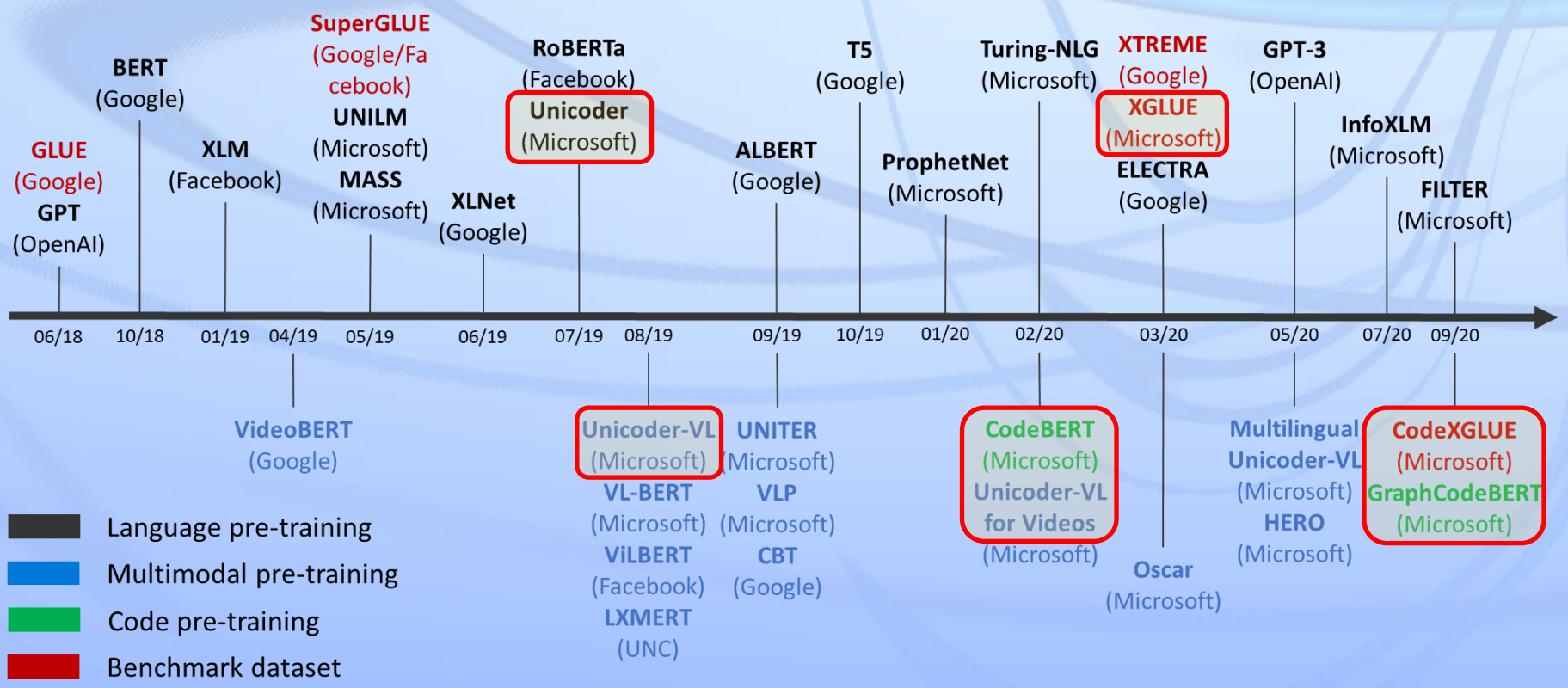
CNCC

Learning Universal Representations via Multitask Multilingual Multimodal Pre-training

Dr. Nan DUAN (段楠)
Principal Researcher
Microsoft Research Asia
CNCC@2020-10-23

CNCC

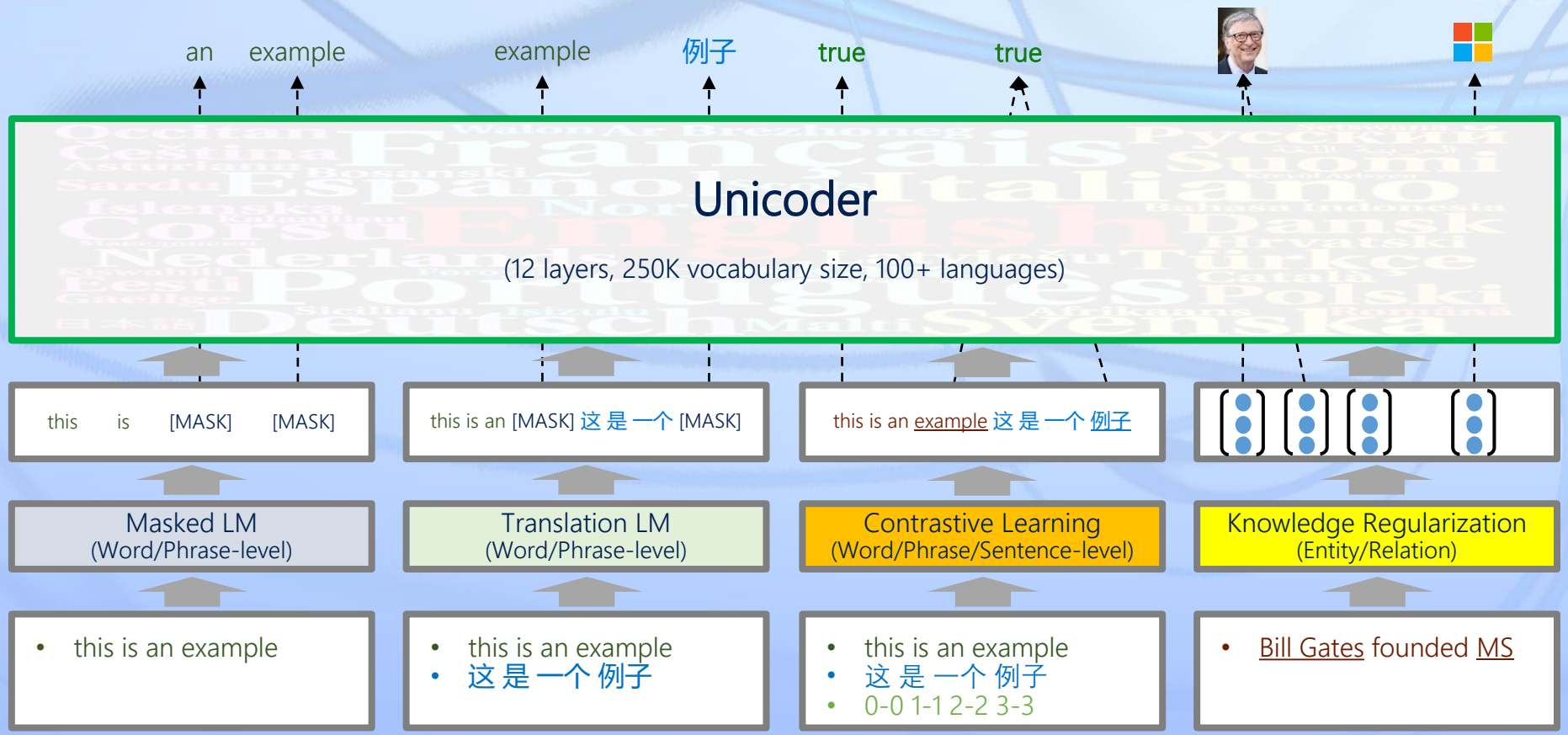
An **Incomplete** Roadmap of Pre-training



Outline

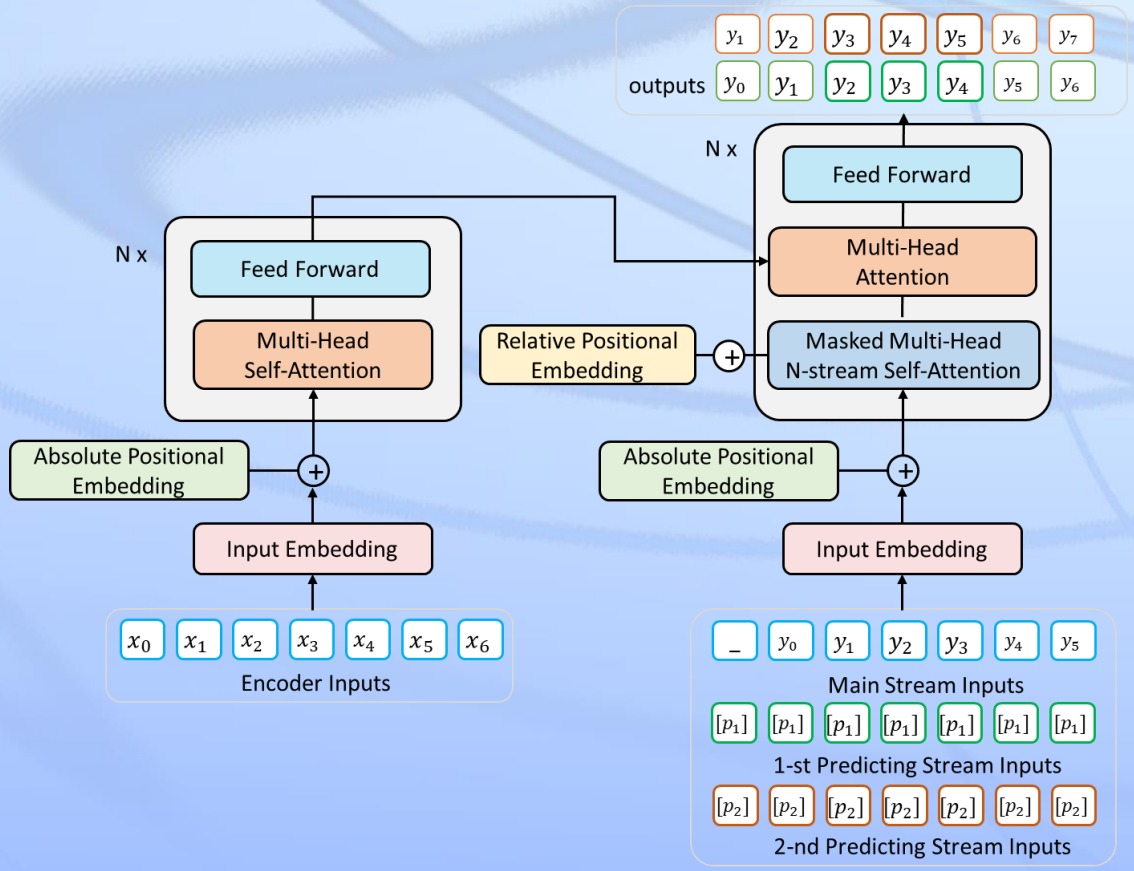
- **Multilingual/Multimodal Pre-trained Models**
  - Unicoder for multilingual language tasks
  - Unicoder-VL for vision-language tasks
- **From Natural Language to Programming Language**
- **Summary & Future Work**

Unicoder for Multilingual NLU



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.

Multilingual NLG 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.

CNCC

XGLUE: A Benchmark for Multilingual Tasks

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!

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

Relevant Links

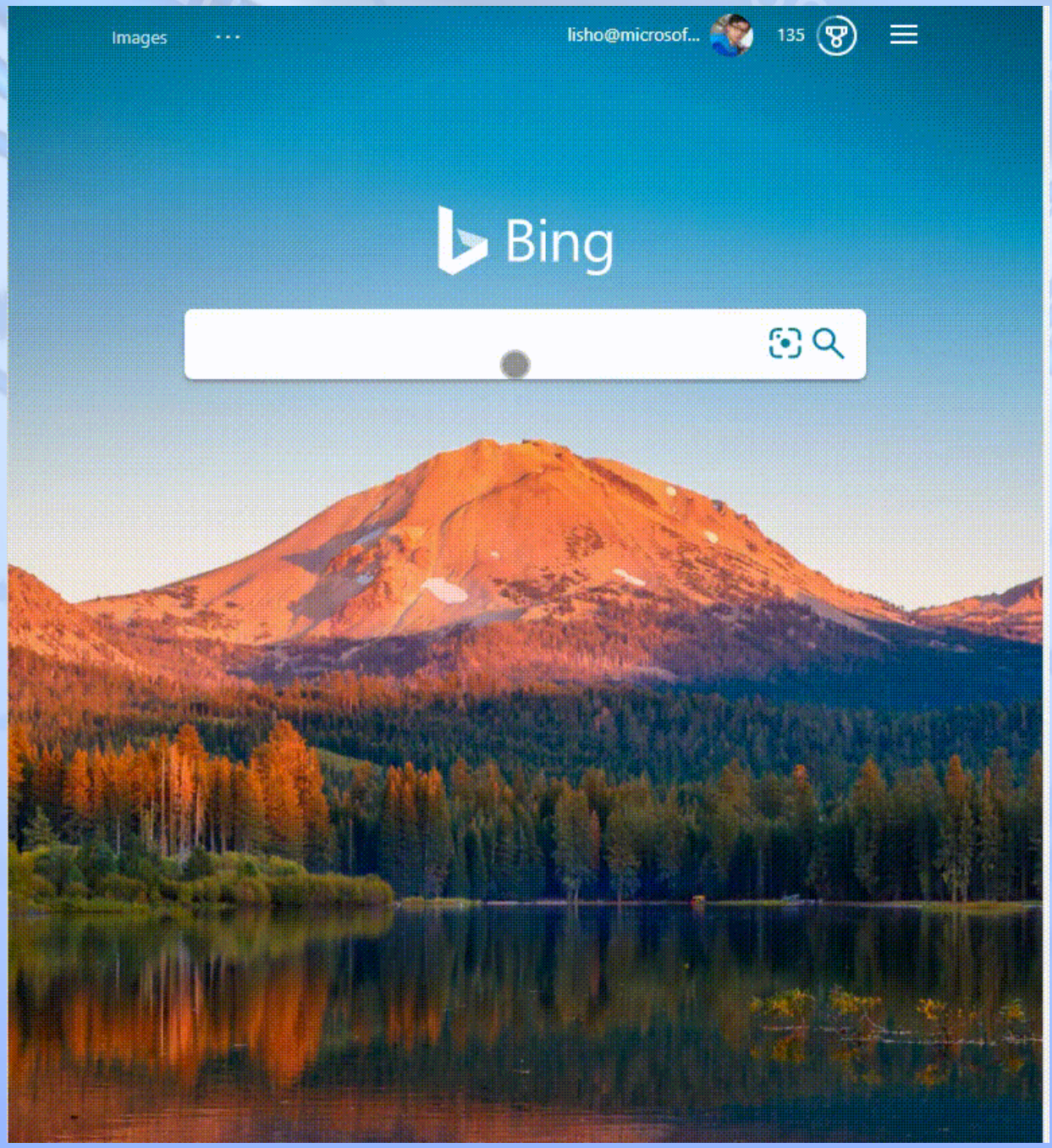
XGLUE Submission Guideline/Github | XGLUE Paper | Unicoder Paper(Baseline)

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

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

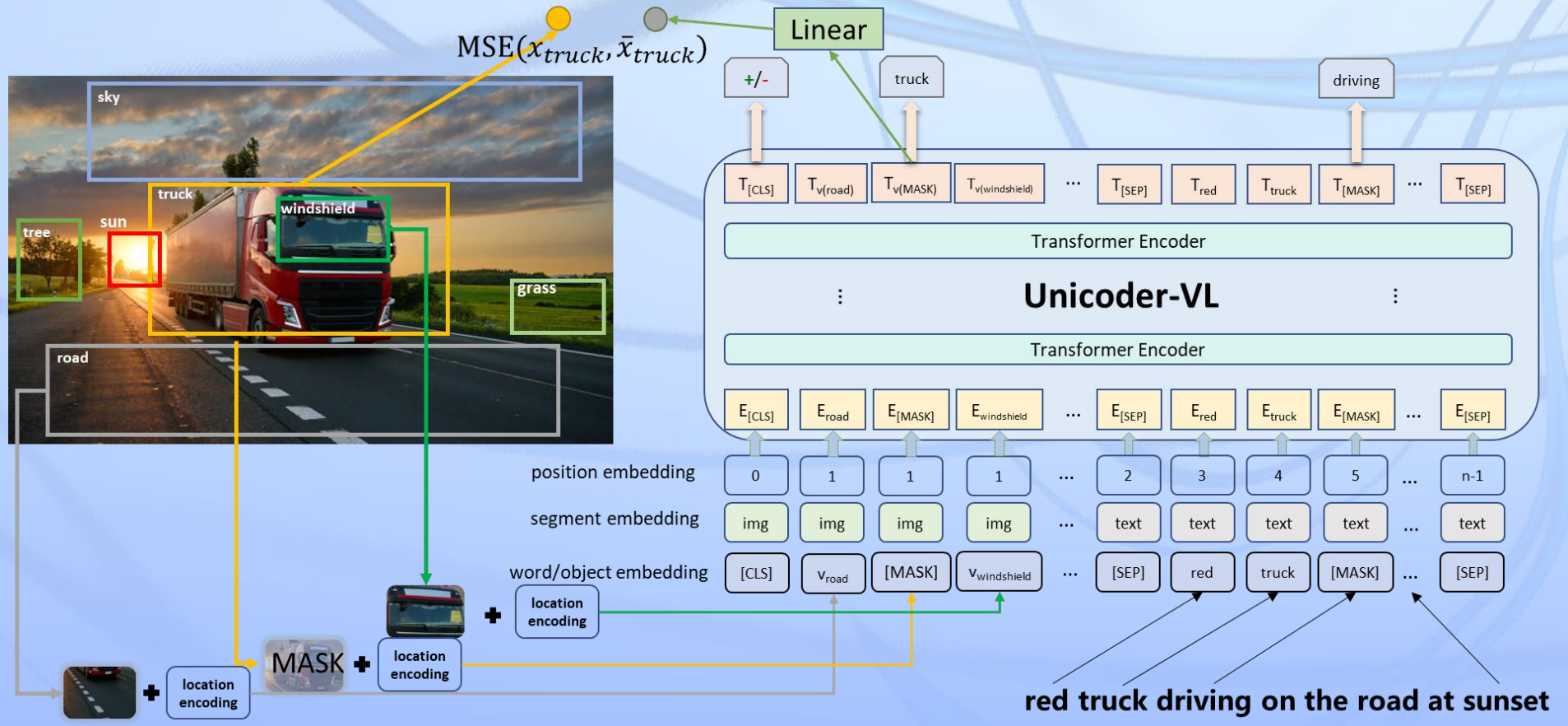
Unicoder scaled Bing intelligent question answering to 100 languages and 200 regions in the world.



Outline

- **Multilingual/Multimodal Pre-trained Models**
  - Unicoder for multilingual language tasks
  - Unicoder-VL for vision-language tasks
- **From Natural Language to Programming Language**
- **Summary & Future Work**

Unicoder-VL for Image-Language Tasks



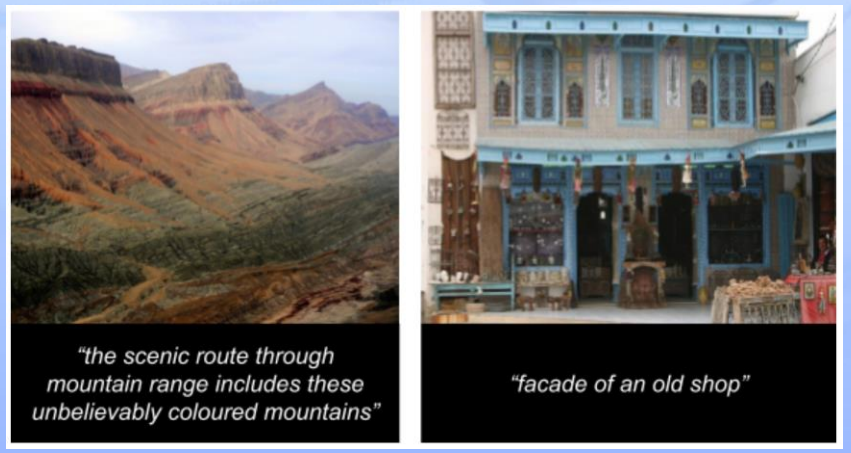
Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. AAAI 2020.

Evaluation: 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: 3.3M image-caption pairs from Google's Conceptual Captions (https://ai.google.com/research/ConceptualCaptions)



Evaluation: 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

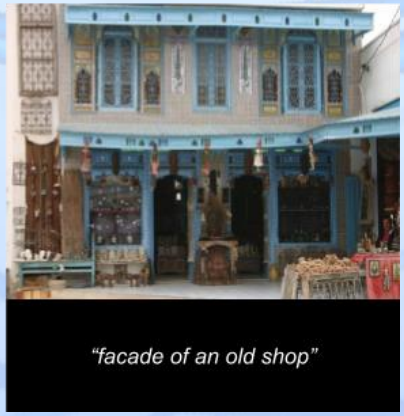
1/97

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

Evaluation: Image Captioning

Methods	Image Caption (MSCOCO)			
	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**

- Pre-training dataset: 3.3M image-caption pairs from Google's Conceptual Captions (https://ai.google.com/research/ConceptualCaptions)



Multilingual Unicoder-VL (a.k.a. $\mathbf{M^3P}$)

Pre-training Overview

Multilingual Image-Text Retrieval | Multilingual Image Captioning | Multimodal Machine Translation

Multitask Multilingual Multimodal Pre-trained Model
($\mathbf{M^3P}$ 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: Multilingual Multimodal Tasks

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)

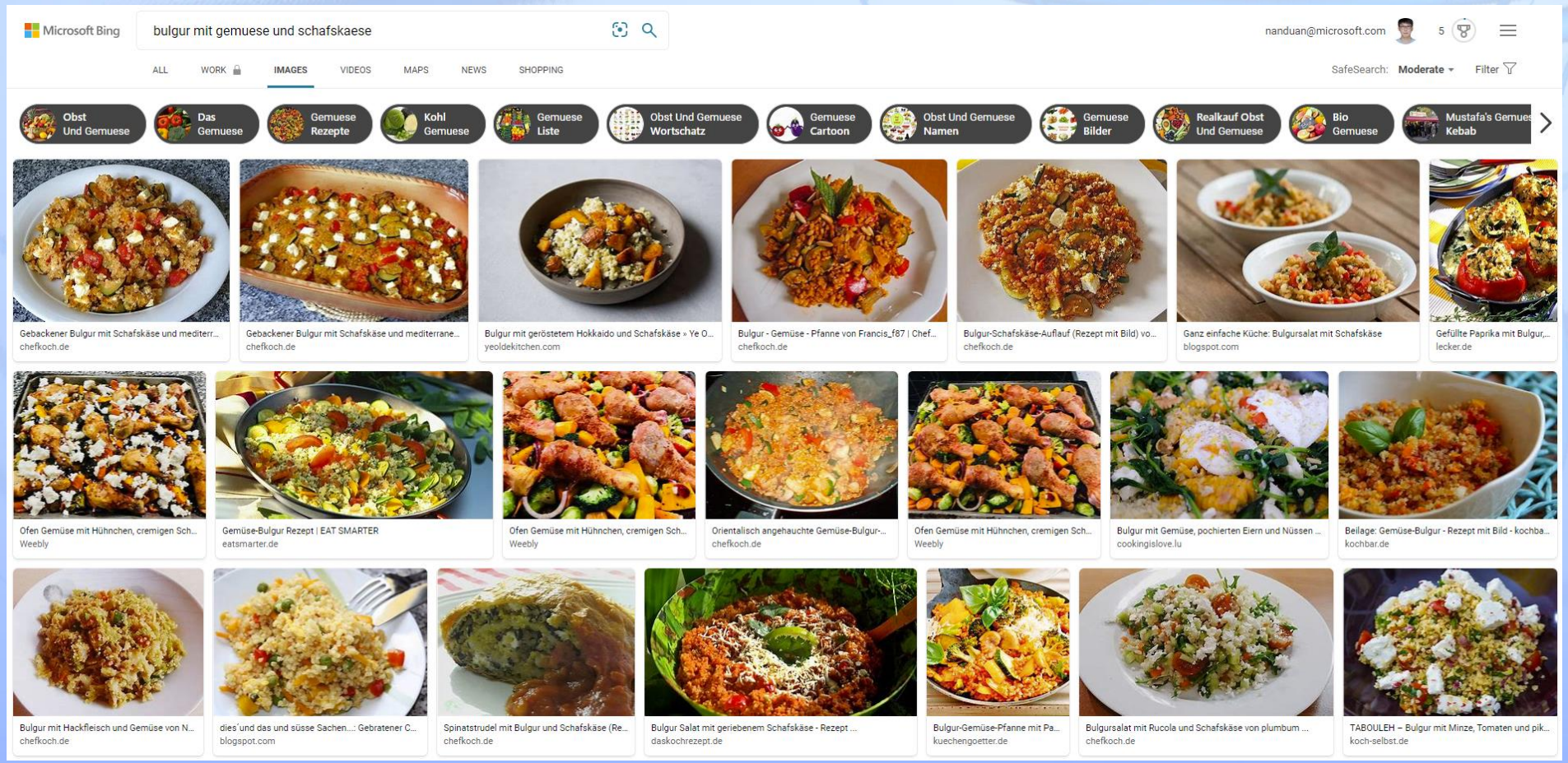
image caption input (**en**): A Boston Terrier is running on lush green grass in front of a white fence.

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.)

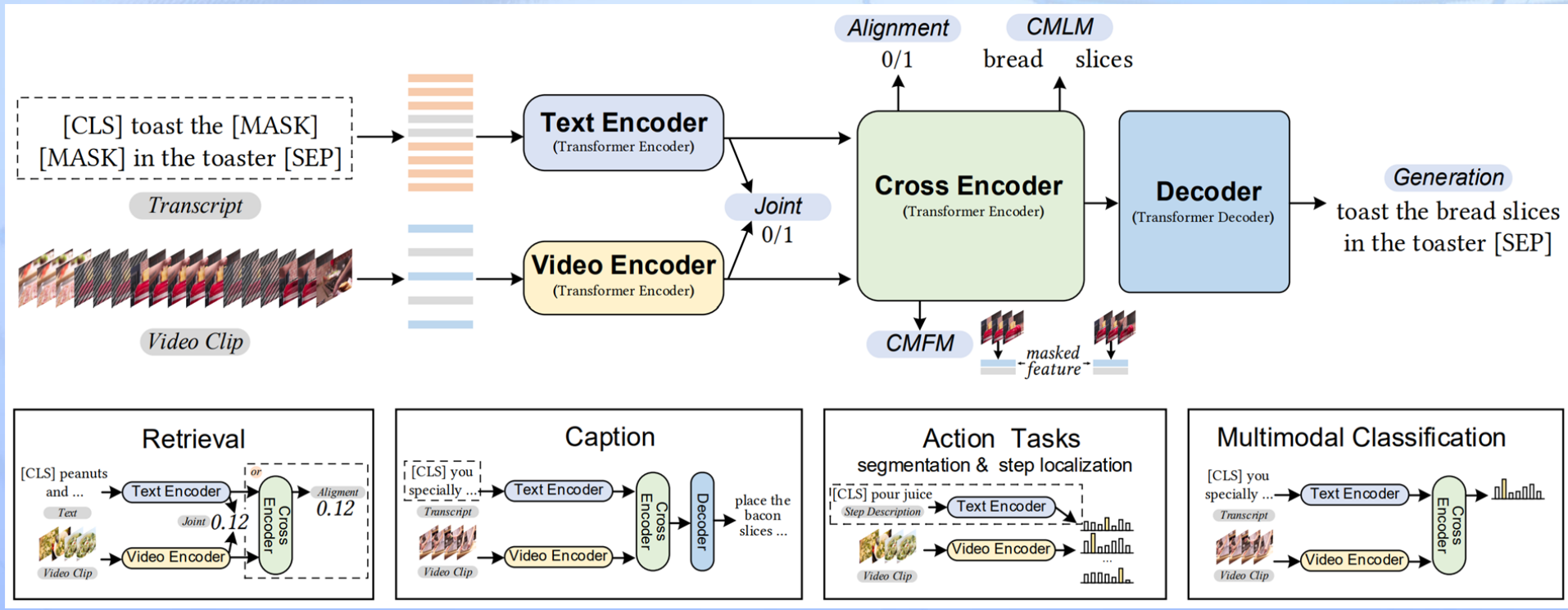
CNCC

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



Unicoder-VL for Video-Language Tasks

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: 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: 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.

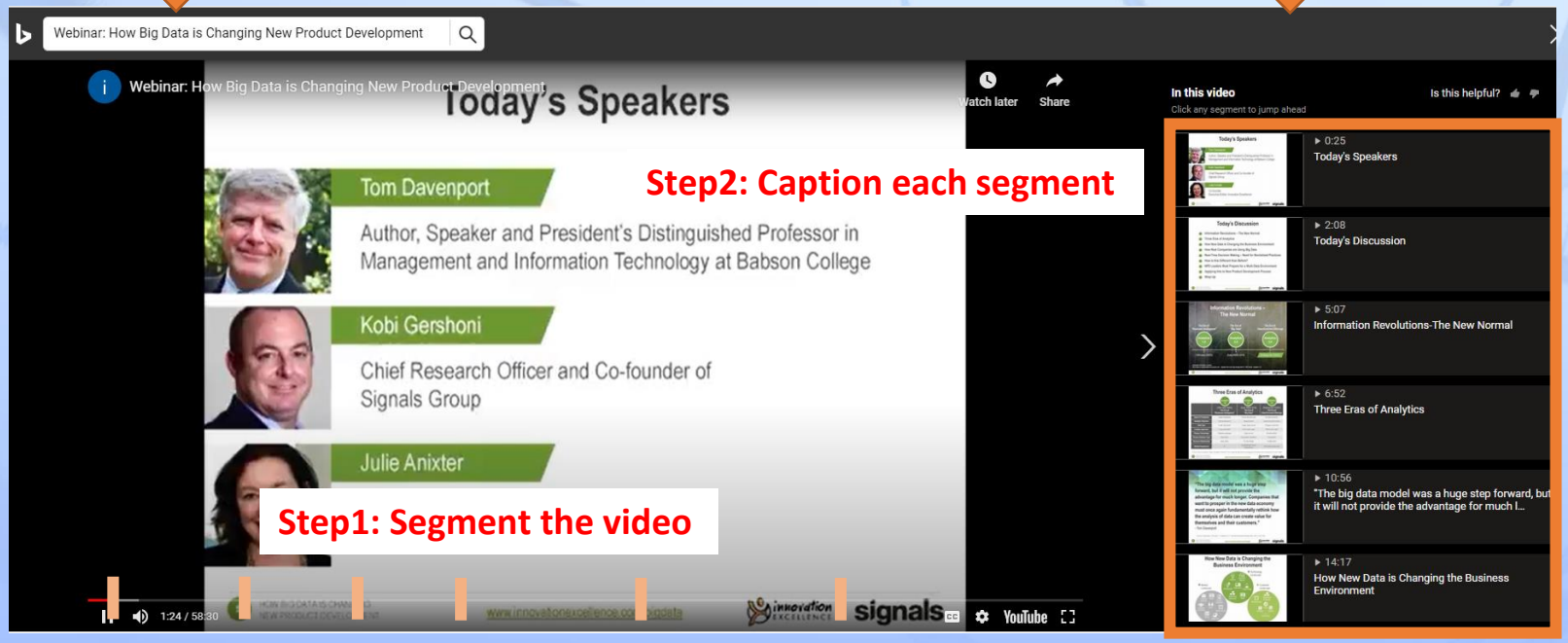
Input:

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

Unicoder-VL enables video chaptering.

Input a video

Output video chapters

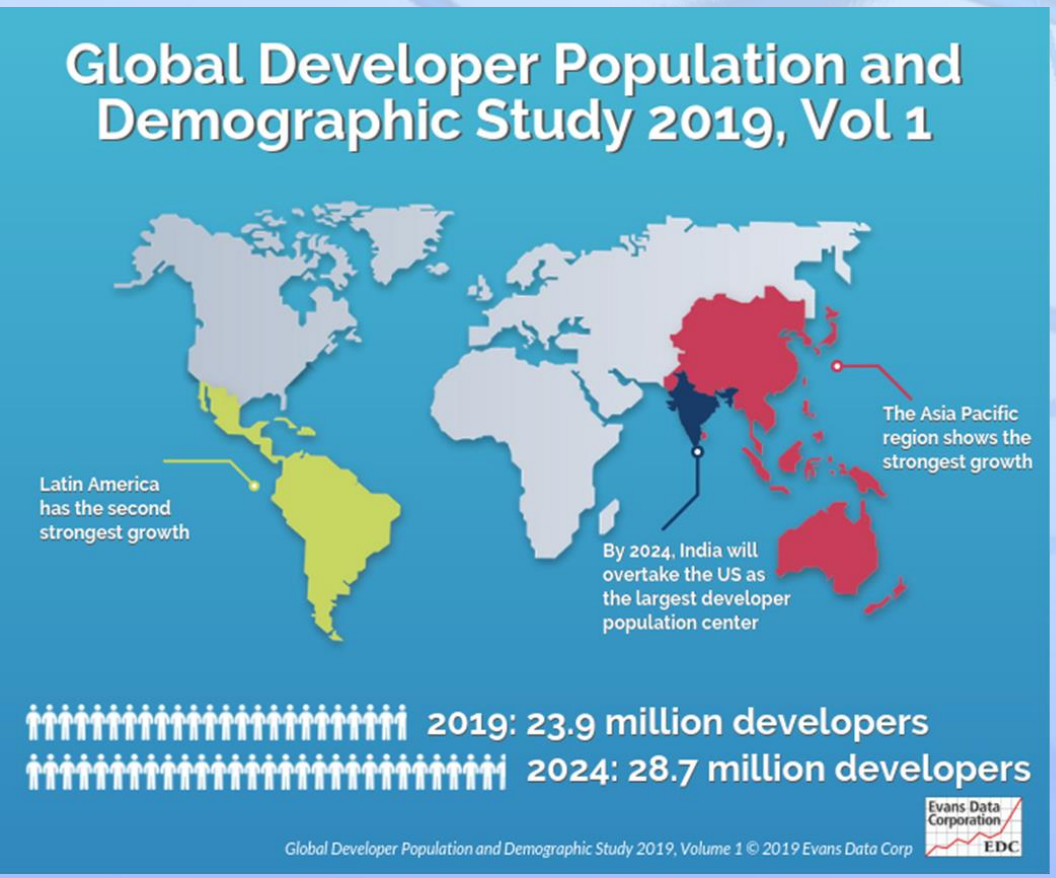


[Webinar: How Big Data is Changing New Product Development]

Outline

- **Multilingual/Multimodal Pre-trained Models**
  - Unicoder for multilingual language tasks
  - Unicoder-VL for vision-language tasks
- **From Natural Language to Programming Language**
- **Summary & Future Work**

Why Important



"There are **23.9 million** professional developers in 2019, and the population is expected to reach **28.7 million** in 2024."

https://evansdata.com/press/viewRelease.php?pressID=278

CodeBERT

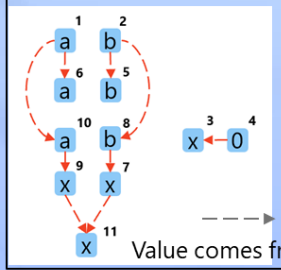
Source code

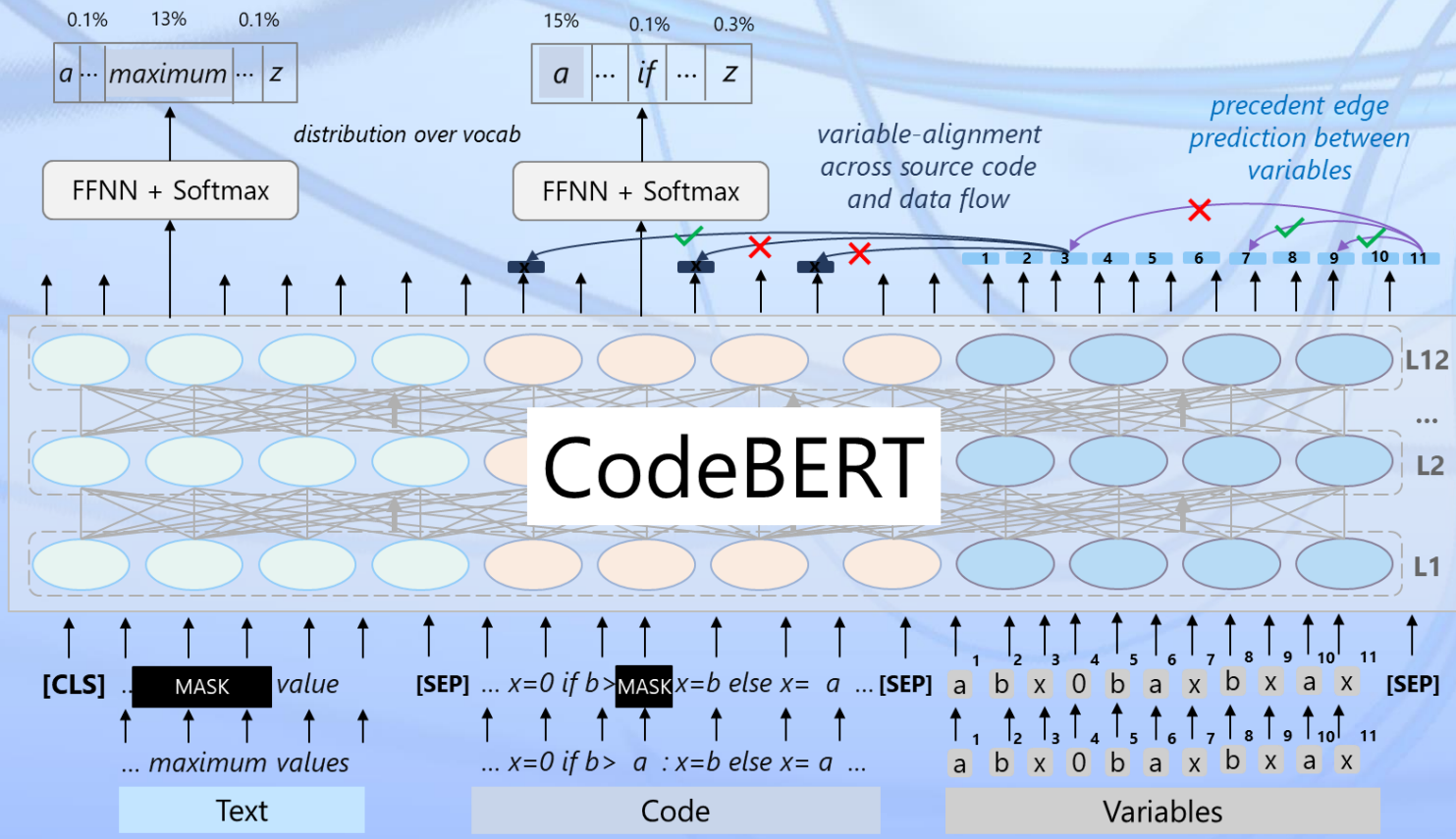
```
def max(a, b):
    x=0
    if b>a:
        x=b
    else:
        x=a
    return x
```

Comment

Return maximum value

Structure





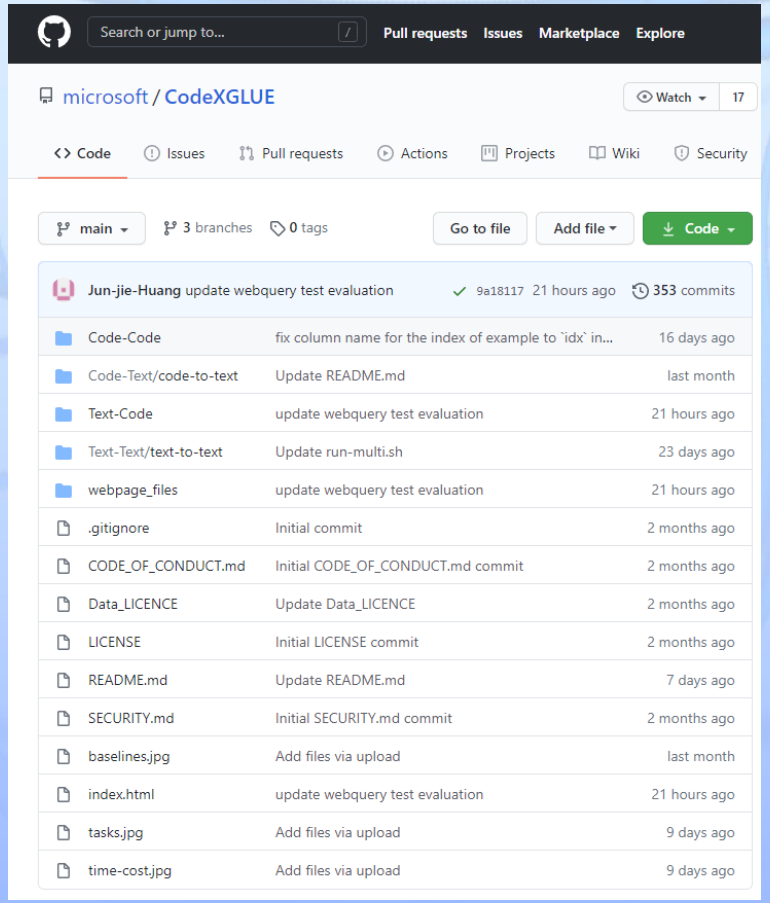
Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, Ming Zhou. **CodeBERT: A Pre-Trained Model for Programming and Natural Languages**. EMNLP 2020.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, Ming Zhou. **GraphCodeBERT: Pre-training Code Representations with Data Flow**. arXiv 2020.

CodeXGLUE: A Benchmark for Code Tasks

(https://github.com/microsoft/CodeXGLUE)

Category	Task	Dataset Name	Language	Train/Dev/Test Size	Baselines
Code-Code	Clone Detection	BigCloneBench	Java	900K/416K/416K	CodeBERT
		POJ-104	C/C++	32K/8K/12K	
	Defect Detection	Defects4J	C	21k/2.7k/2.7k	
	Cloze Testing	CT-all	Python, Java, PHP, Javascript, Ruby, Go	-/-/176k	
		CT-max/min	Python, Java, PHP, Javascript, Ruby, Go	-/-/2.6k	
	Code Completion	PY150	Python	100k/5k/50k	CodeGPT
		GitHub Java Corpus	Java	13k/7k/8k	
	Code Refinement	Bugs2Fix	Java	98K/12K/12K	Encoder-Decoder
	Code Translation	CodeTrans	Java-C#	10K/0.5K/1K	
Text-Code	NL Code Search	CodeSearchNet, AdvTest	Python	251K/9.6K/19K	CodeBERT
		StacQC, WebQueryTest	Python	2.9k/0.9k/1.9k	
	Text-to-Code Generation	CONCODE	Java	100K/2K/2K	CodeGPT
Code-Text	Code Summarization	CodeSearchNet	Python, Java, PHP, Javascript, Ruby, Go	908K/45K/53K	Encoder-Decoder
Text-Text	Document Translation	Microsoft Docs	English-Latvian/Danish/Norwegian/Chinese	156K/4K/4K	



Outline

- **Multilingual/Multimodal Pre-trained Models**
  - Unicoder for multilingual language tasks
  - Unicoder-VL for vision-language tasks
- **From Natural Language to Programming Language**
- **Summary & Future Work**

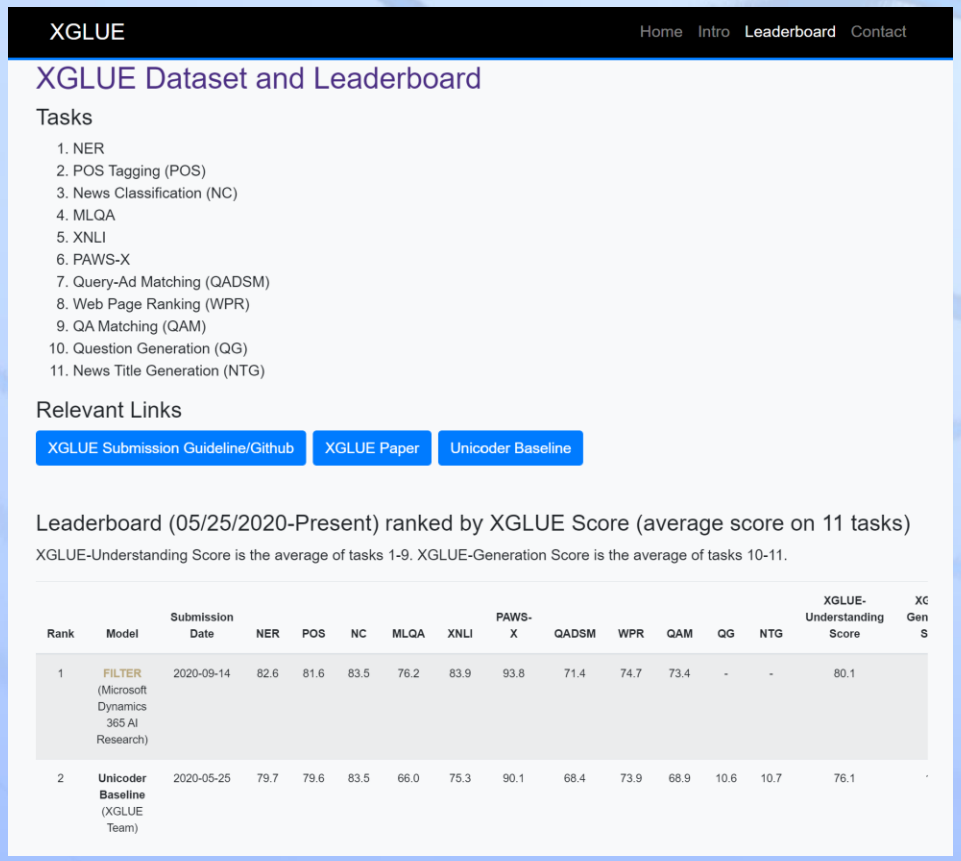
Summary & Future Work

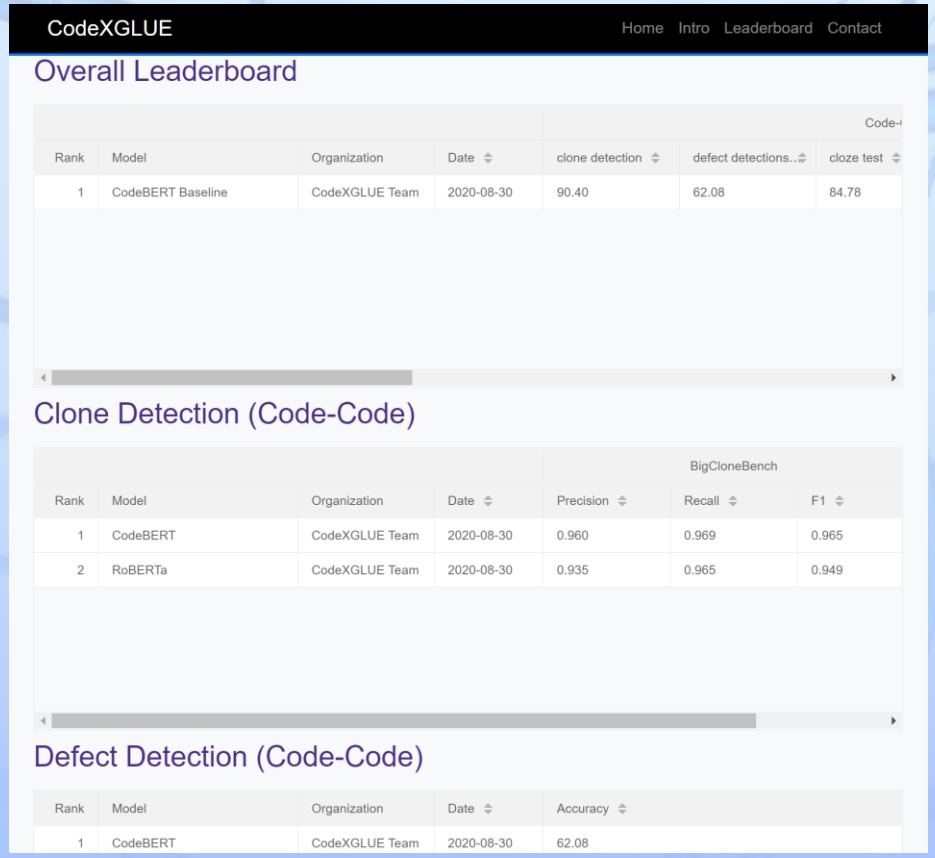
- Summary
  - alleviate low-resource issues and achieve state-of-the-art results
  - push forward AI research to new fields such as videos, code, etc.
  - have high computational costs and bad interpretability issues

- Future work
  - new tasks/architectures for multilingual/multimodality models
  - smaller sizes and faster training/inference
  - integration of knowledge for better interpretability

CNCC

Thank you and welcome to use our datasets!





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

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