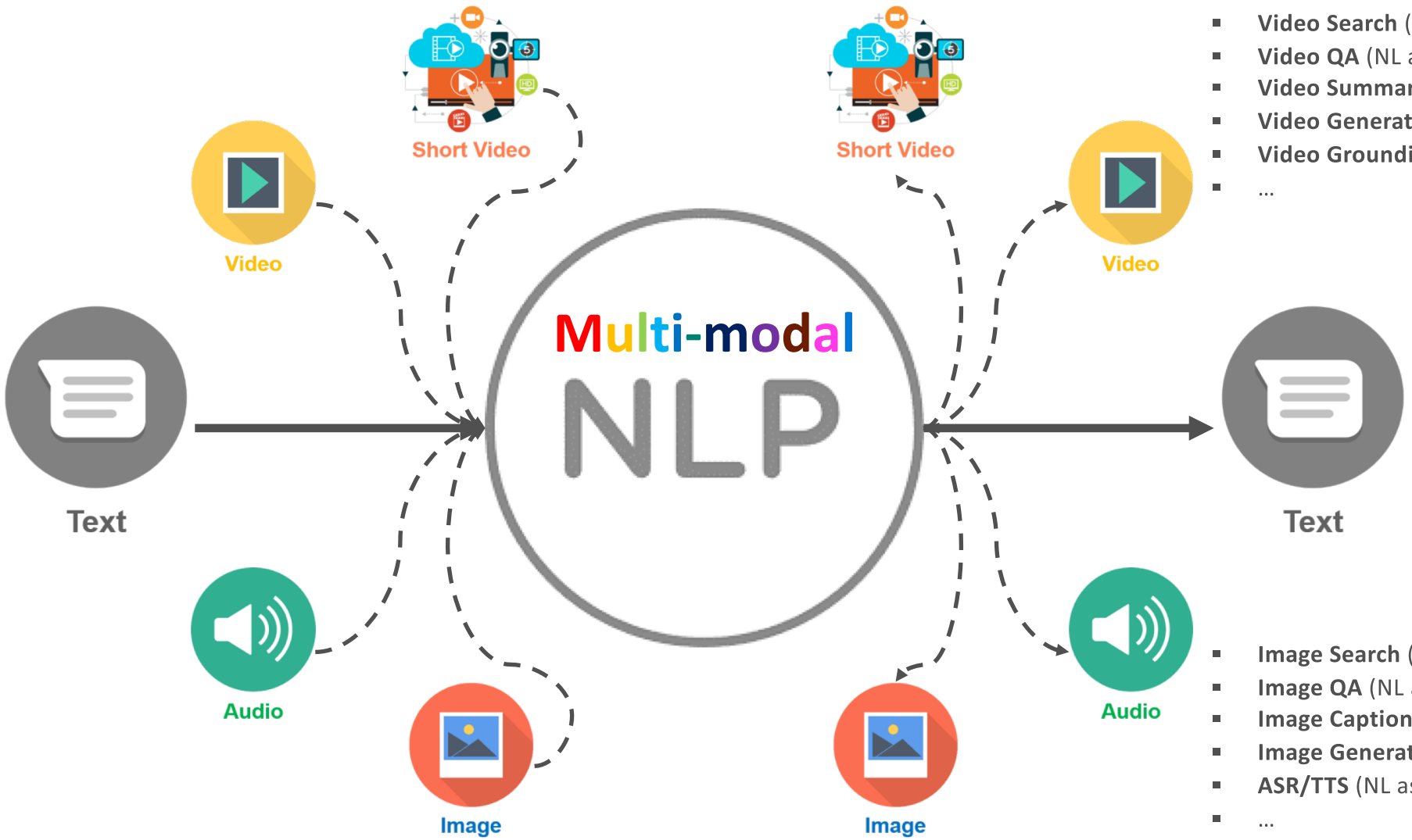
Microsoft

When Language Meets Vision: Multi-modal NLP with Visual Contents

Dr. Nan DUAN (段楠)
Natural Language Computing Group
Microsoft Research Asia
2019
(Joint work with Chenfei WU, Gen LI, Duyu TANG, Yanzhao ZHOU, Lei JI, Botian SHI, Pan LU, Yikang LI, Ming ZHOU)

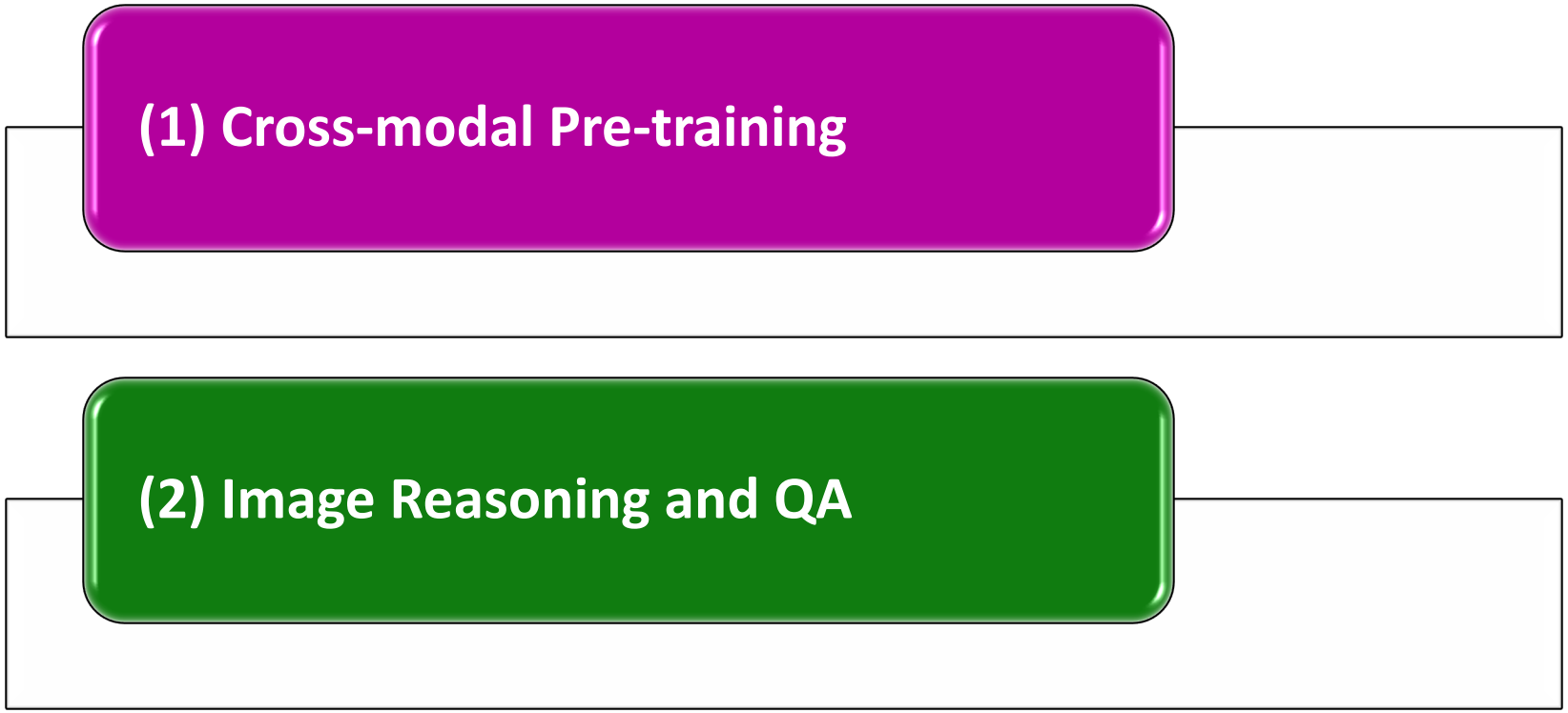
Multi-modal NLP: to solve language-centric multi-modal tasks



- **Video Search** (NL as input)
- **Video QA** (NL as input)
- **Video Summarization** (NL as output)
- **Video Generation** (NL as input)
- **Video Grounding** (NL as input)
- …

- **Image Search** (NL as input)
- **Image QA** (NL as input)
- **Image Captioning** (NL as output)
- **Image Generation** (NL as input)
- **ASR/TTS** (NL as input/output)
- …

Our Work on Multi-modal NLP (Today's Agenda)



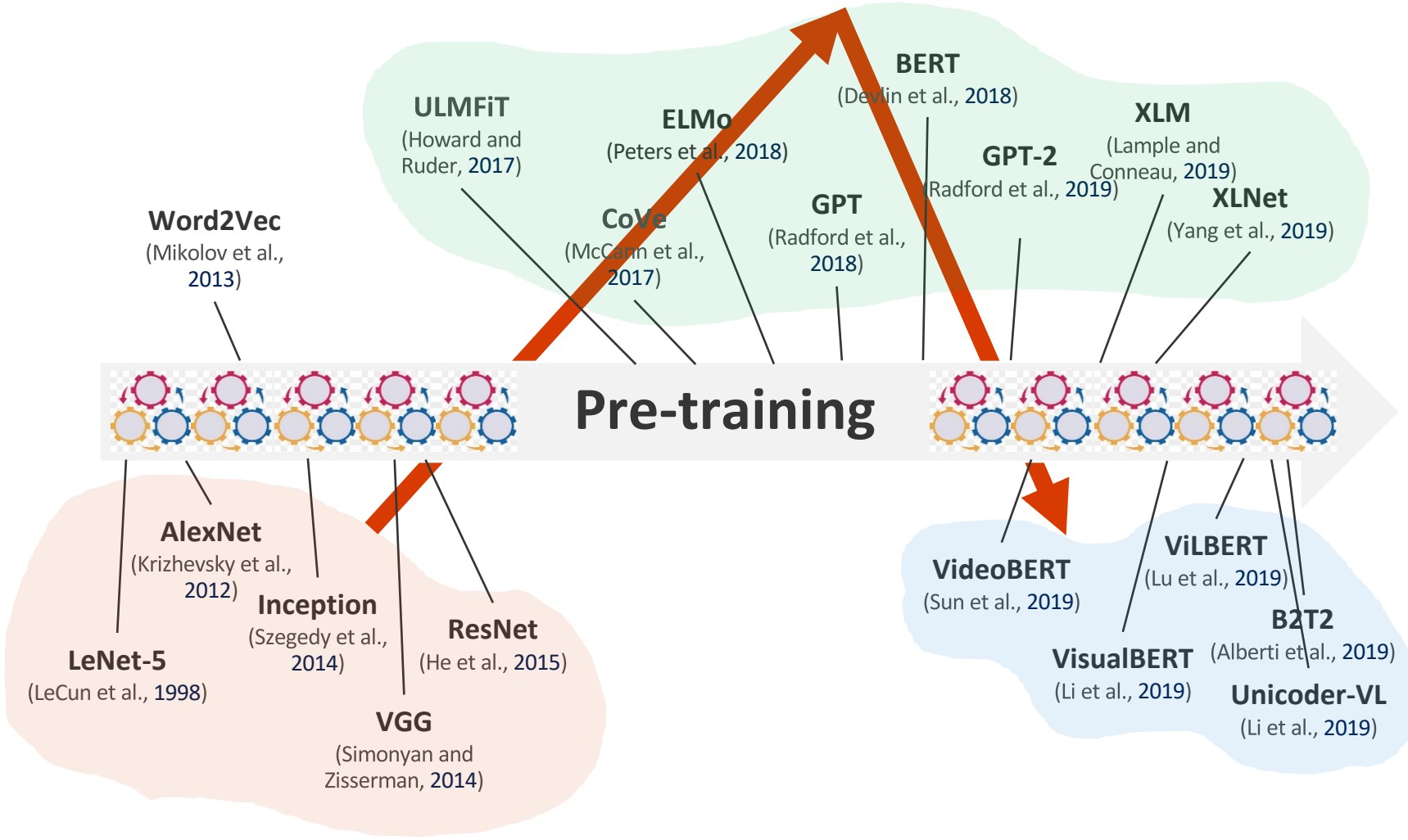
Pre-training Models in NLP and CV: Reinforce Each Other

Natural Language Data

IMAGENET

YouTube

Image/Video Data

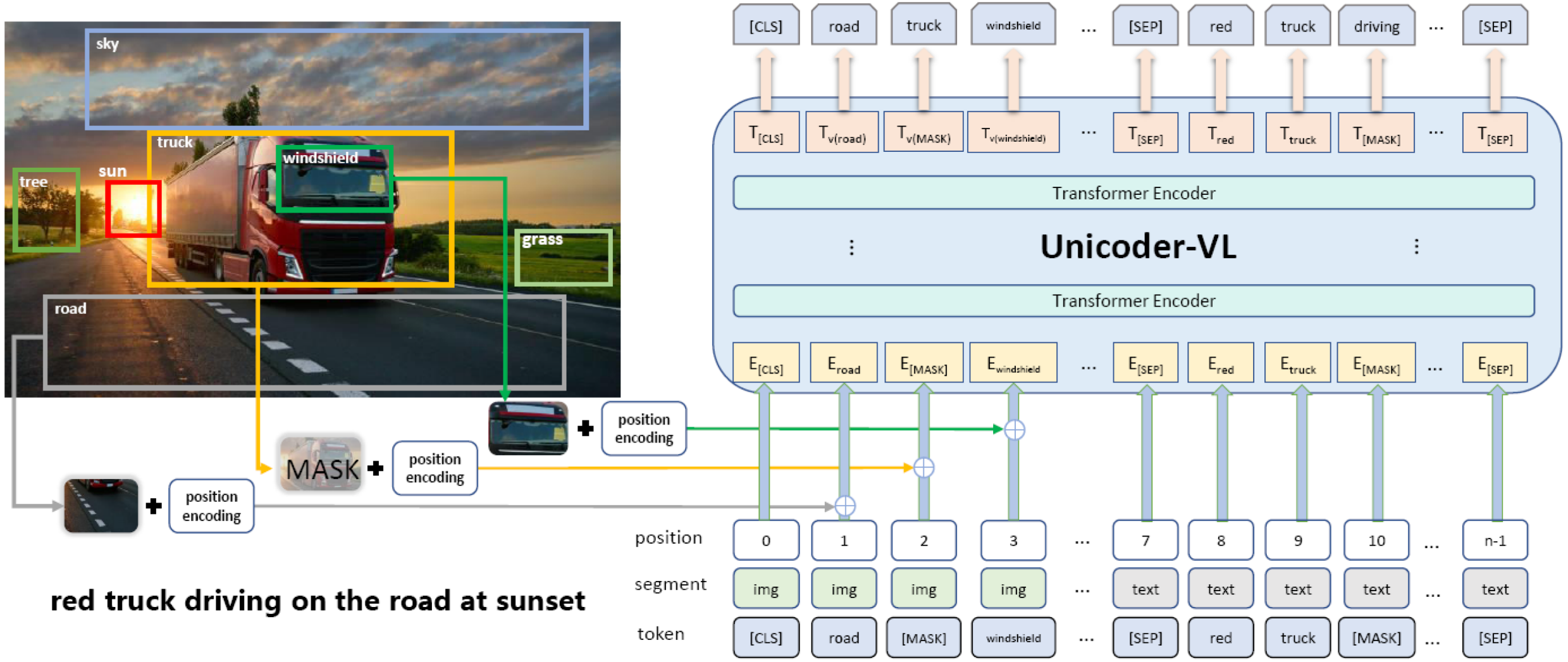


NLP/CV Tasks

- Machine Translation
- Question Answering
- Text Summarization
- Semantic Parsing
- ... ...
- Image Classification
- Image QA
- Image Retrieval
- Video Captioning

Transfer general knowledge learnt from large-scale datasets to downstream NLP/CV Tasks.

(1) Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training



Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, Ming Zhou. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. arXiv, 2019.

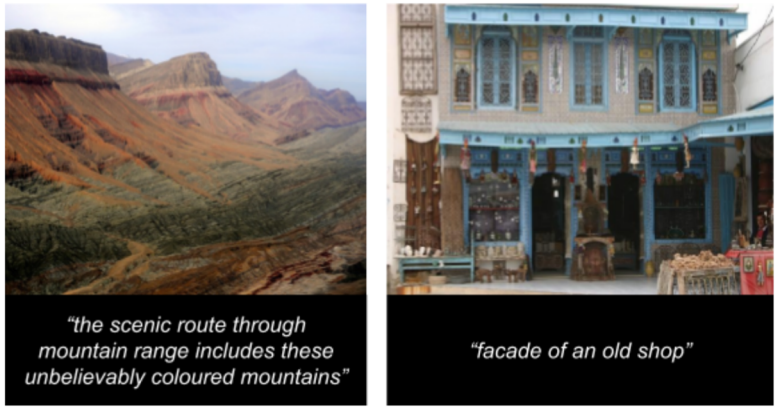
Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. **Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks**. EMNLP, 2019.

State-of-the-Art on Image Retrieval !

MSCOCO	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Two Branch Network (Wang et al., 2018)	43.3	76.8	87.6	54.0	84.0	91.2
SCAN (Lee et al., 2018)	58.8	88.4	94.8	72.7	94.8	98.4
Scene Concept Graph (Shi et al., 2019)	61.4	88.9	95.1	76.6	96.3	**99.2**
Unicoder-VL (Ours)	**69.1 (+7.7)**	**93.4**	**97.2**	**85.4 (+8.8)**	**97.5**	99.0

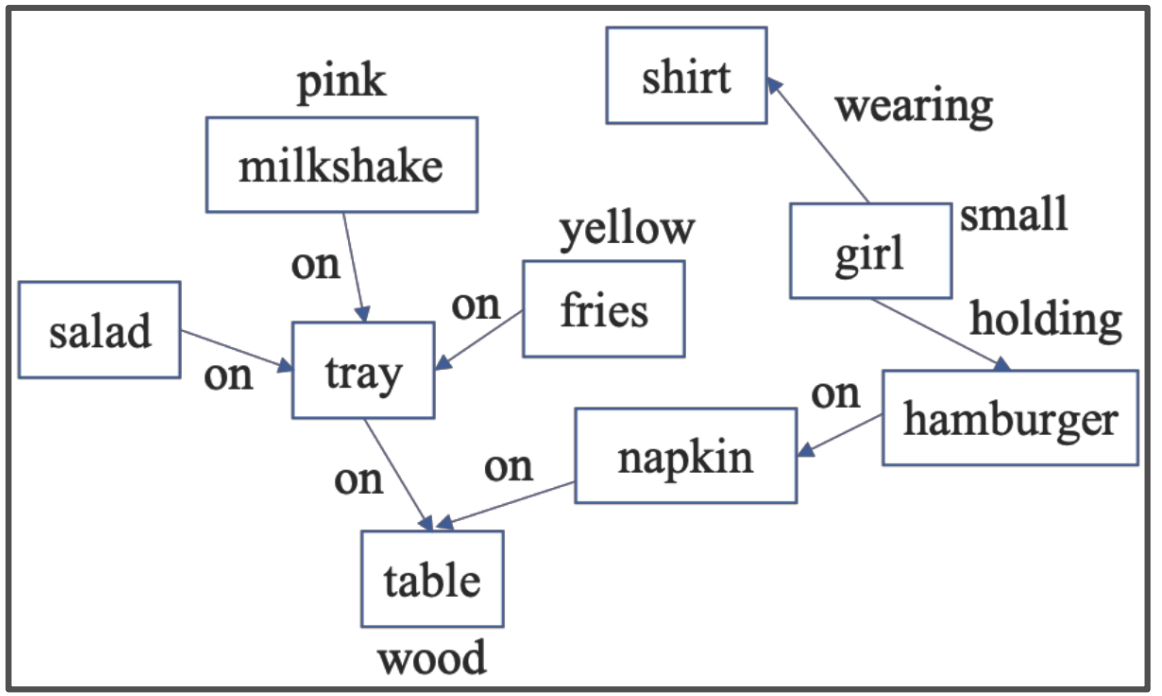
Flickr30K	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Two Branch Network (Wang et al., 2018)	28.4	56.3	67.4	37.5	64.7	75.0
SCAN (Lee et al., 2018)	48.6	77.7	85.2	67.4	90.3	95.8
Scene Concept Graph (Shi et al., 2019)	49.3	76.4	85.6	71.8	90.8	94.8
Unicoder-VL (Ours)	**68.5 (+19.2)**	**90.5**	**94.9**	**83.0 (+11.2)**	**95.9**	**98.0**

- Pre-training dataset
  - 3,318,333 image-caption pairs from Google's Conceptual Captions



- Our findings/thoughts
  1. **Transformer** architecture works in cross-modal pre-training as well.
  2. **Faster R-CNN** is better than ResNet in cross-modal pre-training.
  3. **Adding mono-modality tasks** into pre-training is **non-trivial**.
  4. **Image encoder** should be re-designed for language-vision tasks.

(2) Image Reasoning and QA

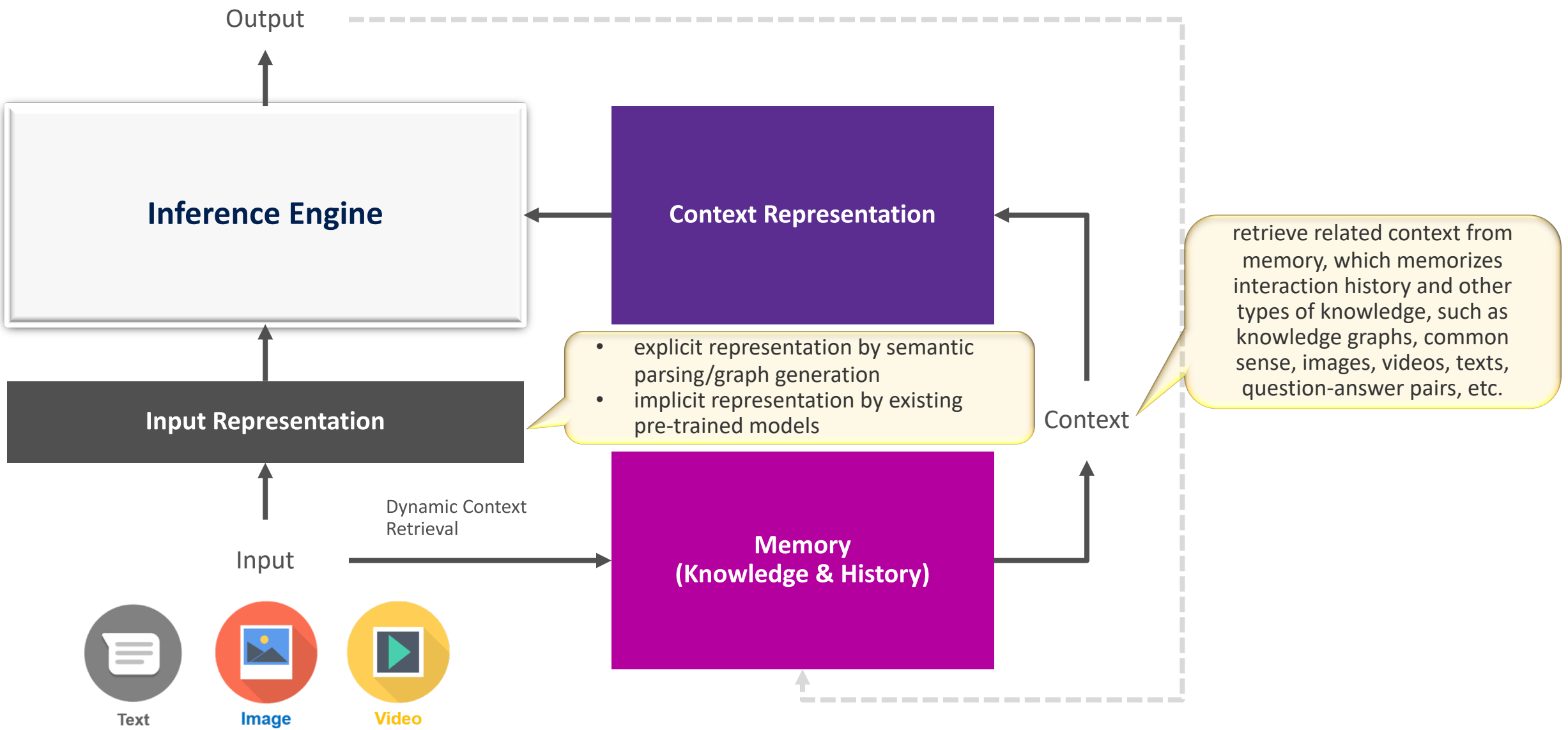


What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

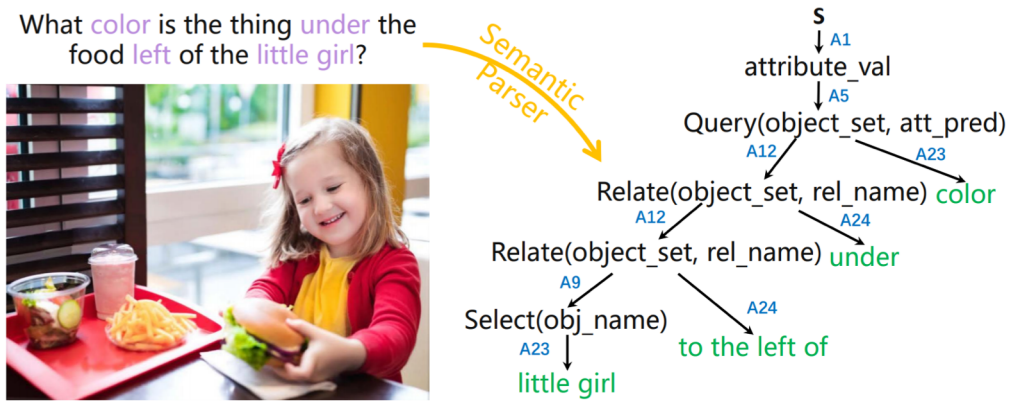
Select: hamburger → **Relate:** girl, holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food, on → **Choose** color: yellow | brown

Hudson and Manning. **GQA: a New Dataset for Compositional Question Answering over Real-World Images**. 2019.

Dynamic **REA**soning **M**achine (**DREAM**)



DREAM for GQA: Question Semantic Parsing



	LF Exact Match
Seq2Seq	77.4%
Seq2Action	**85.6%**

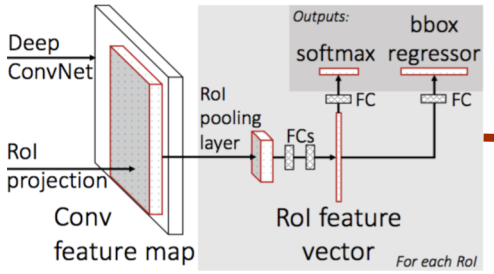
Semantic Parsing Result on GQA Questions

Action	Operation	Description
A1-A4	S->attribute_val | attribute_pred | relation_name | bool	Possible types of answer
A5	attribute_val->Query(object_set, att_pred)	Get attribute of object
A6	attribute_val->Chooseatt(object_set, att_pred, att_val, att_val)	Choose between attribute values
A7	attribute_pred -> Common(object_set set)	Get the predicate that the all objects have same value
A8	relation_name -> Chooserel(object_set, obj_name, rel_name, rel_name)	Choose between relation names
A9	object_set -> Select(obj_name)	Get objects with name of 'name' from the scene graph
A10	object_set -> Filter(object_set, att_pred?, att_val)	Subset of objects that have attribute 'pred' equals 'val'
A11	object_set -> Group(object_set, object_set)	Union of two object sets
A12	object_set -> Relate(object_set, rel_name)	Get object that is linked with the first object of 'set' with relation 'rel'
A13	boolean -> Logic(lg_op, boolean, boolean)	Logical operation
A14	boolean ->Exist(object_set)	True if set is not empty
A15	boolean ->Verifyatt(object_set, att_pred?, att_val)	True if objects have attribute 'pred' equals 'val'
A16	boolean ->Verifyrel(object_set, obj_name, rel_name)	True if the first object of 'set' have the relation with object
A17	boolean ->Same(object_set, att_pred)	True if objects have same value for the attribute 'pred'
A18	boolean ->Different(object_set, att_pred)	True if objects do not have same value for the attribute 'pred'
A19-20	lg_op -> And | Or	Logical types
A21-A24	att_pred, att_val, obj_name, rel_name -> instantiation	Instantiation

Daya Guo, Duyu Tang, Nan Duan, Jian Yin, Ming Zhou. **Dialog-to-Action: Conversational Question Answering over a Large-Scale Knowledge Base**. NeurIPS, 2018.

DREAM for GQA: Dynamic Context Retrieval

1 detected objects



Faster R-CNN

girl 0.91; **table** 0.87; **salad** 0.73; **fries** 0.95; **hamburger** 0.74; **beverage** 0.69; …

knowledge extension by common sense

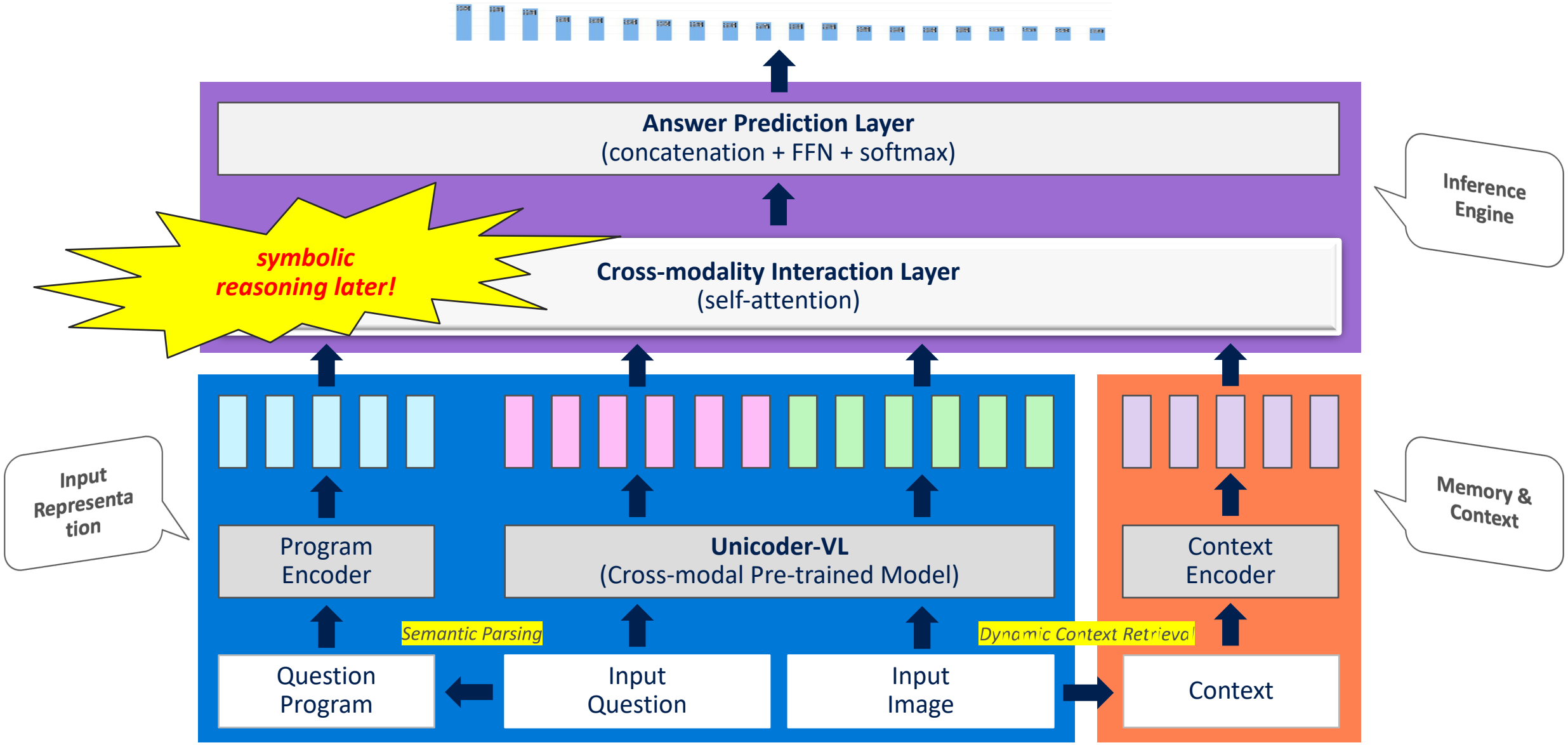
2 extended triples

table is wood
fries is yellow
hamburger contains vegetables
napkin is white
beverage can drink
…

- **what color are the chips?**
- **what is located to the left of the fries?**
- **where is the tray located in?**
- **what color is the thing under the food left of the little girl with the yellow shirt?**
- **…**

3 Co-click queries of the same image

DREAM for GQA: Inference Engine



Chenfei Wu, Gen Li, Yanzhao Zhou, Duyu Tang, Nan Duan. **DREAM: Dynamic REAsoning Machine for GQA**. Will submit to AAAI, 2020.

State-of-the-Art on Visual QA !

B - Baseline submission

Rank	Participant team	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy	Last submission at
1	Human Performance (human)	91.2	87.4	98.4	97.2	98.9	0	89.3	5 months ago
2	DREAM (MSRA)	80.538	68.603	91.987	83.755	96.422	3.678	74.198	27 days ago
3	Kakao Brain	79.675	67.726	77.021	83.704	96.360	2.458	73.328	2 months ago
4	270	77.502	63.819	86.941	83.774	96.653	1.486	70.234	2 months ago
5	NSM ensemble (updated)	80.447	56.161	93.825	84.161	96.534	2.784	67.547	1 month ago
6	NSM single (updated)	78.938	49.250	93.251	84.283	96.407	3.705	63.168	1 month ago
7	LXRT (LXR955, Ensemble)	79.790	47.641	93.095	85.211	96.360	6.419	62.713	2 months ago
8	DL-61 (GRN)	78.688	45.809	90.313	85.434	96.356	6.773	61.223	3 months ago
9	MSM@MSRA	77.839	46.300	88.916	85.494	96.427	5.678	61.086	2 months ago
10	SK T-Brain (Ensemble10)	79.122	44.763	92.613	85.631	96.355	8.558	60.871	2 months ago
11	snudm-clovaai (Musan)	79.095	43.019	93.721	85.923	96.410	10.100	59.932	2 months ago
12	PVR (PVR)	78.017	43.750	91.433	84.774	96.500	6.002	59.815	2 months ago
13	rishabh_test	77.533	43.348	88.627	84.708	96.177	6.057	59.375	4 months ago
14	USTB (glimple_all)	75.071	44.585	84.637	84.863	96.231	5.540	58.877	3 months ago
15	REAGQA (Partial-MSP)	77.391	41.672	90.293	84.534	95.574	7.859	58.418	2 months ago

GQA Leaderboard on 2019-07-13

https://evalai.cloudcv.org/web/challenges/challenge-page/225/leaderboard/733

DREAM for Document-level QA

Google AI Home Competition Leaderboard Visualization Download FAQ Sign In

Long Answer Leaderboard

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall	R@P = 90	R@P = 75	R@P = 50
1	BERT-dm_v2-ensemble	DREAM	Anonymous	6/25/19	0.74498	0.73691	0.75324	0.38539	0.73678	0.84317
2	bert-dm	DREAM	Anonymous	5/21/19	0.72613	0.72017	0.73218	0.31783	0.69884	0.82518
3	bert_dm	DREAM	Anonymous	5/13/19	0.7196	0.71952	0.71968	0.31191	0.68304	0.81268
4	bert-dm	DREAM	Anonymous	4/27/19	0.70248	0.70795	0.69708	0.23799	0.63632	0.805
5	BERT	LEE_SJ	LAIR	6/19/19	0.68803	0.66035	0.71814	0.23492	0.57118	0.78943

Short Answer Leaderboard

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall	R@P = 90	R@P = 75	R@P = 50
1	BERT-dm_v2-ensemble	DREAM	Anonymous	6/25/19	0.58689	0.62578	0.55256	0.11179	0.43118	0.61731
2	BERT-mnlp-ensemble	GAAMA	IBM Research AI	6/15/19	0.58088	0.63477	0.53542	0.17364	0.4277	0.59146
3	BERT	LEE_SJ	LAIR	6/19/19	0.57402	0.63531	0.52352	0.14866	0.42102	0.5842
4	bert-dm	DREAM	Anonymous	5/21/19	0.5699	0.62001	0.52729	0.0903	0.38647	0.58856

https://ai.google.com/research/NaturalQuestions/

DREAM for Commonsense QA and Fact Checking

(Contributor: **Duyu TANG** from MSRA-NLC)

Version 1.11 Random Split Leaderboard

(12,102 examples with 5 answer choices)

Model	Affiliation	Date	Accuracy
Human		03/10/2019	88.9
DREAM (single model)	Microsoft Research Asia and Bing	08/08/2019	66.9
CSR-KG (AI2 IR, single model)	Microsoft Dynamics 365 AI Research	07/19/2019	65.3
AristoBERTv7 (single model)	Aristo team at Allen Institute for AI	07/18/2019	64.6
BERT+OMCS (single model)	Allen Institute for Artificial Intelligence - Israel	07/30/2019	62.5
BERT+AMS (single model)	Alibaba DAMO Speech Lab	07/10/2019	62.2
CSR-KG (single model)	Microsoft Dynamics 365 AI Research	06/28/2019	61.8
BECON (ensemble)	Singapore University of Technology and Design	07/01/2019	59.6
KagNet (single model)	Anonymous	05/14/2019	58.9
CoS-E (single model)	Salesforce Research	04/12/2019	58.2

https://www.tau-nlp.org/csqa-leaderboard

FEVER

Fact Extraction and VERification (FEVER) Challenge

Organized by oana - Current server time: Aug. 9, 2019, 3:26 a.m. UTC

Previous	▶ Current	End
Competition (Blind Test Set Evaluation)	After Competition: Perpetual Evaluation (Test Set)	Competition Ends
July 24, 2018, midnight UTC	July 28, 2018, 7 a.m. UTC	Never

Learn the Details | Phases | Participate | Results

Before Competition (Development Set Evaluation) | Competition (Blind Test Set Evaluation) | After Competition: Perpetual Evaluation (Test Set)

Phase description

None

Max submissions per day: 2

Max submissions total: 1000

Download CSV

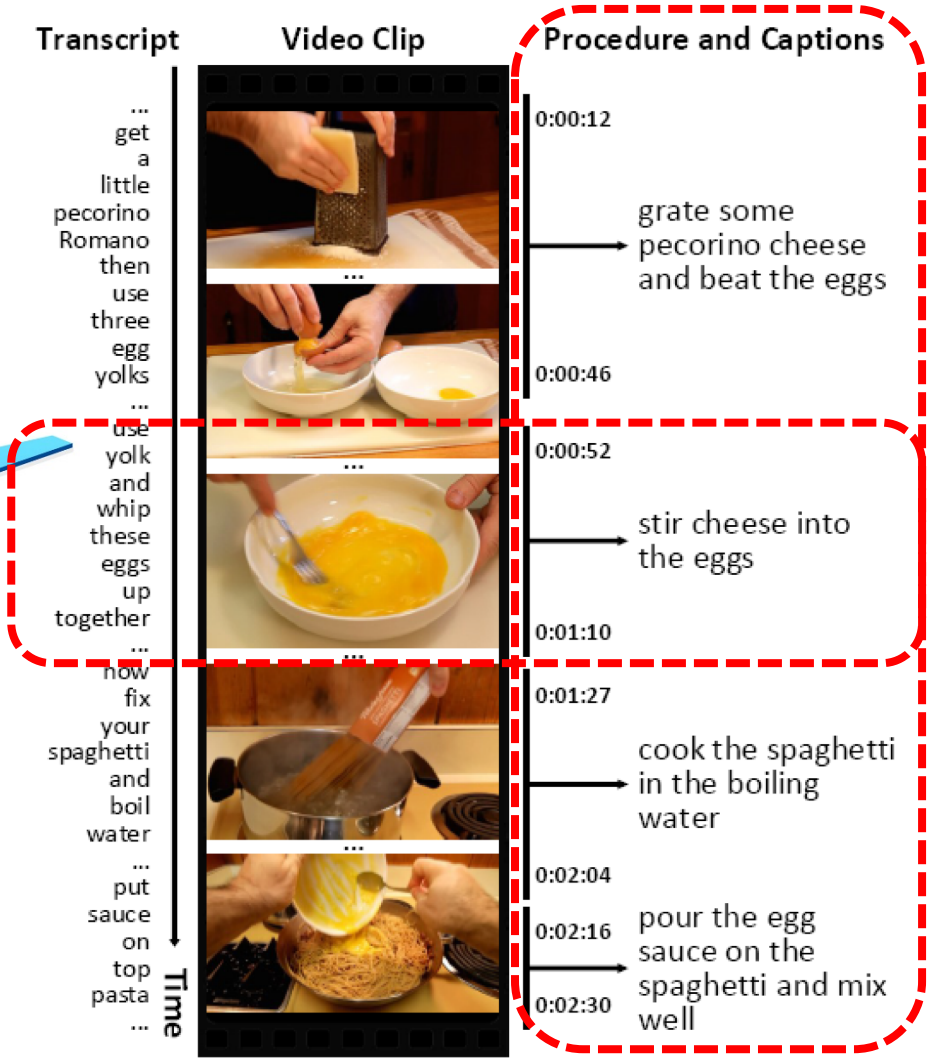
Results

#	User	Entries	Date of Last Entry	Team Name	Evidence F1 ▲	Label Accuracy ▲	FEVER Score ▲
1	DREAM	11	08/07/19	DREAM (MSRA+MSNews)	0.3920 (6)	0.7642 (1)	0.6962 (1)
2	abcd_zh	4	07/28/19		0.3914 (7)	0.7281 (2)	0.6940 (2)
3	cunlp	1	07/25/19		0.3765 (8)	0.7247 (3)	0.6880 (3)
4	dominiks	2	07/25/19		0.3626 (11)	0.7154 (6)	0.6846 (4)
5	a.soleimani.b	1	07/25/19		0.3619 (12)	0.7041 (7)	0.6836 (5)

https://competitions.codalab.org/competitions/18814#results

Beyond Image: Video Summarization and QA

- Motivation
  - **Video Summarization**: to make video contents ***readable***
  - **Video QA**: to make video contents ***searchable***
- Application
  - Query-1: *How to cook Spaghetti Carbonara?*
    Summarize the video content into short sentences.
  - Query-2: *How to make Spaghetti sauce?*
    Answer the query by returning the most related video clip.



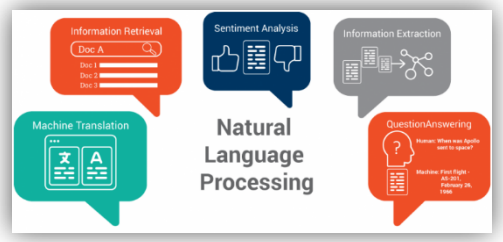
https://www.youtube.com/watch?v=dEBUJ6MZ6e0

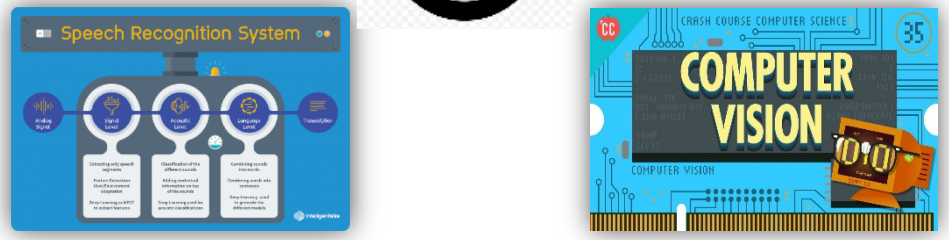
Botian Shi, Lei Ji, Yaobo Liang, Zhendong Niu, Nan Duan, Ming Zhou. **Dense Procedure Captioning in Narrated Instructional Videos**. ACL, 2019.

Kyungjae Lee, Nan Duan, lei ji, Jason Li, Seung-won Hwang, Ming Zhou. **Segment-then-Rank: Non-factoid Question Answering on Video Contents**. To appear in arXiv, 2019

Future Work: ***Multi-modal Fundamentals***

- Redesign of image/video encoders
- Cross-modal pre-training with images/videos
- Knowledge-based visual understanding/reasoning
- Language-to-Vision generation
- Fake content identification





TO enable NLP systems to look, listen, comprehend and reply!

Microsoft

Thank You!