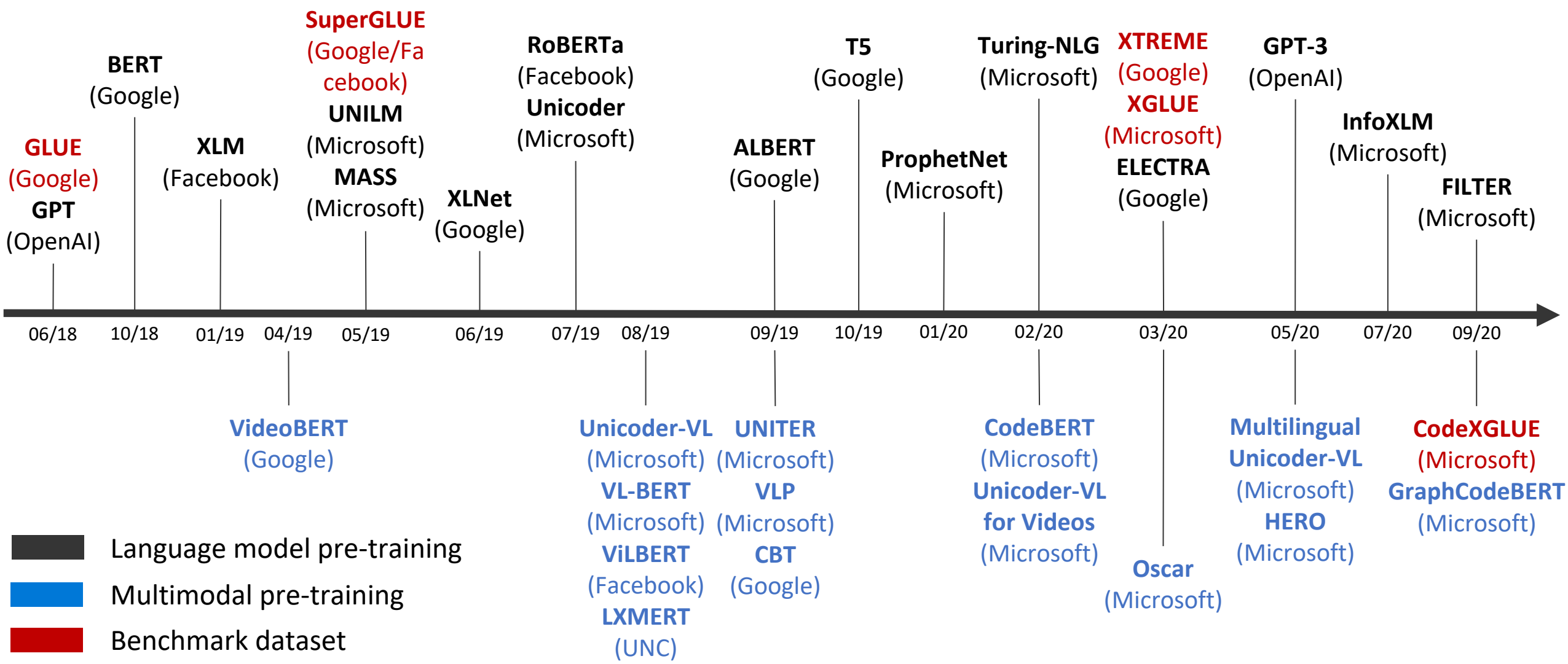
Multilingual Multimodal Pre-training

Nan DUAN
Principal Researcher
Microsoft Research Asia
2020-11-20

Pre-training: A New NLP Paradigm



Our Goal

Learn universal representations

- map objects occurred in different modalities or expressed in different languages to vectors in a common semantic space

1). Self-Supervised Learning

Self-Supervised Learning

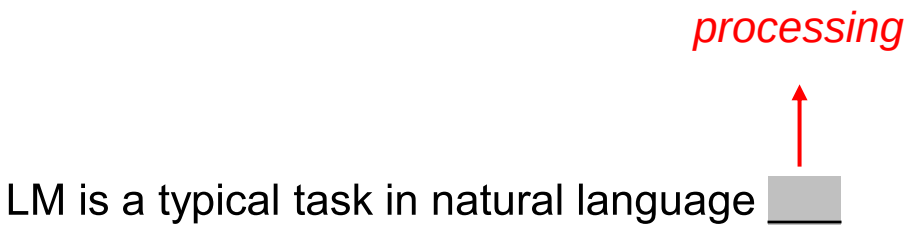
A form of unsupervised learning where the data itself provides the supervision.

(1) Auto-regressive; (2) Denoising Auto-encoding; (3) Contrastive Learning.

1. Auto-regressive

Maximize the likelihood under the forward auto-regressive factorization.

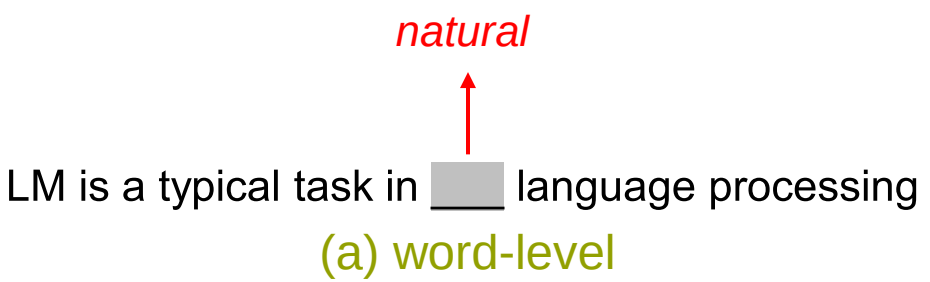
$$\max_{\theta} \mathbb{E}_{\mathbf{w}\sim D} \sum_{t=1}^{|\mathbf{w}|} \log p_\theta(w_t | w_{<t})$$



2. Denoising Auto-encoding

Reconstruct masked words from corrupted inputs.

$$\max_{\theta} \mathbb{E}_{\mathbf{w}\sim D} \log p_\theta(w_t|\mathbf{w}_{\setminus t})$$



(a) word-level

2. Denoising Auto-encoding

Reconstruct original inputs from corrupted inputs.

$$\max_{\theta} \mathbb{E}_{\mathbf{w}\sim D} \sum_{t=1}^{|\mathbf{w}|} \log p_\theta(w_t | w_{<t}, corrupt(\mathbf{w}))$$

LM is a typical task in natural language processing

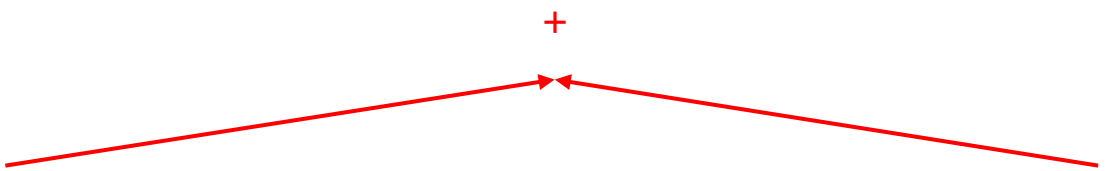
LM is a ~~typical~~ task in processing ___ language

(b) sentence-level

3. Contrastive Learning

Learn to compare via the Noise Constrative Estimation (NCE) objective.

$$\max_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{x}^+,\{\mathbf{x}^-\}_k \sim D} \log \frac{\exp(\mathrm{sim}(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}^+))/\tau)}{\exp(\mathrm{sim}(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}^+))/\tau) + \sum_k \exp(\mathrm{sim}(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}_k^-))/\tau)}$$



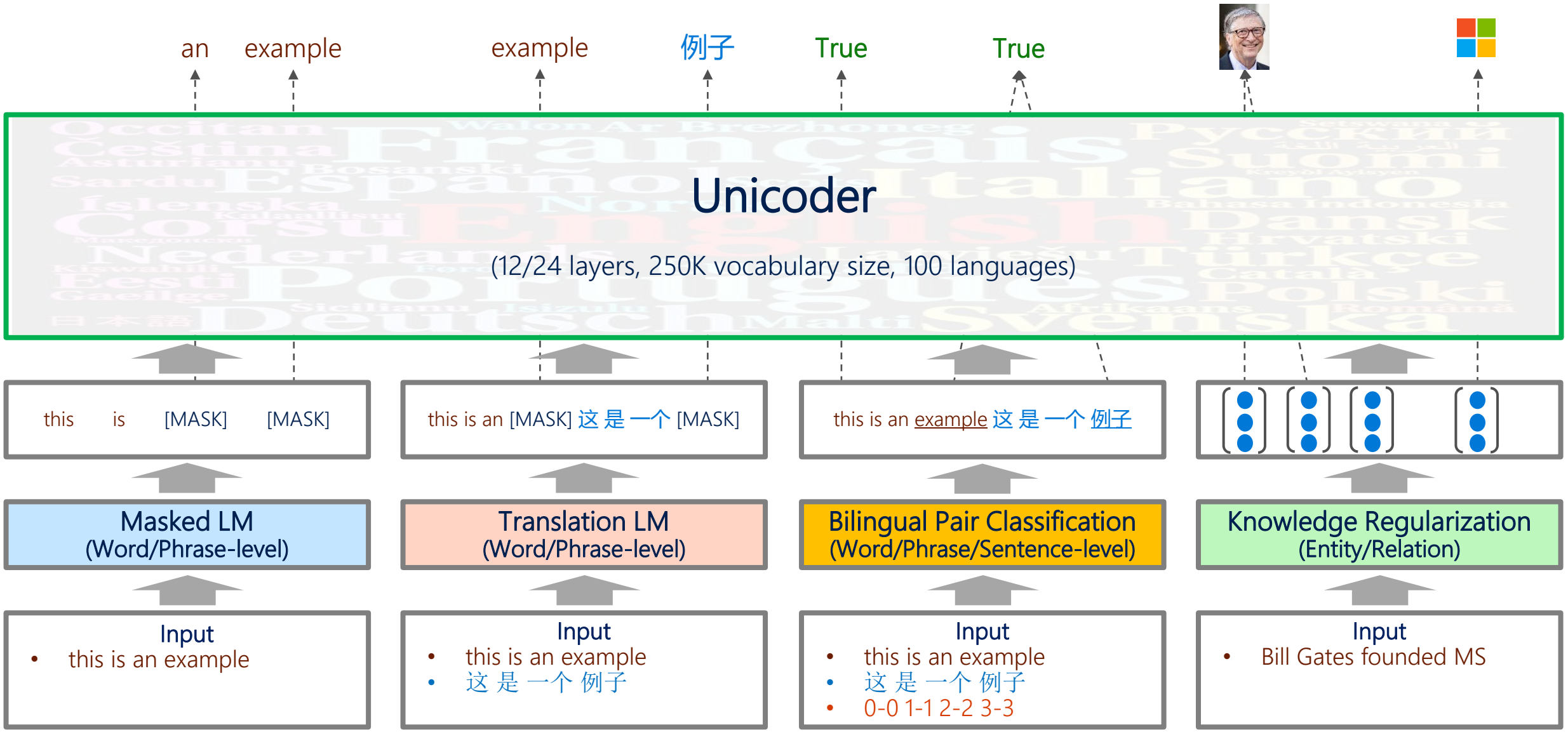
LM is a typical task in natural language processing

语言模型是一个典型的自然语言处理任务。

2). Multilingual Pre-training

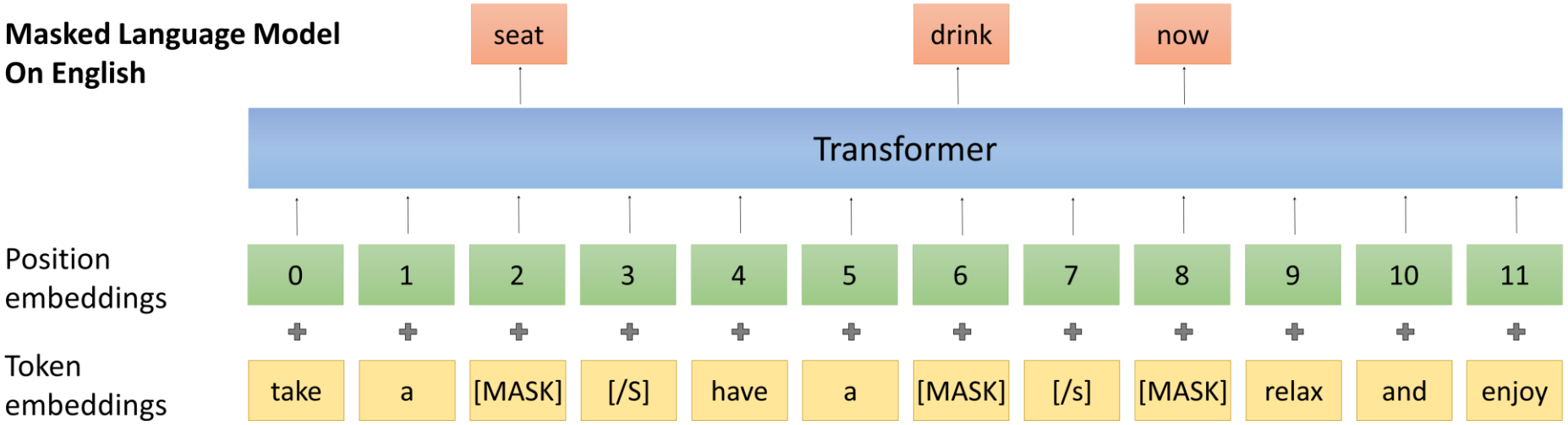
UNICODER: a UNIversal enCODER for multiple languages

(Contact: Yaobo LIANG, Nan DUAN)

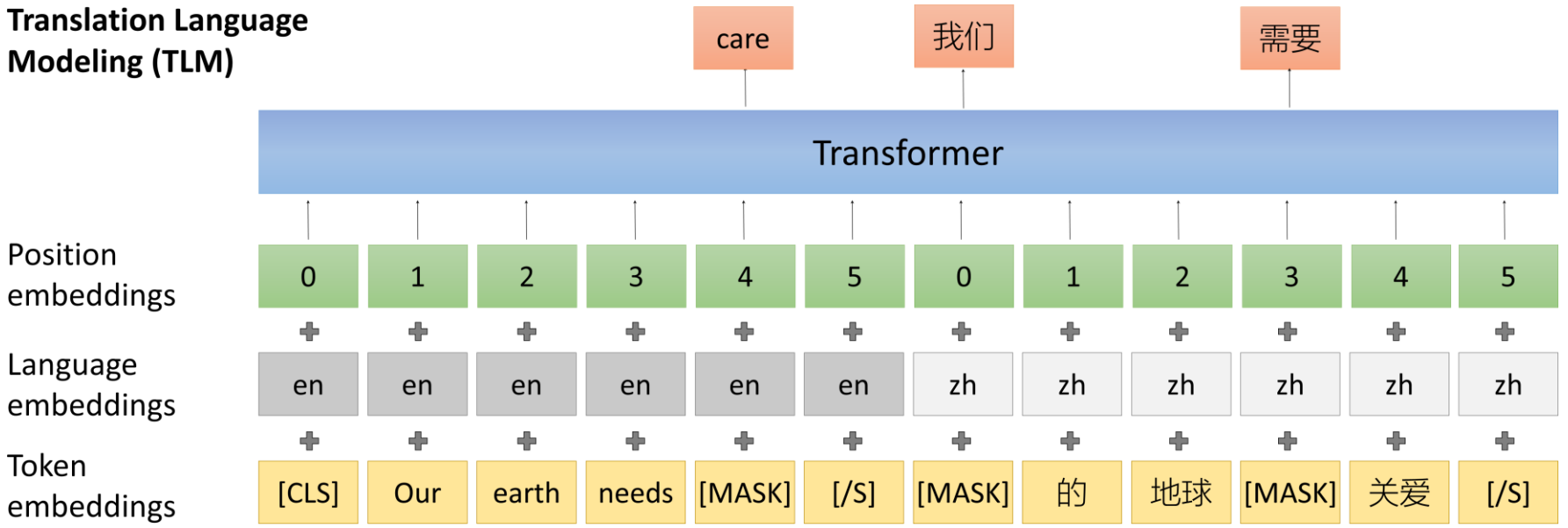


Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. **Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks**. EMNLP, 2019.

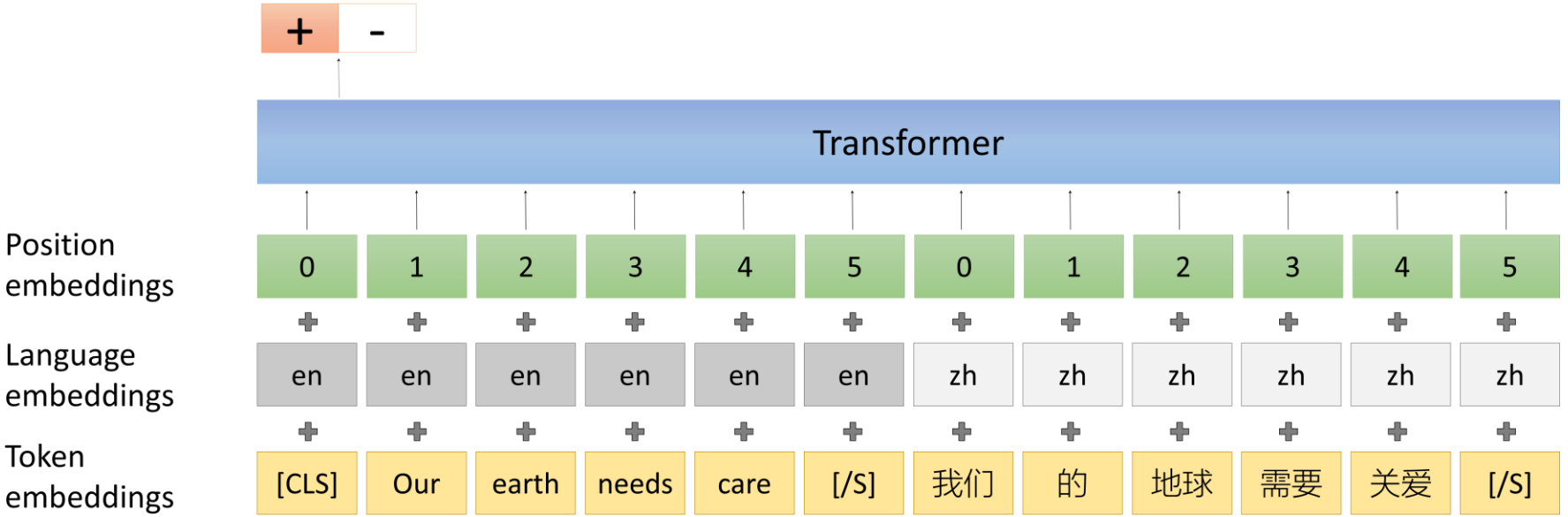
Pre-training Task (1): Masked Language Model



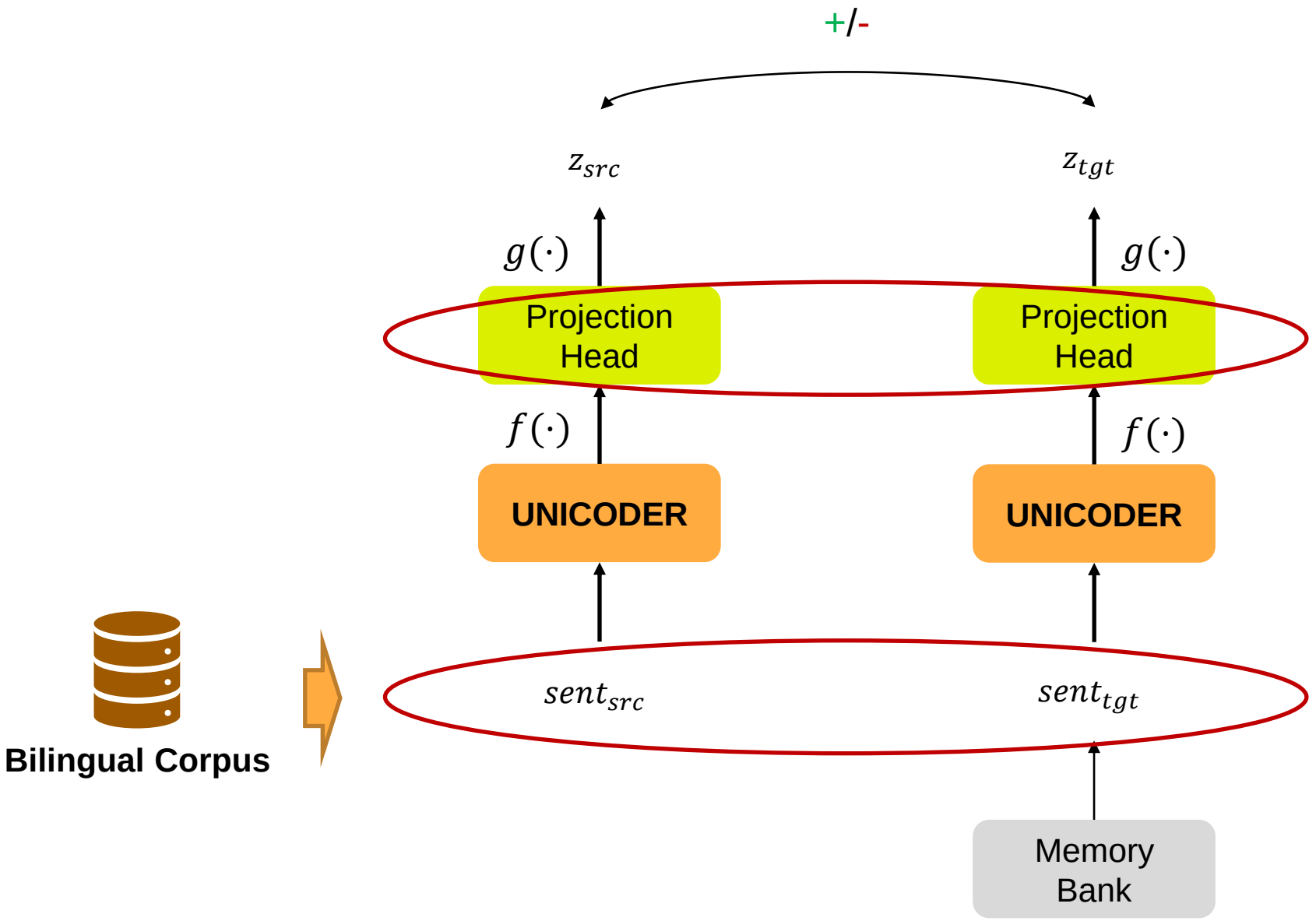
Pre-training Task (2): Translation Language Model



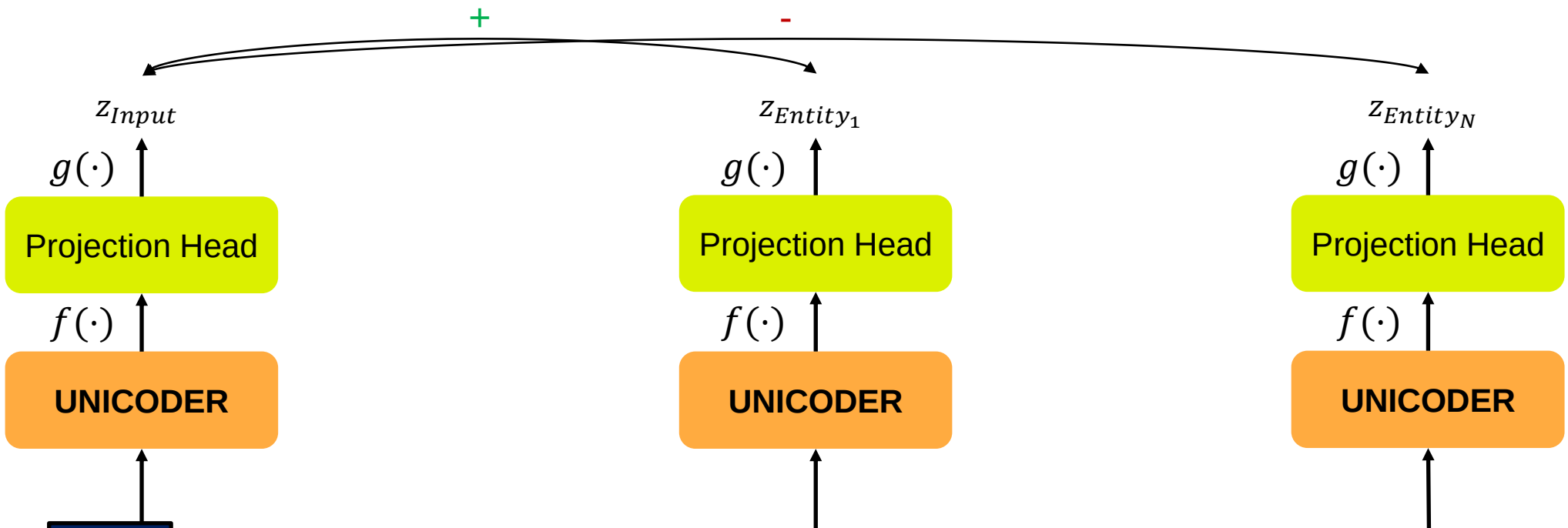
Pre-training Task (3): Bilingual Pair Classification – End-to-End



Pre-training Task (3): Bilingual Pair Classification – Memory Bank



Pre-training Task (4): Knowledge Regularization



The latest model in [MASK] 's line of Windows PCs, the Surface Laptop Go, forces buyers to confront why they want a Surface machine in the first place.

Microsoft

From Wikipedia, the free encyclopedia

Microsoft Corporation (/ˈmaɪkrəsɒft/ *MY-kroh-soft*) is an American multinational technology company with headquarters in Redmond, Washington. It develops, manufactures, licenses, supports, and sells computer software, consumer electronics, personal computers, and related services. Its best known software products are the Microsoft Windows line of operating systems, the Microsoft Office suite, and the Internet Explorer and Edge web browsers. Its flagship hardware products are the Xbox video game consoles and the Microsoft Surface lineup of touchscreen personal computers. Microsoft ranked No. 21 in the 2020 Fortune 500 rankings of the largest United States corporations by total revenue;[3] it was the world's largest software maker by revenue as of 2016[4] and is considered one of the domestic Big Five technology companies.

Google

From Wikipedia, the free encyclopedia

This article is about the company. For the search engine, see Google Sea (disambiguation).
Not to be confused with Googol or Goggles.

Google, LLC is an American multinational technology company that specializes in Internet-related services and products, which include online advertising technologies, a search engine, cloud computing, software, and hardware. It is considered one of the Big Five technology companies in the U.S. information technology industry, alongside Amazon, Facebook, Apple, and Microsoft.[9][10][11]

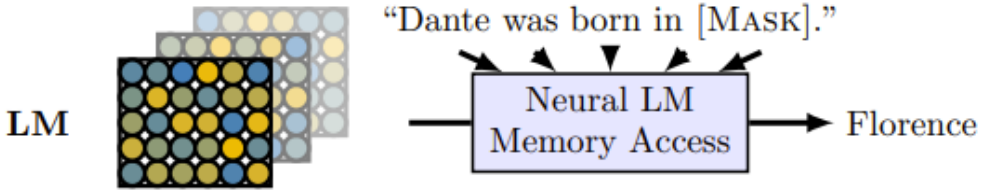
Evaluation: XNLI

XNLI task: to predict the entailment relation between two sentences
https://www.nyu.edu/projects/bowman/xnli/

Language	Premise / Hypothesis	Label
English	You don't have to stay there. You can leave.	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنأللوكاتات أ اتعرف ما إذا كانت ناجحة أم لا	Contradiction

Model	Training data	# Lan.	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
XLM	wiki+MT	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
mBERT	wiki	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM-R	CC	100	84.8	78.6	79.4	77.2	76.5	78.1	76.0	73.4	72.5	75.4	72.2	74.2	70.4	65.5	66.6	74.7
UNICODER	CC+MT	100	**84.9**	**80.2**	**80.4**	**79.3**	**78.3**	**79.1**	**77.7**	**75.3**	**75.4**	**77.5**	**74.8**	**76.7**	**72.4**	**70.0**	**68.8**	**76.7**

Evaluation: Knowledge Probing



Probe the factual and commonsense knowledge in language models.

Dataset	Relation	XLM-R (base)	UNICODER (base)
		P@1	P@1
Google-RE	birth-place	9.32	**13.81**
	birth-date	0.60	**0.84**
	death-place	7.98	**14.97**
	Total	7.37	**9.87**

Dataset	Relation	XLM-R (base)	UNICODER (base)
		P@1	P@1
T-Rex	1-1	48.37	**51.16**
	N-1	21.96	**25.51**
	N-M	17.87	**23.10**
	Total	21.65	**25.82**

Dataset	Relation	XLM-R (base)	UNICODER (base)
		P@1	P@1
SQuAD	Total	5.45	**7.88**

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller. Language Models as Knowledge Bases? EMNLP, 2019.

UNICODER for Multilingual NLG

(Contact: Yeyun GONG, Nan DUAN)

Unicoder Encoder

(12/24 layers, shared 250K vocabulary size, 100 languages)

Unicoder Decoder

(12 layers, shared 256K vocabulary size, 100 languages)

Text Noising Method

Sentence Permutation	*could this be sentence a in . any language*
Token Deletion	*this be a in any language*
Token Masking	*[MASK] could be a [MASK] in any [MASK] .*
Text Infilling	*this could be [MASK] in [MASK] .*

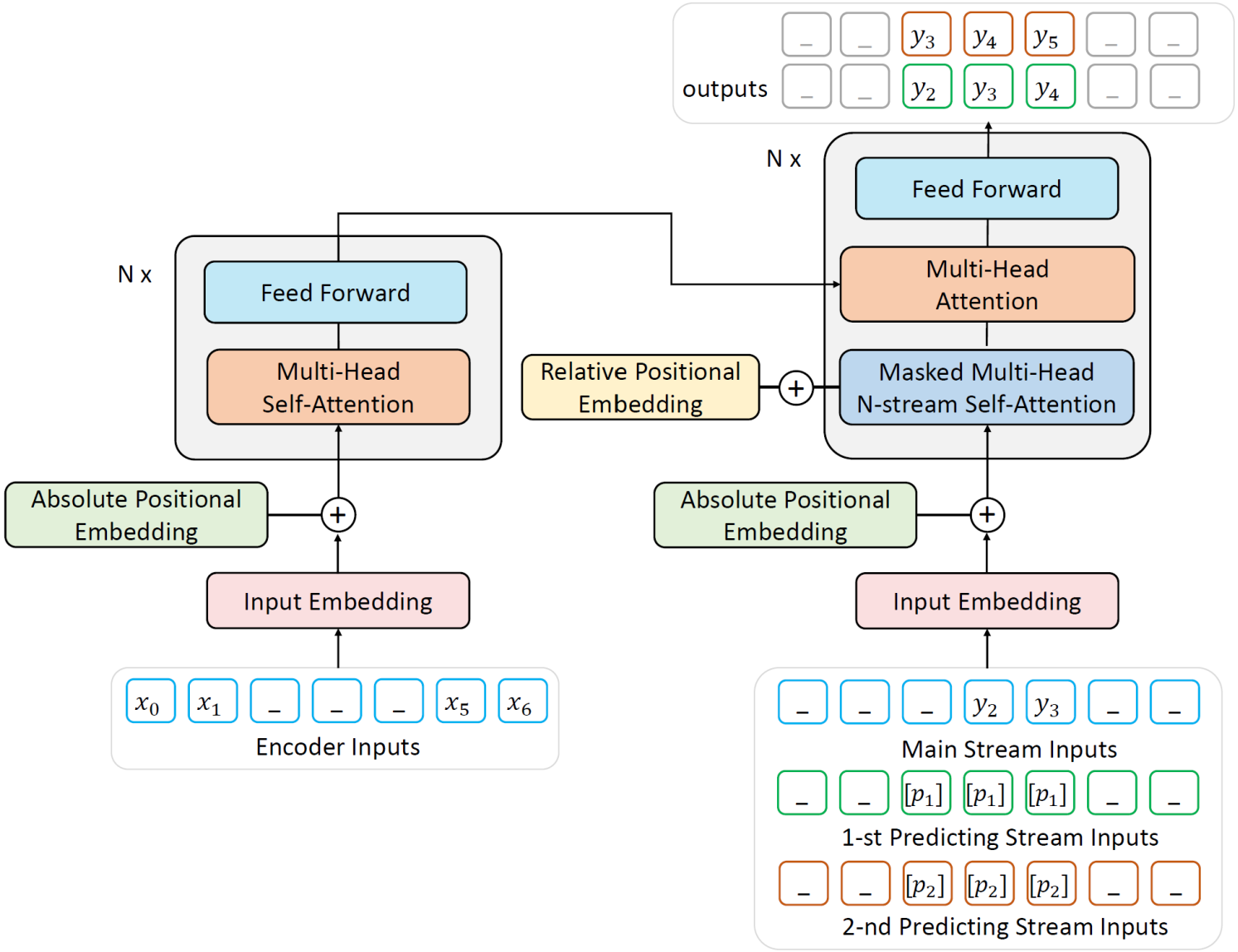
Text Denoising Method

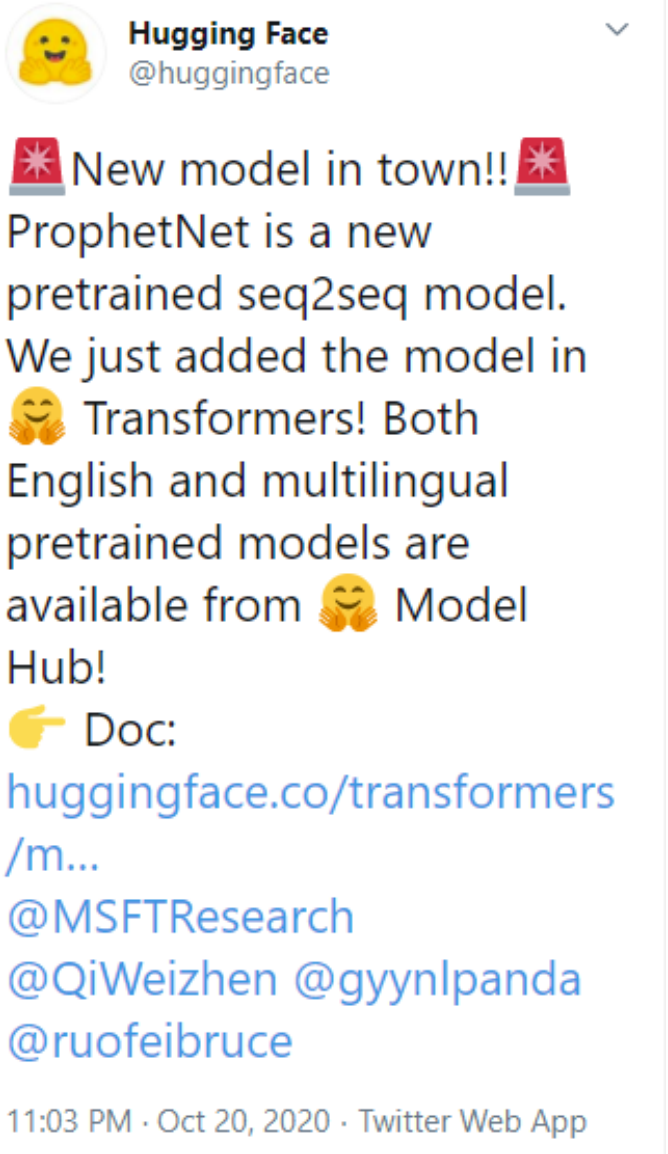
this could be a sentence in any language .

this could be a sentence in any language .

Yaobo Liang, Nan Duan, Yeyun Gong and Others. **XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation**. EMNLP, 2020.

xFNP is based on ProphetNet





Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, Ming Zhou. **ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training**. EMNLP, 2020.

Evaluation

	CNN/DM	**Gigaword**	**Xsum**	**MSNews**	**SQuAD QG**
	R1/R2/RL	R1/R2/RL	R1/R2/RL	R1/R2/RL	RL/B-4/MTR
Transformer	39.56/16.79/36.71	36.48/17.71/33.89	22.63/5.79/18.08	31.45/14.09/28.65	29.47/4.46/10.05
BART	44.16/21.28/40.90	37.53/17.67/34.34	45.14/22.27/37.25	43.86/24.00/39.21	50.31/22.04/26.40
ProphetNet	44.20/21.17/41.30	39.51/20.42/36.69	44.44/21.38/36.42	44.14/24.48/40.27	51.73/23.45/26.96

	de	fr	es	ru	AVG
mBART (reproduce)	6.8	8.7	9.0	7.7	8.1
xProphetNet	8.4	10.9	12	7.7	9.8

Multilingual News Title Generation (XGLUE)

	de	fr	es	it	pt	AVG
mBART (reproduce)	3.0	4.9	12.4	15.8	8.3	8.9
xProphetNet	4.2	5.7	17.4	18.9	10.7	11.4

Multilingual Question Generation (XGLUE)

XGLUE: A Benchmark Dataset for Multilingual NLU and NLG

XGLUE Home Intro **Leaderboard** Contact

XGLUE Dataset and Leaderboard

Tasks

1. NER
2. POS Tagging (POS)
3. News Classification (NC)
4. MLQA
5. XNLI
6. PAWS-X
7. Query-Ad Matching (QADSM)
8. Web Page Ranking (WPR)
9. QA Matching (QAM)
10. Question Generation (QG)
11. News Title Generation (NTG)

New Tasks!

Relevant Links

XGLUE Submission Guideline/Github | XGLUE Paper | Unicoder Paper(Baseline)

Leaderboard (05/25/2020-Present) ranked by XGLUE Score (average score on 11 tasks)

Rank	Model	Submission Date	NER	POS	NC	MLQA	XNLI	PAWS-X	QADSM	WPR	QAM	QG	NTG	XGLUE Score
1	**Unicoder Baseline** (XGLUE Team)	2020-05-25	79.7	79.6	83.5	66.0	75.3	90.1	68.4	73.9	68.9	10.6	10.7	64.2

https://microsoft.github.io/XGLUE/

Unicoder scaled Bing QnA to 100 languages and 200 regions in the world.

Bing blogs

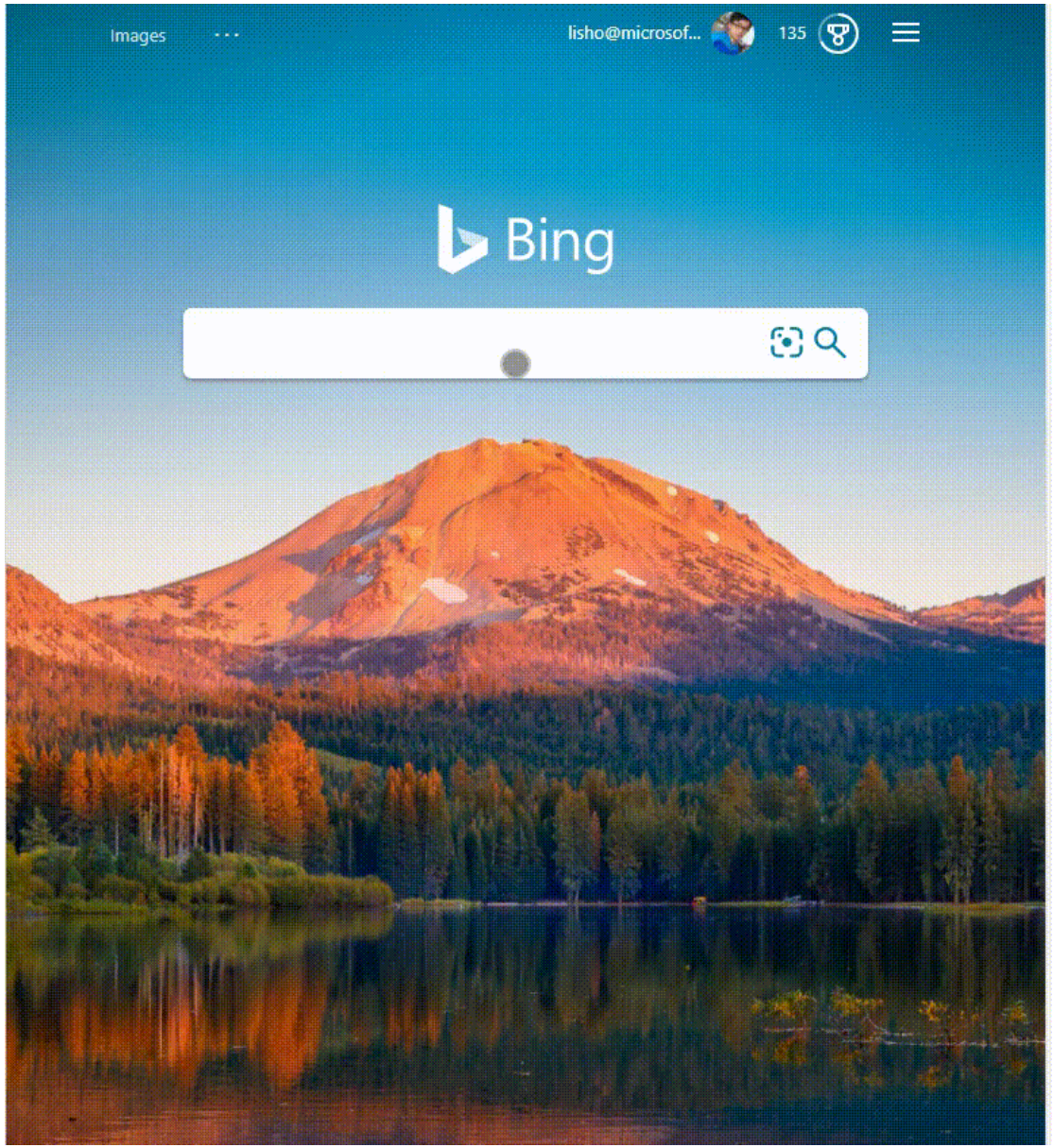
OCTOBER 1 2020

Bing Releases Intelligent Question-Answering Feature to 100+ Languages

Intelligent question-answering is one of the most useful and delightful features of search. As a user, you ask a question (e.g., "what are the benefits of eating apricots") and can get the answer directly (e.g., info about health and nutrition benefits of apricots) at the top of the page without further need to search for relevant content by yourself. The feature aims to direct users to the most concise and precise answers from web documents, thus saving users time and efforts.

English-language question answering from web has been enabled on Bing for several years, and another dozen of languages, like French and German, have been added within the last year. But our work isn't done - there are thousands of languages in the world! Not all of them have rich enough web content to derive good answers, but for those that do, uses of those spoken languages deserve the same useful, delightful, time-saving experience.

Recently, Bing expanded its intelligent question-answering feature to more than 100 languages, making AI and Bing itself more inclusive and accessible. What is amazing is this is achieved by using a language agnostic approach. In other words, the AI model generating the intelligent question-answering in Urdu is the same one generating the intelligent question-answering in Romanian. Here are some examples of this experience in various languages (if you speak a language other than English, feel free to give it a try, but be reminded to set your browser to the relevant language):

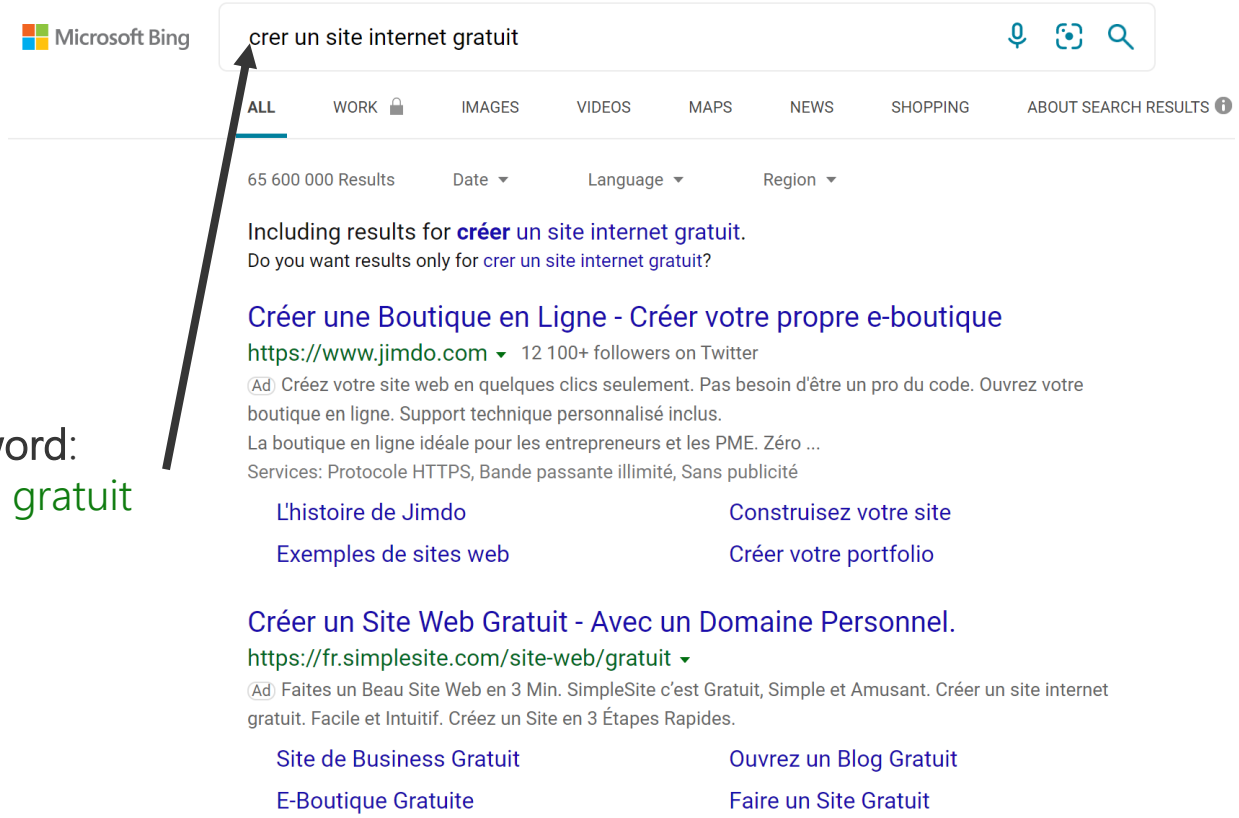


ProphetNet supports multilingual ad keyword generation.

Input Query: créer un site gratuitement

UNICODER based on ProphetNet

Generated Ad keyword:
crer un site internet gratuit



Summary

Multilingual pre-trained models can learn joint language representations from multilingual/bilingual corpus.

Multilingual pre-trained models can significantly alleviate the low-resource issues in multiple languages.

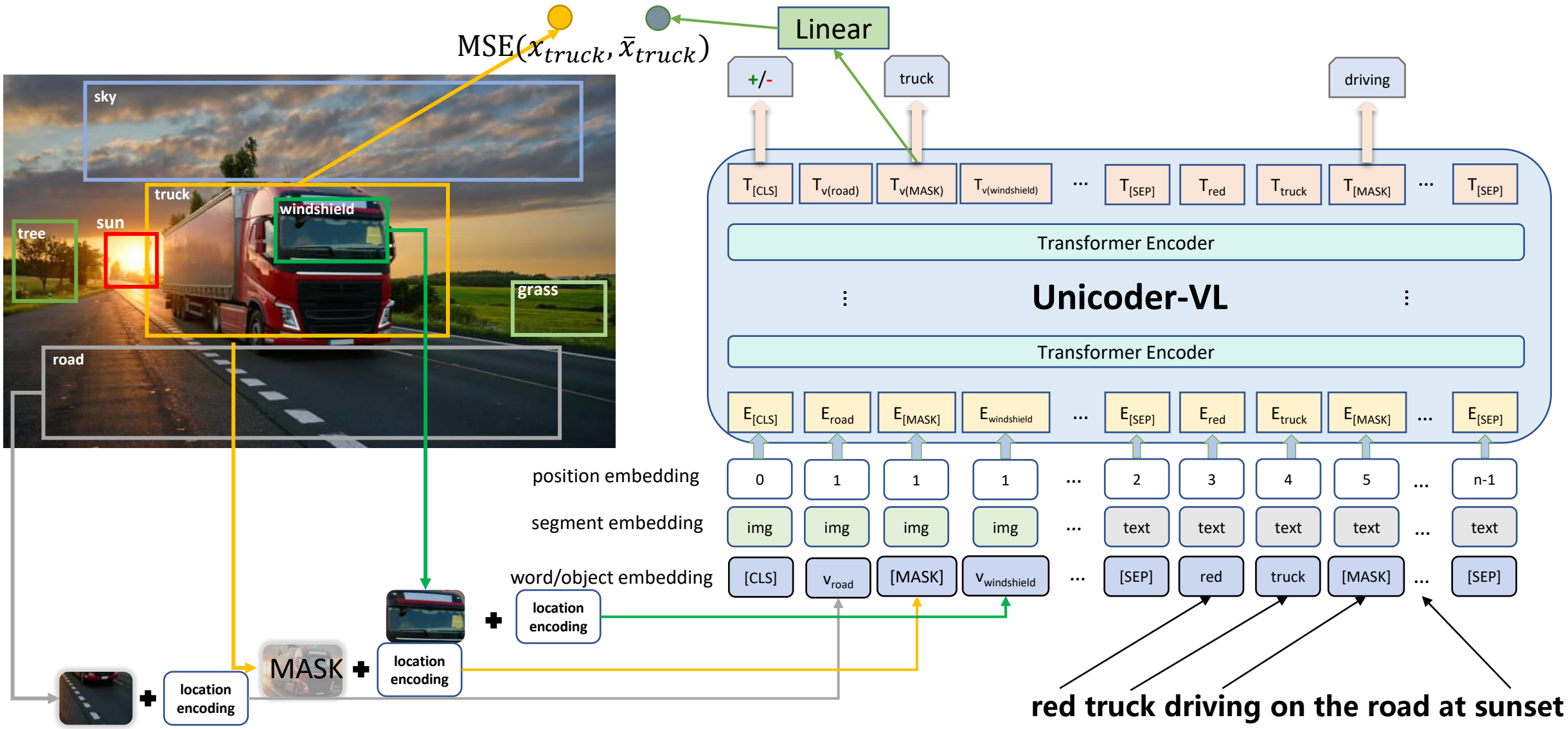
Multilingual pre-trained models can be successfully applied in real-world scenarios, such as search and ads.

MSRA released XGLUE (https://microsoft.github.io/XGLUE/) as a new benchmark for multilingual NLP.

3). Multimodal Pre-training

Unicoder-VL for Image-Language Tasks

(Contact: Haoyang HUANG, Nan DUAN)



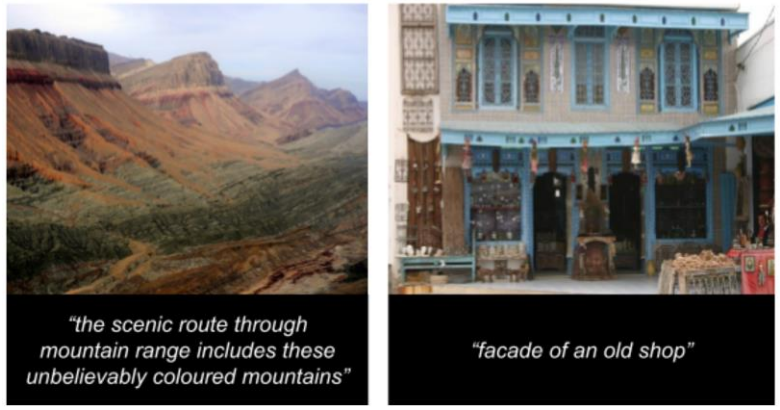
Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. AAAI 2020.

Evaluation Results: Image-Text Retrieval

Model	Text-to-Image Retrieval (Flickr30k)			Image-to-Text Retrieval (Flickr30k)		
	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT (Lu et al., 2019)	58.2	84.9	91.5	-	-	-
UNITER (Chen et al., 2019)	71.5	91.2	95.2	84.7	97.1	**99.0**
Unicoder-VL (Li et al., 2020)	**73.1**	**92.3**	**95.9**	**88.0**	**97.3**	98.6

Model	Text-to-Image Retrieval (MSCOCO)			Image-to-Text Retrieval (MSCOCO)		
	R@1	R@5	R@10	R@1	R@5	R@10
UNITER (Chen et al., 2019)	48.4	76.7	85.9	63.3	87.0	93.1
Unicoder-VL (Li et al., 2020)	**50.5**	**78.7**	**87.1**	**66.4**	**89.8**	**94.4**

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



Evaluation Results: Visual QA & Reasoning (GQA)

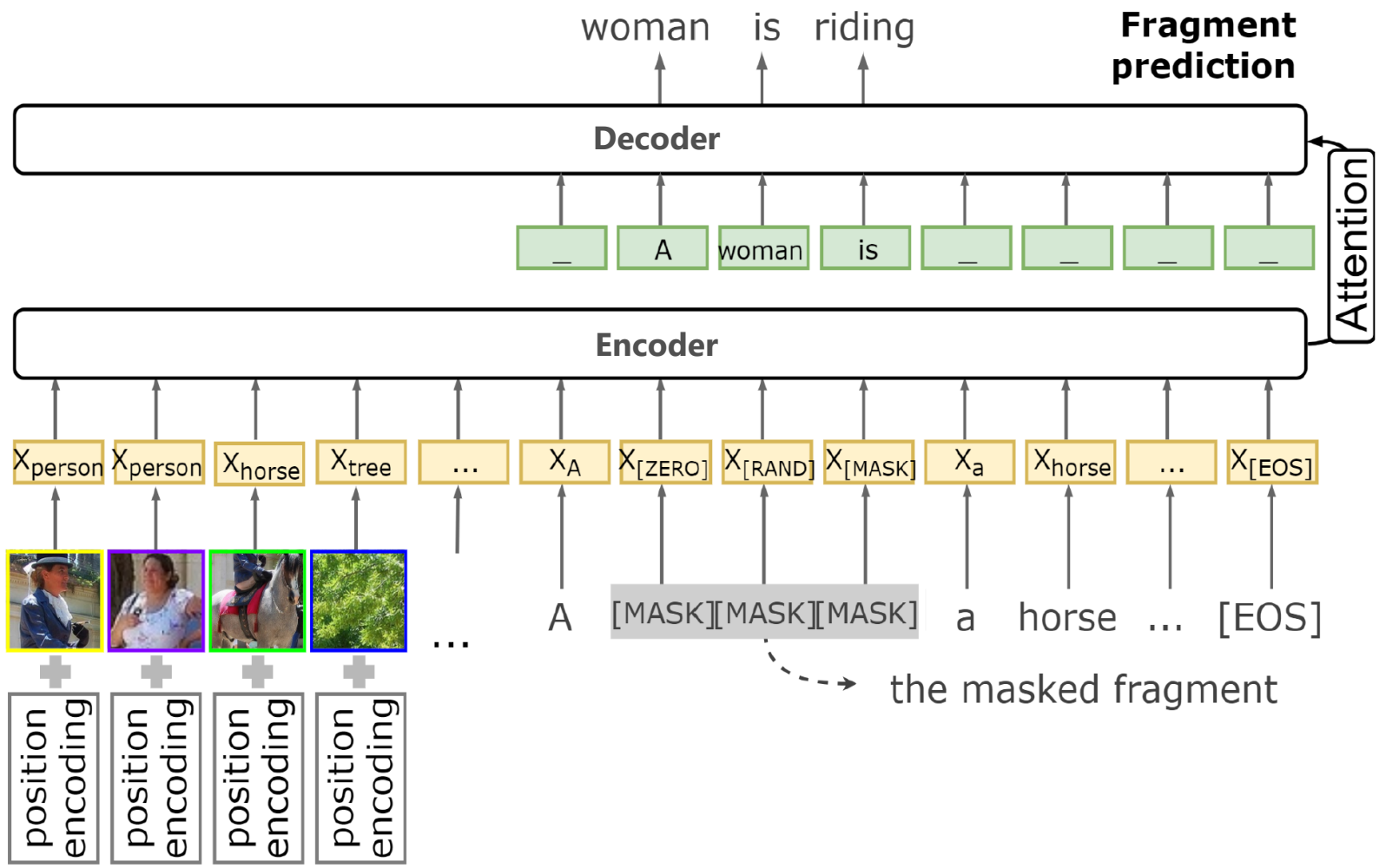
What color is the food on the red object left of the small girl that is holding a hamburger, yellow or green?

Rank	Participant team	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy	Last submission at
1	Human Performance (human)	91.20	87.40	98.40	97.20	98.90	0.00	89.30	2 years ago
2	DREAM+Unicoder-VL (MSRA)	84.46	68.60	91.47	83.75	96.42	3.68	76.04	1 year ago
3	TRRNet (Ensemble)	82.12	66.89	89.00	83.58	96.76	1.29	74.03	8 months ago
4	MIL-nbgao	80.80	67.64	91.76	83.90	96.73	1.70	73.81	24 days ago
5	Kakao Brain	79.68	67.73	77.02	83.70	96.36	2.46	73.33	1 year ago
6	AIOZ (Coarse-to-Fine Reasoning, Sing)	81.16	64.19	90.96	84.81	96.77	2.39	72.14	10 months ago
7	270	77.50	63.82	86.94	83.77	96.65	1.49	70.23	1 year ago
8	NSM ensemble (updated)	80.45	56.16	93.83	84.16	96.53	2.78	67.55	1 year ago
9	TRRNet (Single)	77.91	50.22	89.84	85.15	96.47	5.25	63.20	7 months ago
10	NSM single (updated)	78.94	49.25	93.25	84.28	96.41	3.71	63.17	1 year ago

1/100

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

Extend Unicoder-VL to Image Captioning

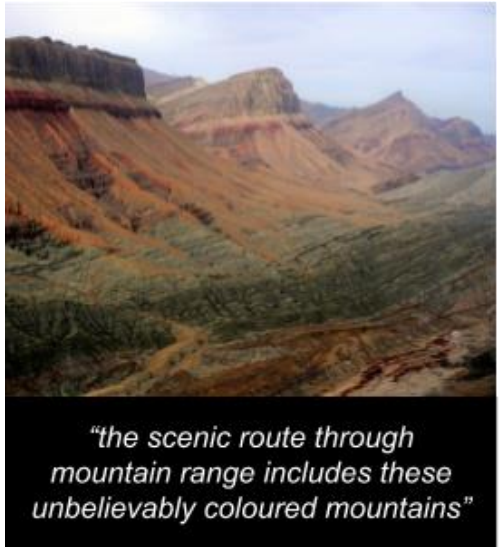


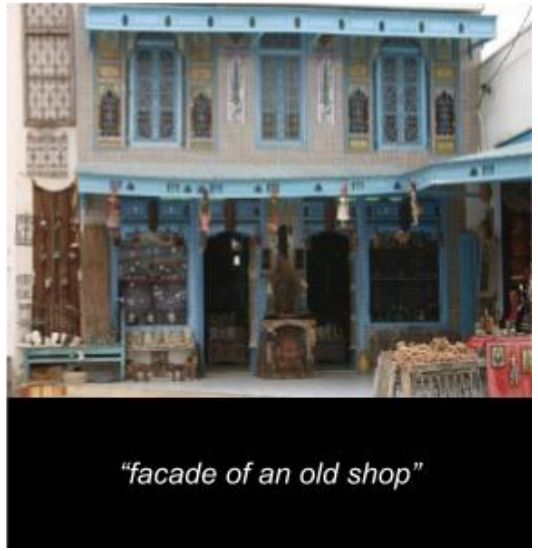
Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, Ming Zhou. **XGPT: Cross-modal Generative Pre-Training for Image Captioning**. arXiv 2020.

Evaluation Results: Image Captioning

Pre-trained with Conceptual Captions dataset
~3.3M images annotated with captions
Evaluated on MSCOCO dataset

Methods	Image Caption			
	BLEU@4	METEOR	CIDEr	SPICe
BUTD (Anderson et al. 2018)	36.2	27.0	113.5	20.3
NBT (Lu et al. 2018)	34.7	27.1	107.2	20.1
VLP (Zhou et al. 2018)	36.5	28.4	116.9	20.8
Unicoder-VL (Huang et al., 2020)	**37.2**	**28.6**	**120.1**	**21.8**





Extend Unicoder-VL to Multilingual Scenarios (a.k.a. M^3P)

(Contact: Haoyang HUANG, Nan DUAN)

Pre-training Overview

Multilingual Image-Text Retrieval

Multilingual Image Captioning

Multimodal Machine Translation

Multitask Multilingual Multimodal Pre-trained Model
(M^3P for 100 languages)

Multilingual Pre-trained Model
(**Unicoder** for 100 languages)

- masked language model
- translation language model
- text denoising auto-encoding

- Multilingual corpus (text only)
- Bilingual corpus (text only)

Vision-Language Pre-trained Model
(**Unicoder-VL** for English)

- masked token loss
- contrastive loss
- image captioning

- Image-Text (English) pairs

(Research) Evaluation Datasets

En - Two cars are racing on a track while the audience watches from behind a fence
De - Zwei Rennautos fahren auf der Restricken in die Kurve (Tr: Two race cars drive on the race track in the curve)
Fr – Deux voitures roulent sur un circuit. (Tr: Two race cars drive on the race track in the curve)
Cs– Dvě auta jedou po závodní dráze (Tr: Two cars ride the race track)

Multi30K dataset (en/de/fr/cs):

- 31,783 images in total
- 5 captions per image in **English (en)** and **German (de)**
- 1 caption per image in **French (fr)** and **Czech (cs)**

En - A young man playing frisbee in a grassy park

Cn - 两个男人在公园的草地上跳起来接飞盘
(Tr: Two men jump on the grass in the park and pick up the Frisbee)

Ja - 芝生の上で女性がフリスビーで遊んでいます
(Tr: A woman is playing frisbee on the grass)

MSCOCO dataset (en/ja/zh):

- 123,287 images in total
- 5 captions per image in **English (en)** and **Japanese (ja)**
- 1~2 captions per image in **Chinese (zh)**

Haoyang Huang, Lin Su, Di Qi, Nan Duan and Others. **M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training**. arXiv 2020.

Evaluation Results

Task	Multilingual Image-Text Retrieval (Multi30K + MSCOCO)						Multilingual Image Captioning (Multi30K + MSCOCO)						Multimodal MT (Multi30K)	
	en	de	fr	cs	ja	zh	en	de	fr	cs	ja	zh	en→fr	en→de
SoTA	**92.7**	72.1	65.9	64.8	76.0	74.8	**37.4**	3.8	5.0	2.8	38.5	36.7	53.8	31.6
M^3P_B	88.0	**82.0**	**73.5**	**70.2**	**86.8**	**81.8**	34.7	**16.6**	**8.7**	**5.4**	**40.2**	**39.7**	**55.5**	**35.7**
Δ	4.7 ↓	9.9 ↑	7.6 ↑	5.4 ↑	10.8 ↑	7.0 ↑	3.7 ↓	12.8 ↑	3.7 ↑	2.6 ↑	1.7 ↑	3.0 ↑	1.7 ↑	4.1 ↑

Blue numbers indicates the best result for a task. For retrieval tasks, we use *mean Recall* as the metric, which is an average score of R@1, R@5 and R@10 on i2t and t2i tasks. For captioning and translation tasks, we use *BLEU-4* as the metric.

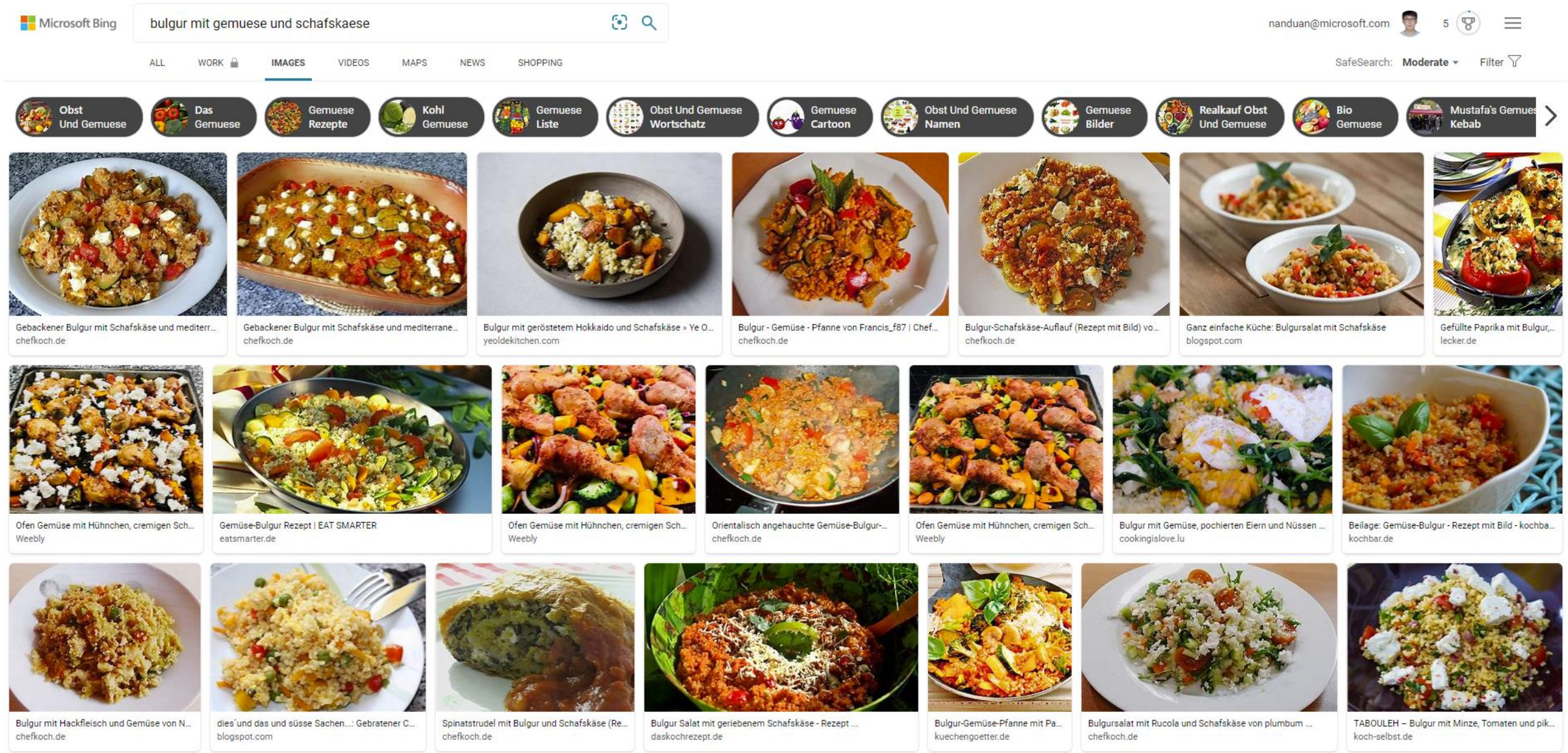
image caption output (**zh**): 一辆载着人和纸糊的房子的卡车行驶在街道上
(translation: a truck carrying people and paper houses travels down the street)

image caption input (**en**): A Boston Terrier is running on lush green grass in front of a white fence.

caption translation output (**fr**): Le Boston Terrier court sur l'herbe verte luxurie devant une clôture blanche.
(translation: The Boston Terrier runs on lush green grass in front of a white fence.)

caption translation output (**de**): Ein Hund läuft auf grünem Rasen vor einem weißen Zaun.
(translation: A dog runs on green grass in front of a white fence.)

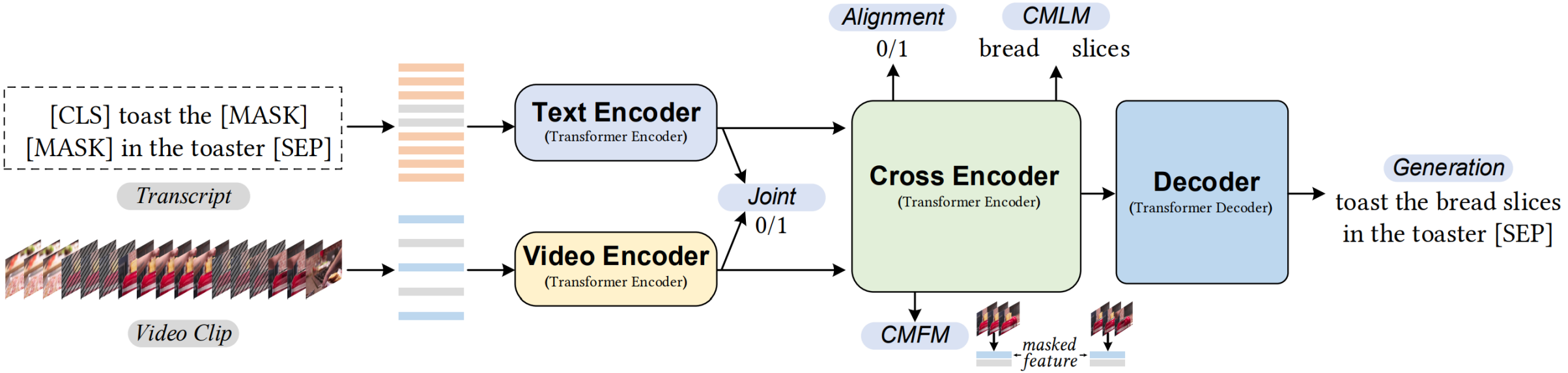
Unicoder-VL scaled Bing image search to 8 top-tier languages and 17 markets.

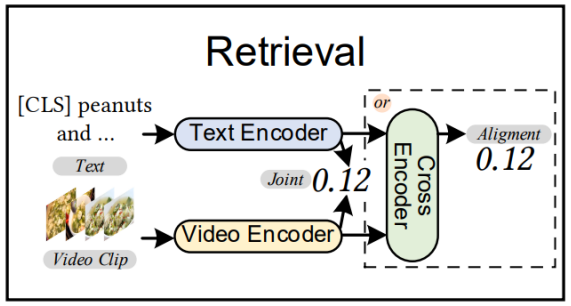


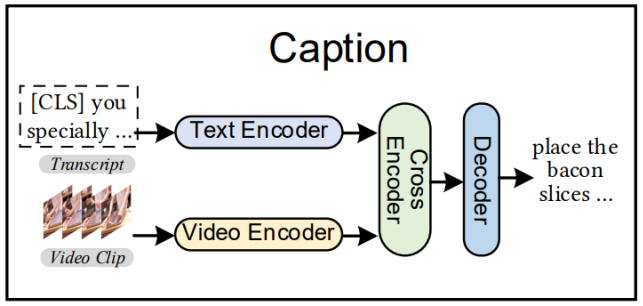
Unicoder-VL for Video-Language Tasks

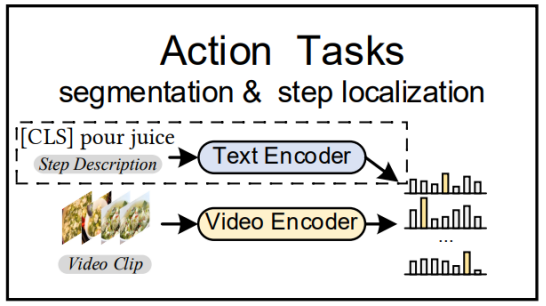
(Contact: Lei JI, Nan DUAN)

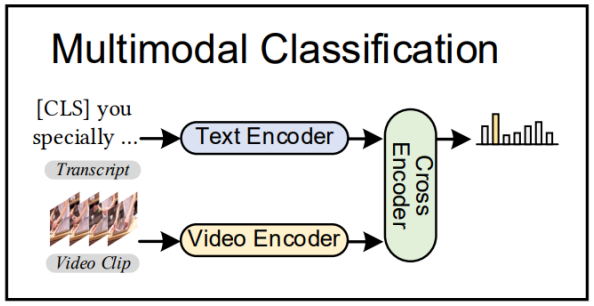
1. Video-Text Joint Embedding **2.** Video-Text Alignment **3.** Masked Frame Model **4.** Masked Language Model **5.** Caption Generation







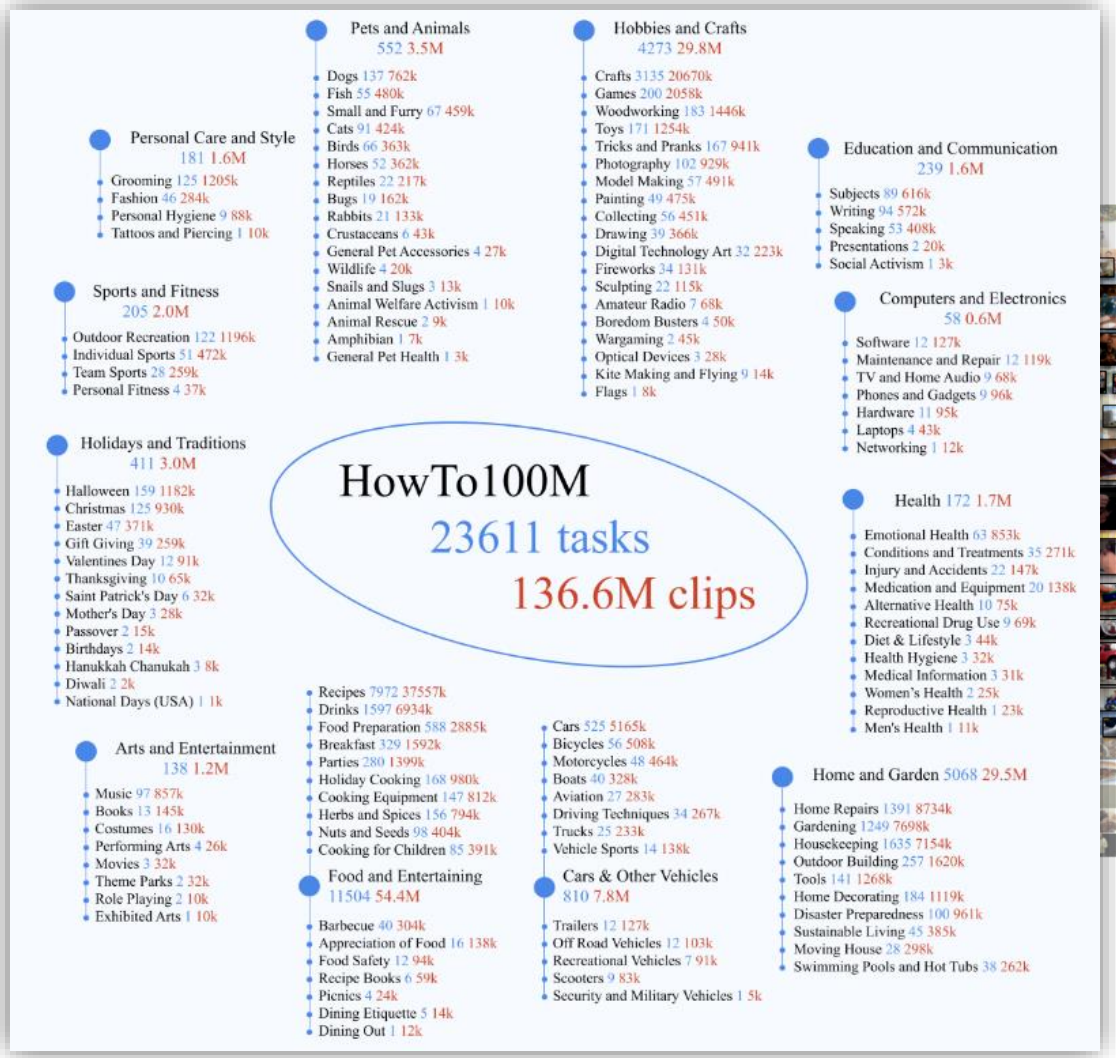




Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, Ming Zhou. **UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation**. arXiv 2020.

Pre-training Corpus

HowTo100M (Miech et al., 2019): 136M video clips with captions from 1.2M Youtube videos.



Evaluation Results: Text-based Video Retrieval

MSR-VTT (Xe et al., 2016): 200K clip-text pairs from 10K videos in 20 categories

YouCook2 (Zhou et al., 2018): 14k clip-text pairs from 2k videos.

Input: **Query:** cook a pizza

Video:

Output: Yes

Methods	R@1	R@5	R@10	Median R
Random	0.03	0.15	0.3	1675
HGLMM (Klein et al., 2015)	4.6	14.3	21.6	75
HowTo100M (Miech et al., 2019)	8.2	24.5	35.3	24
MIL-NCE (Miech et al., 2020)	15.1	38.0	51.2	10
ActBERT (Zhu and Yang, 2020)	9.6	26.7	38.0	19
VideoAsMT (Korbar et al., 2020)	11.6	-	43.9	-
UniVL (FT-Joint)	22.2	52.2	66.2	5
UniVL (FT-Align)	**28.9**	**57.6**	**70.0**	**4**

Table 1: Results of text-based video retrieval on Youcook2 dataset.

Methods	R@1	R@5	R@10	Median R
Random	0.1	0.5	1.0	500
C+LSTM+SA (Torabi et al., 2016)	4.2	12.9	19.9	55
VSE (Kiros et al., 2014)	3.8	12.7	17.1	66
SNUVL (Yu et al., 2016)	3.5	15.9	23.8	44
Kaufman et al. (2017)	4.7	16.6	24.1	41
CT-SAN (Yu et al., 2017)	4.4	16.6	22.3	35
JSFusion (Yu et al., 2018)	10.2	31.2	43.2	13
HowTo100M (Miech et al., 2019)	14.9	40.2	52.8	9
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
ActBERT (Zhu and Yang, 2020)	8.6	23.4	33.1	36
VideoAsMT (Korbar et al., 2020)	14.7	-	52.8	-
UniVL (FT-Joint)	20.6	49.1	62.9	6
UniVL (FT-Align)	**21.2**	**49.6**	**63.1**	**6**

Table 2: Results of text-based video retrieval on MSR-VTT dataset.

Evaluation Results: Video Captioning

YouCook2 (Zhou et al., 2018): 14k clip-text pairs from 2k videos.

Methods	Input	B-3	B-4	M	R-L	CIDEr
Bi-LSTM (Zhou et al., 2018a)	V	-	0.87	8.15	-	-
EMT (Zhou et al., 2018b)	V	-	4.38	11.55	27.44	0.38
VideoBERT (Sun et al., 2019b)	V	6.80	4.04	11.01	27.50	0.49
CBT (Sun et al., 2019a)	V	-	5.12	12.97	30.44	0.64
ActBERT (Zhu and Yang, 2020)	V	8.66	5.41	13.30	30.56	0.65
VideoAsMT (Korbar et al., 2020)	V	-	5.3	13.4	-	-
AT (Hessel et al., 2019)	T	-	8.55	16.93	35.54	1.06
DPC (Shi et al., 2019)	V + T	7.60	2.76	18.08	-	-
AT+Video (Hessel et al., 2019)	V + T	-	9.01	17.77	36.65	1.12
UniVL	V	16.46	11.17	17.57	40.09	1.27
UniVL	T	20.32	14.70	19.39	41.10	1.51
UniVL	V + T	**23.87**	**17.35**	**22.35**	**46.52**	**1.81**

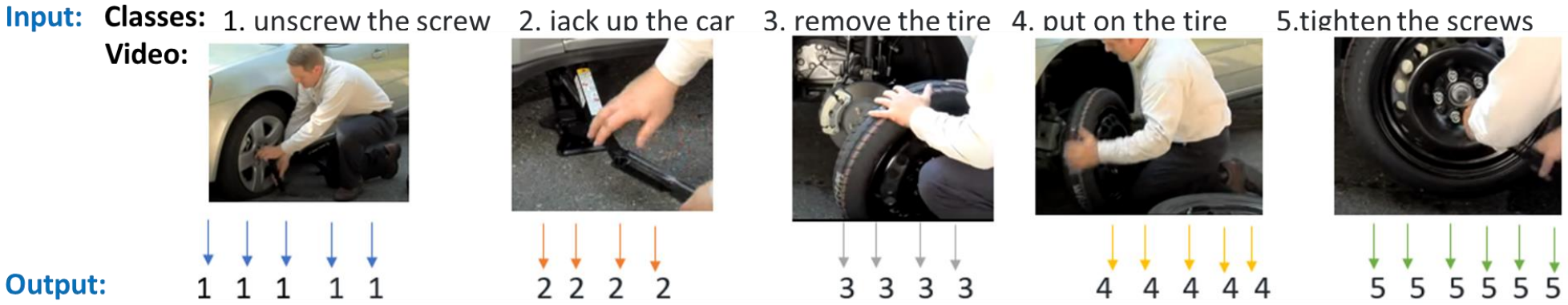
Table 3: The multimodal video captioning results on Youcook2 dataset. 'V' means video and 'T' means Transcript.

Input:

Output: sprinkle some cheese on top of pizza and bake them in the oven

Evaluation Results: Frame-wise Action Classification

COIN (Tang et al., 2019): 11,827 videos related to 180 different tasks in 12 domains.



Methods	Frame Accuracy (%)
NN-Viterbi (Richard et al., 2018)	21.17
VGG (Simonyan and Zisserman, 2014)	25.79
TCFPN-ISBA (Ding and Xu, 2018)	34.30
CBT (Sun et al., 2019a)	53.90
MIL-NCE (Miech et al., 2020)	61.00
ActBERT (Zhu and Yang, 2020)	56.95
UniVL	**70.02**

Table 4: Action segmentation results on COIN.

Evaluation Results: Video Sentiment Analysis

CMU-MOSI (Zadeh et al., 2018): 2,199 videos for multimodal sentiment analysis.

Input:

Output:

- Highly Positive
- Positive
- Weakly Positive
- Neutral
- Weakly Negative
- Negative
- Highly Negative

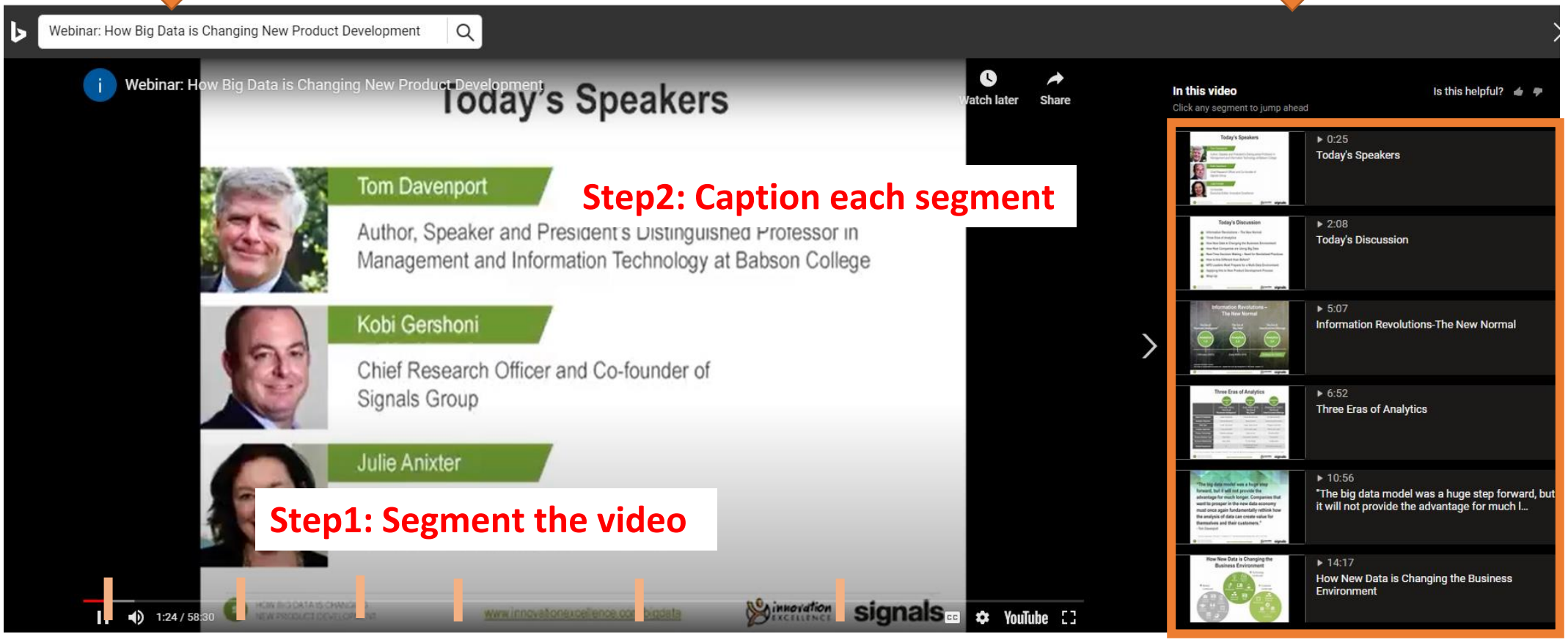
Methods	BA	F1	MAE	Corr
MV-LSTM (Rajagopalan et al., 2016)	73.9/-	74.0/-	1.019	0.601
TFN (Zadeh et al., 2017)	73.9/	73.4/-	1.040	0.633
MARN (Zadeh et al., 2018b)	77.1/	77.0/-	0.968	0.625
MFN (Zadeh et al., 2018a)	77.4/	77.3/-	0.965	0.632
RMFN (Liang et al., 2018)	78.4/	78.0/-	0.922	0.681
RAVEN (Wang et al., 2019)	78.0/	-/-	0.915	0.691
MulT (Tsai et al., 2019)	/83.0	-/82.8	0.870	0.698
FMT (Zadeh et al., 2019)	81.5/83.5	81.4/83.5	0.837	0.744
UniVL	**83.2/84.6**	**83.3/84.6**	**0.781**	**0.767**

Table 6: Multimodal sentiment analysis results on CMU-MOSI dataset. BA means binary accuracy, MAE is Mean-absolute Error, and Corr is Pearson Correlation Coefficient. For BA and F1, we report two numbers following Zadeh et al. (2019): the number on the left side of / is calculated based on the approach from Zadeh et al. (2018b), and the right side is by Tsai et al. (2019).

Unicoder-VL enables Bing video chaptering.

Input a video

Output video chapters



Webinar: How Big Data is Changing New Product Development

Summary

NLP pre-training pipeline can be applied to vision-language scenarios.

VL pre-trained models improve VL downstream tasks significantly.

Vision pre-trained models are critical to multimodal pre-training.

4). Future Work

Future Work: Machine Reasoning

Machine Reasoning: Technology, Dilemma and Future

Nan Duan, Duyu Tang, Ming Zhou

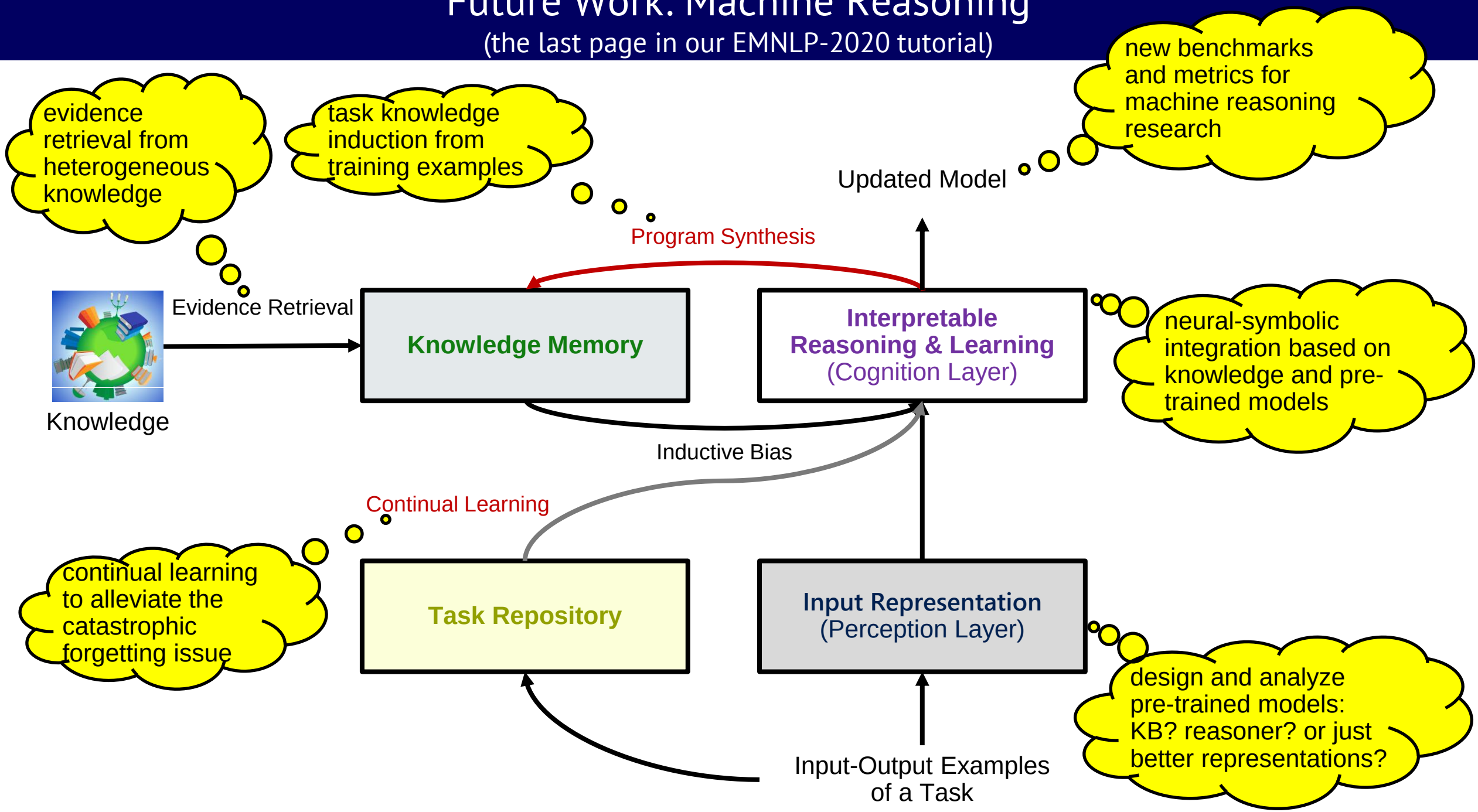


Machine reasoning research aims to build interpretable AI systems that can solve problems or draw conclusions from what they are told (i.e. facts and observations) and already know (i.e. models, common sense and knowledge) under certain constraints. In this tutorial, we will (1) describe the motivation of this tutorial and give our definition on machine reasoning; (2) introduce typical machine reasoning frameworks, including symbolic reasoning, probabilistic reasoning, neural-symbolic reasoning and neural-evidence reasoning, and show their successful applications in real-world scenarios; (3) talk about the dilemma between black-box neural networks with state-of-the-art performance and machine reasoning approaches with better interpretability; (4) summarize the content of this tutorial and discuss possible future directions.

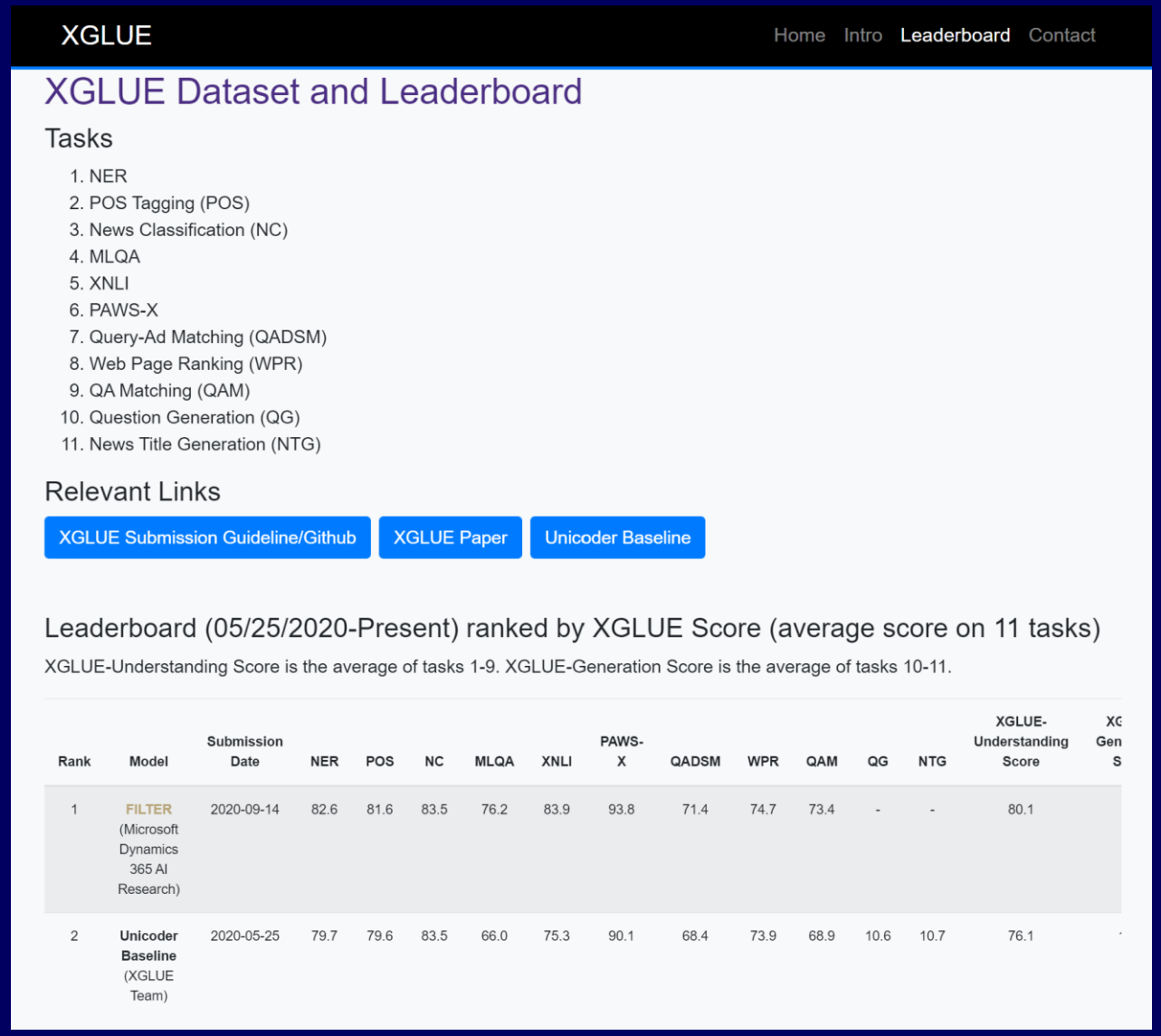
https://slideslive.com/38940827/t4-machine-reasoning-technology-dilemma-and-future

Future Work: Machine Reasoning

(the last page in our EMNLP-2020 tutorial)



Thank You and Welcome to Use Our Datasets!



XGLUE Home Intro Leaderboard Contact

XGLUE Dataset and Leaderboard

Tasks

1. NER
2. POS Tagging (POS)
3. News Classification (NC)
4. MLQA
5. XNLI
6. PAWS-X
7. Query-Ad Matching (QADSM)
8. Web Page Ranking (WPR)
9. QA Matching (QAM)
10. Question Generation (QG)
11. News Title Generation (NTG)

Relevant Links

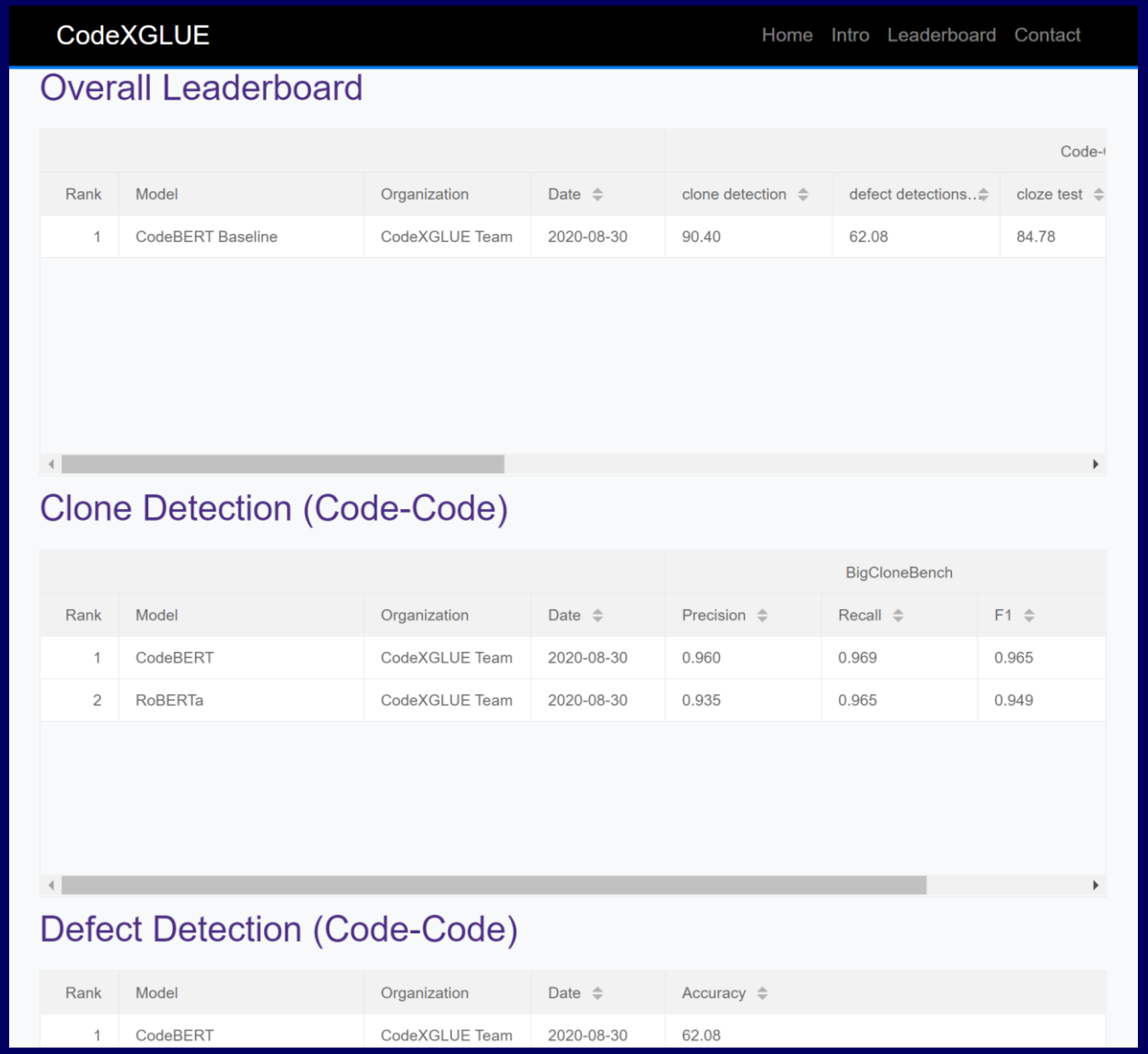
XGLUE Submission Guideline/Github | XGLUE Paper | Unicoder Baseline

Leaderboard (05/25/2020-Present) ranked by XGLUE Score (average score on 11 tasks)

XGLUE-Understanding Score is the average of tasks 1-9. XGLUE-Generation Score is the average of tasks 10-11.

Rank	Model	Submission Date	NER	POS	NC	MLQA	XNLI	PAWS-X	QADSM	WPR	QAM	QG	NTG	XGLUE-Understanding Score	XG Gen S
1	FILTER (Microsoft Dynamics 365 AI Research)	2020-09-14	82.6	81.6	83.5	76.2	83.9	93.8	71.4	74.7	73.4	-	-	80.1	
2	**Unicoder Baseline** (XGLUE Team)	2020-05-25	79.7	79.6	83.5	66.0	75.3	90.1	68.4	73.9	68.9	10.6	10.7	76.1	

XGLUE: https://microsoft.github.io/XGLUE/



CodeXGLUE Home Intro Leaderboard Contact

Overall Leaderboard

				Code-		
Rank	Model	Organization	Date	clone detection	defect detections..	cloze test
1	CodeBERT Baseline	CodeXGLUE Team	2020-08-30	90.40	62.08	84.78

Clone Detection (Code-Code)

				BigCloneBench		
Rank	Model	Organization	Date	Precision	Recall	F1
1	CodeBERT	CodeXGLUE Team	2020-08-30	0.960	0.969	0.965
2	RoBERTa	CodeXGLUE Team	2020-08-30	0.935	0.965	0.949

Defect Detection (Code-Code)

Rank	Model	Organization	Date	Accuracy
1	CodeBERT	CodeXGLUE Team	2020-08-30	62.08

CodeXGLUE: https://microsoft.github.io/CodeXGLUE/