State-of-the-Arts, Benchmarks and Future of Pre-trained Models for Multilingual, Multimodal, Code and Generation

Nan DUAN and Yeyun GONG
Microsoft Research Asia
CCF-ADL第116期:《大规模预训练模型》
2021-05-24

Agenda

(1) Opening (Nan, 10min)

(2) Pre-trained Model & Benchmark for Multiple Languages (Nan, 30min)

(3) Pre-trained Model & Benchmark for Language + Vision (Nan, 30min)

(4) Pre-trained Model & Benchmark for Language + Code (Nan, 20min)

---------- 10min break ----------

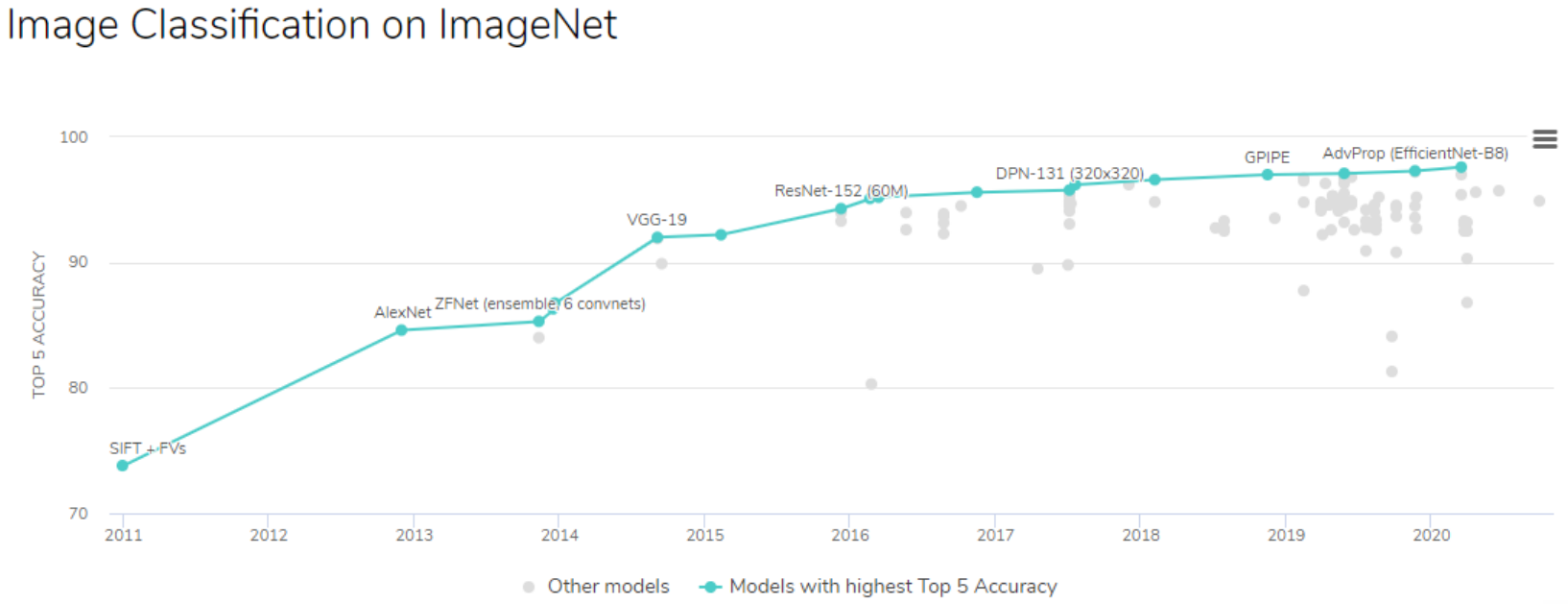
(4) Pre-trained Model & Benchmark for Language Generation (Yeyun, 80min)

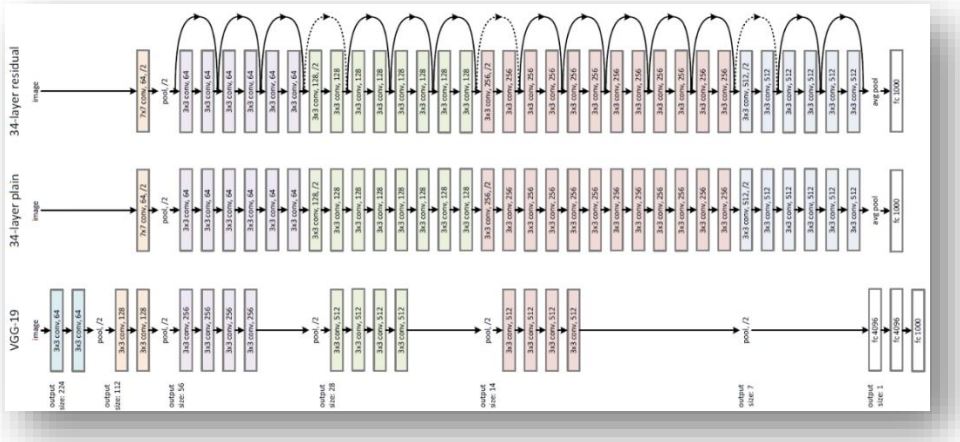
(5) Conclusion & Future Work (Yeyun, 10min)

1). Opening

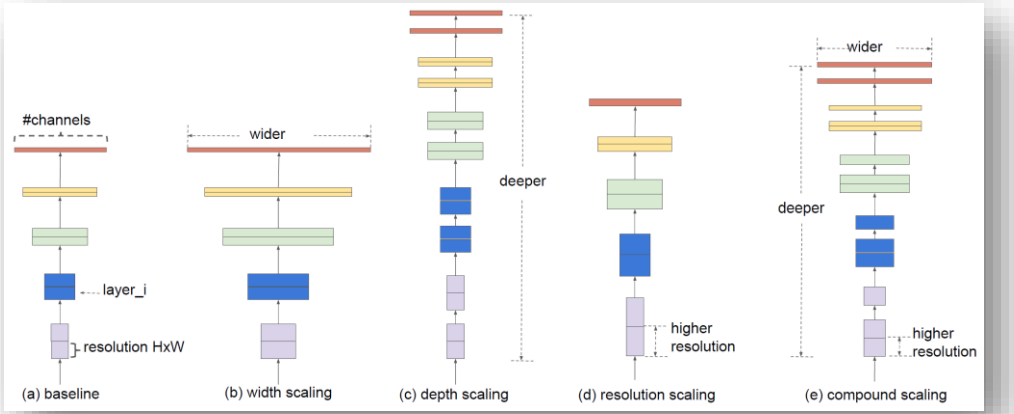
Current AI Paradigm: Pre-trained Deep Neural Networks

(achieving SOTA results on CV tasks)





ResNet (He et al., 2015)



EfficientNet (Tan and Le, 2020)

SimCLR (Chen et al., 2020)

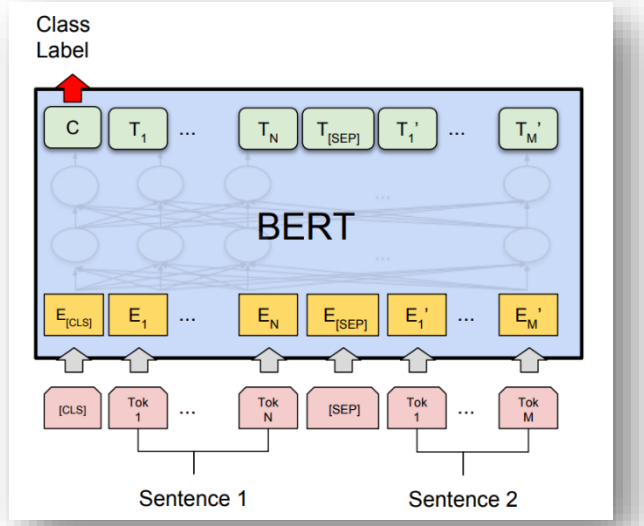
Current AI Paradigm: Pre-trained Deep Neural Networks

(achieving SOTA results on monolingual NLP tasks)

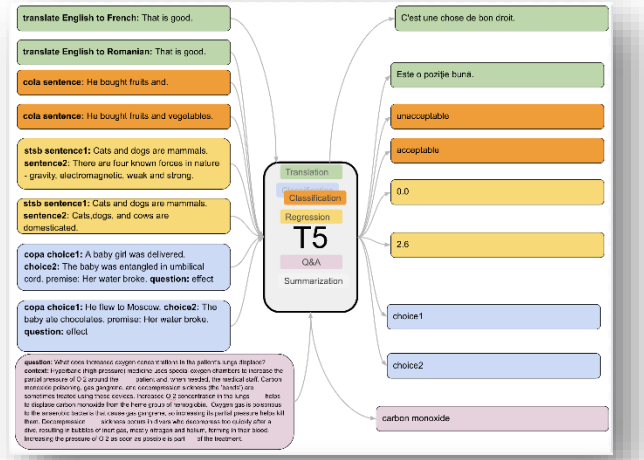
SuperGLUE GLUE | Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
✚	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
✚	3	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
✚	4	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
✚	5	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	6	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	7	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
✚	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
✚	9	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4	81.2	79.9/88.8	90.8	74.1/31.7	85.9/85.4	70.8	49.3	88.4	36.2	97.8/57.9
	10	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
	11	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
	12	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
			Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
			CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0
			Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
	-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-



BERT (Devlin et al., 2018)



T5 (Raffel et al., 2020)

Current AI Paradigm: Pre-trained Deep Neural Networks

(achieving SOTA results on multilingual NLP tasks)

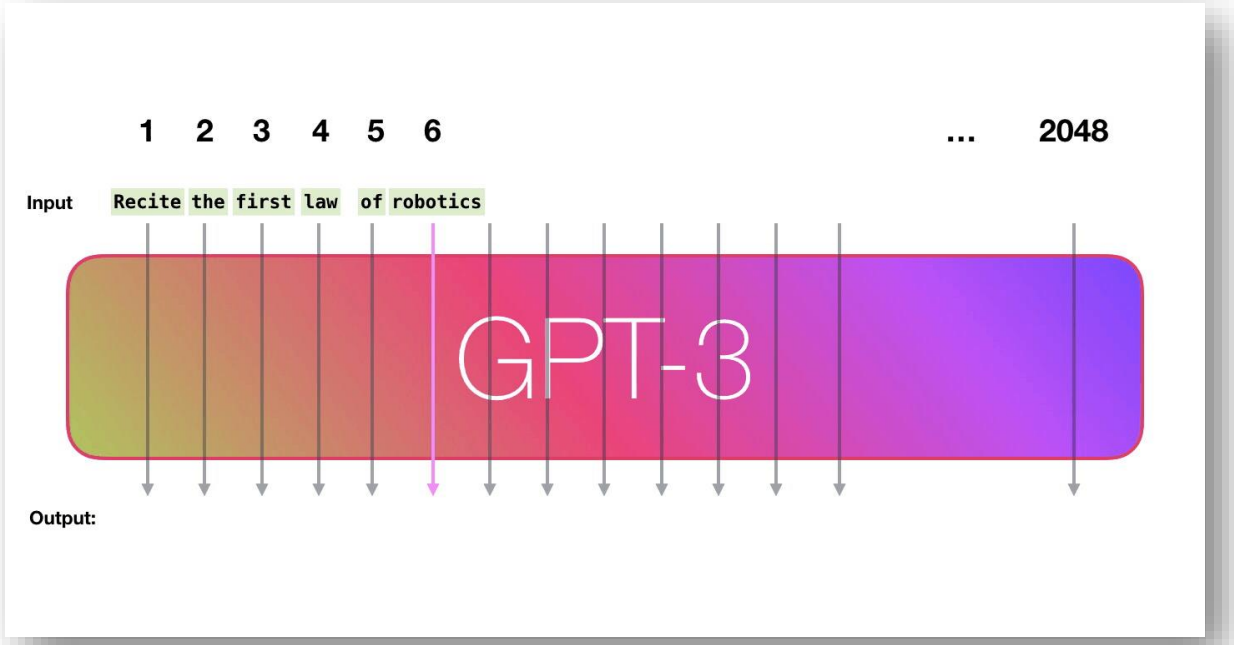
Leaderboard results

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	Unicoder + ZCode	MSRA + Cognition	Microsoft	Apr 26, 2021	81.6	88.4	76.2	72.5	93.7
2	VECO	DAMO NLP Team	Alibaba	Mar 22, 2021	81.4	88.9	75.6	72.9	92.7
3	Polyglot	MLNLC	ByteDance	Feb 25, 2021	81.3	88.1	80.6	71.8	89.4
4	ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
5	HiCTL	DAMO MT Team	Alibaba	Mar 21, 2021	80.8	89.0	74.4	71.9	92.6
6	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
7	Anonymous3	Anonymous3	Anonymous3	Jan 3, 2021	79.9	88.2	74.6	71.7	89.0
8	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
9	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
10	XLM-R	XTREME	Alphabet,						

Participate in Competition

Current AI Paradigm: Pre-trained Deep Neural Networks

(achieving surprisingly good zero/one/few-shot results on various tasks)



GPT-3 (Brown et al., 2020)

http://jalammar.github.io/how-gpt3-works-visualizations-animations/

			Fine-tune		Zero-Shot								One-Shot								Few-Shot								
Name	Metric	Split	SOTA	K	Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B	Small	Med	Large	XL	2.7B	6.7B	13B	175B	175B (test server)
HellaSwag	acc	dev	85.6	20	33.7	43.6	51.0	54.7	62.8	67.4	70.9	78.9	33.0	42.9	50.5	53.5	61.9	66.5	70.0	78.1	33.5	43.1	51.3	54.9	62.9	67.3	71.3	79.3	
LAMBADA	acc	test	68.0	15	42.7	54.3	60.4	63.6	67.1	70.3	72.5	76.2	22.0	47.1	52.6	58.3	61.1	65.4	69.0	72.5	22.0	40.4	63.2	57.0	78.1	79.1	81.3	86.4	
LAMBADA	ppl	test	8.63	15	18.6	9.09	6.53	5.44	4.60	4.00	3.56	3.00	165.0	11.6	8.29	6.46	5.53	4.61	4.06	3.35	165.0	27.6	6.63	7.45	2.89	2.56	2.56	1.92	
StoryCloze	acc	test	91.8	70	63.3	68.5	72.4	73.4	77.2	77.7	79.5	83.2	62.3	68.7	72.3	74.2	77.3	78.7	79.7	84.7	62.3	70.2	73.9	76.1	80.2	81.2	83.0	87.7	
NQs	acc	test	44.5	64	0.64	1.75	2.71	4.40	6.01	5.79	7.84	14.6	1.19	3.07	4.79	5.43	8.73	9.78	13.7	23.0	1.72	4.46	7.89	9.72	13.2	17.0	21.0	29.9	
TriviaQA	acc	dev	68.0	64	4.15	7.61	14.0	19.7	31.3	38.7	41.8	64.3	4.19	12.9	20.5	26.5	35.9	44.4	51.3	68.0	6.96	16.3	26.5	32.1	42.3	51.6	57.5	71.2	71.2
WebQs	acc	test	45.5	64	1.77	3.20	4.33	4.63	7.92	7.73	8.22	14.4	2.56	6.20	8.51	9.15	14.5	15.1	19.0	25.3	5.46	12.6	15.9	19.6	24.8	27.7	33.5	41.5	
Ro→En 16	BLEU-mb	test	39.9	64	2.08	2.71	3.09	3.15	16.3	8.34	20.2	19.9	0.55	15.4	23.0	26.3	30.6	33.2	35.6	38.6	1.25	20.7	25.8	29.2	33.1	34.8	37.0	39.5	
Ro→En 16	BLEU-sb	test		64	2.39	3.08	3.49	3.56	16.8	8.75	20.8	20.9	0.65	15.9	23.6	26.8	31.3	34.2	36.7	40.0	1.40	21.3	26.6	30.1	34.3	36.2	38.4	41.3	
En→Ro 16	BLEU-mb	test	38.5	64	2.14	2.65	2.53	2.50	3.46	4.24	5.32	14.1	0.35	3.30	7.89	8.72	13.2	15.1	17.3	20.6	1.25	5.90	9.33	10.7	14.3	16.3	18.0	21.0	
En→Ro 16	BLEU-sb	test		64	2.61	3.11	3.07	3.09	4.26	5.31	6.43	18.0	0.55	3.90	9.15	10.3	15.7	18.2	20.8	24.9	1.64	7.40	10.9	12.9	17.2	19.6	21.8	25.8	
Fr→En 14	BLEU-mb	test	35.0	64	1.81	2.53	3.47	3.13	20.6	15.1	21.8	21.2	1.28	15.9	23.7	26.3	29.0	30.5	30.2	33.7	4.98	25.5	28.5	31.1	33.7	34.9	36.6	39.2	
Fr→En 14	BLEU-sb	test		64	2.29	2.99	3.90	3.60	21.2	15.5	22.4	21.9	1.50	16.3	24.4	27.0	30.0	31.6	31.4	35.6	5.30	26.2	29.5	32.2	35.1	36.4	38.3	41.4	
En→Fr 14	BLEU-mb	test	45.6	64	1.74	2.16	2.73	2.15	15.1	8.82	12.0	25.2	0.49	8.00	14.8	15.9	20.3	23.3	24.9	28.3	4.08	14.5	19.3	21.5	24.9	27.3	29.5	32.6	
En→Fr 14	BLEU-sb	test	45.9	64	2.44	2.75	3.54	2.82	19.3	11.4	15.3	31.3	0.81	10.0	18.2	19.3	24.7	28.3	30.1	34.1	5.31	18.0	23.6	26.1	30.3	33.3	35.5	39.9	
De→En 16	BLEU-mb	test	40.2	64	2.06	2.87	3.41	3.63	21.5	17.3	23.0	27.2	0.83	16.2	22.5	24.7	28.2	30.7	33.0	30.4	3.25	22.7	26.2	29.2	32.7	34.8	37.3	40.6	
De→En 16	BLEU-sb	test		64	2.39	3.27	3.85	4.04	22.5	18.2	24.4	28.6	0.93	17.1	23.4	25.8	29.2	31.9	34.5	32.1	3.60	23.8	27.5	30.5	34.1	36.5	39.1	43.0	
En→De 16	BLEU-mb	test	41.2	64	1.70	2.27	2.31	2.43	12.9	8.66	10.4	24.6	0.50	7.00	12.9	13.1	18.3	20.9	22.5	26.2	3.42	12.3	15.4	17.1	20.9	23.0	26.6	29.7	
En→De 16	BLEU-sb	test	41.2	64	2.09	2.65	2.75	2.92	13.7	9.36	11.0	25.3	0.54	7.40	13.4	13.4	18.8	21.7	23.3	27.3	3.78	12.9	16.1	17.7	21.7	24.1	27.7	30.9	
Winograd	acc	test	93.8	7	66.3	72.9	74.7	76.9	82.4	85.7	87.9	88.3	63.4	68.5	72.9	76.9	82.4	84.6	86.1	89.7	63.4	67.4	73.6	76.9	84.3	85.4	82.4	88.6	
Winogrande	acc	dev	84.6	50	52.0	52.1	57.4	58.7	62.3	64.5	67.9	70.2	51.3	53.0	58.3	59.1	61.7	65.8	66.9	73.2	51.3	52.6	57.5	59.1	62.6	67.4	70.0	77.7	
PIQA	acc	dev	77.1	50	64.6	70.2	72.9	75.1	75.6	78.0	78.5	81.0	64.3	69.3	71.8	74.4	74.3	76.3	77.8	80.5	64.3	69.4	72.0	74.3	75.4	77.8	79.9	82.3	82.8
ARC (Challenge)	acc	test	78.5	50	26.6	29.5	31.8	35.5	38.0	41.4	43.7	51.4	25.5	30.2	31.6	36.4	38.4	41.5	43.1	53.2	25.5	28.4	32.3	36.7	39.5	43.7	44.8	51.5	
ARC (Easy)	acc	test	92.0	50	43.6	46.5	53.0	53.8	58.2	60.2	63.8	68.8	42.7	48.2	54.6	55.9	60.3	62.6	66.8	71.2	42.7	51.0	58.1	59.1	62.1	65.8	69.1	70.1	
OpenBookQA	acc	test	87.2	100	35.6	43.2	45.2	46.8	53.0	50.4	55.6	57.6	37.0	39.8	46.2	46.4	53.4	53.0	55.8	58.8	37.0	43.6	48.0	50.6	55.6	55.2	60.8	65.4	
Quac	f1	dev	74.4	5	21.2	26.8	31.0	30.1	34.7	36.1	38.4	41.5	21.1	26.9	31.9	32.3	37.4	39.0	40.6	43.4	21.6	27.6	32.9	34.2	38.2	39.9	40.9	44.3	
RACE-h	acc	test	90.0	10	35.2	37.9	40.1	40.9	42.4	44.1	44.6	45.5	34.3	37.7	40.0	42.0	43.8	44.3	44.6	45.9	34.3	37.0	40.4	41.4	42.3	44.7	45.1	46.8	
RACE-m	acc	test	93.1	10	42.1	47.2	52.1	52.3	54.7	54.4	56.7	58.4	42.3	47.3	51.7	55.2	56.1	54.7	56.9	57.4	42.3	47.0	52.7	53.0	55.6	55.4	58.1	58.1	
SQuADv2	em	dev	90.7	16	22.6	32.8	33.9	43.1	43.6	45.4	49.0	52.6	25.1	37.5	37.9	47.9	47.9	51.1	56.0	60.1	27.5	40.5	39.2	53.5	50.0	56.6	62.6	64.9	
SQuADv2	f1	dev	93.0	16	28.3	40.2	41.4	50.3	51.0	52.7	56.3	59.5	30.1	43.6	44.1	54.0	54.1	57.1	61.8	65.4	32.1	45.5	44.9	58.7	55.9	62.1	67.7	69.8	
CoQA	f1	dev	90.7	5	34.5	55.0	61.8	65.3	71.1	72.8	76.3	81.5	30.6	52.1	61.6	66.1	71.8	75.1	77.9	84.0	31.1	52.0	62.7	66.8	73.2	77.3	79.9	85.0	
DROP	f1	dev	89.1	20	9.40	13.6	14.4	16.4	19.7	17.0	24.0	23.6	11.7	18.1	20.9	23.0	26.4	27.3	29.2	34.3	12.9	18.7	24.0	25.6	29.7	29.7	32.3	36.5	
BoolQ	acc	dev	91.0	32	49.7	60.3	58.9	62.4	67.1	65.4	66.2	60.5	52.6	61.7	60.4	63.7	68.4	68.7	69.0	76.7	43.1	60.6	62.0	64.1	70.3	70.0	70.2	77.5	76.4
CB	acc	dev	96.9	32	0.00	32.1	8.93	19.6	19.6	28.6	19.6	46.4	55.4	53.6	53.6	48.2	57.1	33.9	55.4	64.3	42.9	58.9	53.6	69.6	67.9	60.7	66.1	82.1	75.6
CB	f1	dev	93.9	32	0.00	29.3	11.4	17.4	22.4	25.1	20.3	42.8	60.1	39.8	45.6	37.5	45.7	28.5	44.6	52.5	26.1	40.4	32.6	48.3	45.7	44.6	46.0	57.2	52.0
Copa	acc	dev	94.8	32	66.0	68.0	73.0	77.0	76.0	80.0	84.0	91.0	62.0	64.0	66.0	74.0	76.0	82.0	86.0	87.0	67.0	64.0	72.0	77.0	83.0	83.0	86.0	92.0	92.0
RTE	acc	dev	92.5	32	47.7	49.8	48.4	56.0	46.6	55.2	62.8	63.5	53.1	47.3	49.5	49.5	54.9	54.9	56.3	70.4	52.3	48.4	46.9	50.9	56.3	49.5	60.6	72.9	69.0
WiC	acc	dev	76.1	32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.0	50.3	50.3	49.2	49.4	50.3	50.0	48.6	49.8	55.0	53.0	53.0	51.6	53.1	51.1	55.3	49.4
WSC	acc	dev	93.8	32	59.6	56.7	65.4	61.5	66.3	60.6	64.4	65.4	58.7	58.7	60.6	62.5	66.3	60.6	66.3	69.2	58.7	60.6	54.8	49.0	62.5	67.3	75.0	75.0	80.1
MultiRC	acc	dev	62.3	32	4.72	9.65	12.3	13.6	14.3	18.4	24.2	27.6	4.72	9.65	12.3	13.6	14.3	18.4	24.2	27.6	6.09	11.8	16.8	20.8	24.7	23.8	25.0	32.5	30.5
MultiRC	f1a	dev	88.2	32	57.0	59.7	60.4	59.9	60.0	64.5	71.4	72.9	57.0	59.7	60.4	59.9	60.0	64.5	71.4	72.9	45.0	55.9	64.2	65.4	69.5	66.4	69.3	74.8	75.4
ReCoRD	acc	dev	92.5	32	70.8	78.5	82.1	84.1	86.2	88.6	89.0	90.2	69.8	77.0	80.7	83.0	85.9	88.0	88.8	90.2	69.8	77.2	81.3	83.1	86.6	87.9	88.9	89.0	90.2
ReCoRD	f1	dev	93.3	32	71.9	79.2	82.8	85.2	87.3	89.5	90.4	91.0	70.7	77.8	81.6	83.9	86.8	88.8	89.7	91.2	70.7	77.9	82.1	84.0	87.5	88.8	89.8	90.1	91.1
SuperGLUE	average	dev	89.0		40.6	47.4	46.8	49.6	50.1	52.3	54.4	58.2	54.4	55.1	56.7	57.8	61.2	59.7	64.3	68.9	50.2	56.2	56.8	60.0	64.3	63.6	66.9	73.2	71.8
ANLI R1	acc	test	73.8	50	33.4	34.2	33.4	33.4	34.2	32.3	33.2	34.6	32.1	31.6	31.9	34.6	30.6	31.6	32.7	32.0	32.1	32.5	30.9	32.5	33.5	33.1	33.3	36.8	
ANLI R2	acc	test	50.7	50	33.2	31.9	33.3	33.3	33.8	33.5	33.5	35.4	35.7	33.7	33.2	32.7	32.7	33.9	33.9	33.9	35.7	33.8	32.1	31.4	32.6	33.3	32.6	34.0	
ANLI R3	acc	test	48.3	50	33.6	34.0	33.8	33.4	35.3	34.8	34.4	34.5	35.0	32.6	33.0	33.9	34.1	33.1	32.5	35.1	35.0	34.4	35.1	36.0	32.7	33.9	34.5	40.2	
2D+	acc	n/a		50	0.70	0.65	0.70	0.85	1.10	2.54	15.4	76.9	2.00	0.55	3.15	4.00	12.1	19.6	73.0	99.6	2.00	4.10	3.50	4.50	8.90	11.9	55.5	100.0	
2D-	acc	n/a		50	1.25	1.25	1.25	1.25	1.60	7.60	12.6	58.0	1.15	0.95	1.45	1.95	3.85	11.5	44.6	86.4	1.15	1.45	2.25	2.70	7.35	13.6	52.4	98.9	
3D+	acc	n/a		50	0.10	0.10	0.05	0.10	0.10	0.25	1.40	34.2	0.15	0.00	0.10	0.30	0.45	0.95	15.4	65.5	0.15	0.45	0.30	0.55	0.75	0.90	8.40	80.4	
3D-	acc	n/a		50	0.05	0.05	0.05	0.05	0.05	0.45	1.35	48.3	0.05	0.15	0.25	0.30	0.55	1.60	6.15	78.7	0.05	0.10	0.15	0.35	0.65	1.05	9.20	94.2	
4D+	acc	n/a		50	0.05	0.05	0.00	0.00	0.05	0.05	0.15	4.00	0.00	0.00	0.10	0.00	0.00	0.10	0.80	14.0	0.00	0.05	0.05	0.00	0.15	0.15	0.40	25.5	
4D-	acc	n/a		50	0.00	0.00	0.00	0.00	0.00	0.00	0.10	7.50	0.00	0.00	0.00	0.00	0.05	0.00	0.50	14.0	0.00	0.05	0.00	0.00	0.10	0.05	0.40	26.8	
5D+	acc	n/a		50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.05	3.45	0.00	0.00	0.00	0.00	0.00	0.00	0.05	9.30	
5D-	acc	n/a		50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.05	3.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.90	
2Dx	acc	n/a		50	2.20	2.25	2.65	2.10	2.55	5.80	6.15	19.8	1.35	2.35	3.35	2.35	4.75	9.15	11.0	27.4	1.35	2.90	2.70	2.85	4.25	6.10	7.05	29.2	
1DC	acc	n/a		50	1.25	2.95	2.75	0.05	0.30	2.35	0.75	9.75	1.90	2.80	2.85	3.65	6.45	9.15	8.20	14.3	1.70	2.15	3.90	5.75	6.20	7.60	9.95	21.3	
Cycled Letters	acc	n/a		100	0.62	0.71	2.85	0.00	0.63	1.35	2.58	3.66	1.67	4.36	5.68	6.46	6.25	9.41	15.1	21.7	4.63	9.27	10.7	14.5	16.7	21.9	27.7	37.9	
Anagrams 1	acc	n/a		100	0.10	0.14	0.40	0.00	0.27	0.69	1.16	2.28	0.21	0.61	1.12	1.27	1.60	2.72	3.72	8.62	0.50	1.27	2.13	3.05	3.81	5.49	8.38	15.1	
Anagrams 2	acc	n/a		100	0.81	1.21	2.69	0.01	1.71	3.75	4.53	8.91	1.19	2.62	4.70	4.77	6.97	10.2	14.6	25.9	1.94	4.80	7.59	9.87	12.6	18.9	25.6	39.7	
Symbol Insertion	acc	n/a		100	0.00	0.00	0.10	0.00	0.05	0.42	0.89	8.26	0.03	0.05	0.57	1.18	1.67	3.46	6.62	45.4	0.11	0.28	2.19	4.18	6.61	11.0	27.3	67.2	
Reversed Words	acc	n/a		100	0.00	0.01	0.01	0.01	0.02	0.03	0.03	0.09	0.02	0.01	0.01	0.00	0.05	0.07	0.11	0.48	0.00	0.05	0.00	0.17	0.24	0.30	0.42	0.44	
SAT Analogies	acc	n/a		20	35.6	39.0	45.2	44.1	50.0	49.2	52.7	53.7	30.5	41.2	43.1	46.5	55.1	54.3	53.5	59.1	30.5	40.4	42.8	40.6	48.4	51.9	53.5	65.2	

Current AI Paradigm: Pre-trained Deep Neural Networks

(achieving surprisingly good content creation capabilities)

OpenAI's DALL·E

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES

Edit prompt or view more images

TEXT PROMPT

a store front that has the word 'openai' written on it [...]

AI-GENERATED IMAGES

Edit prompt or view more images

Microsoft's GODIVA

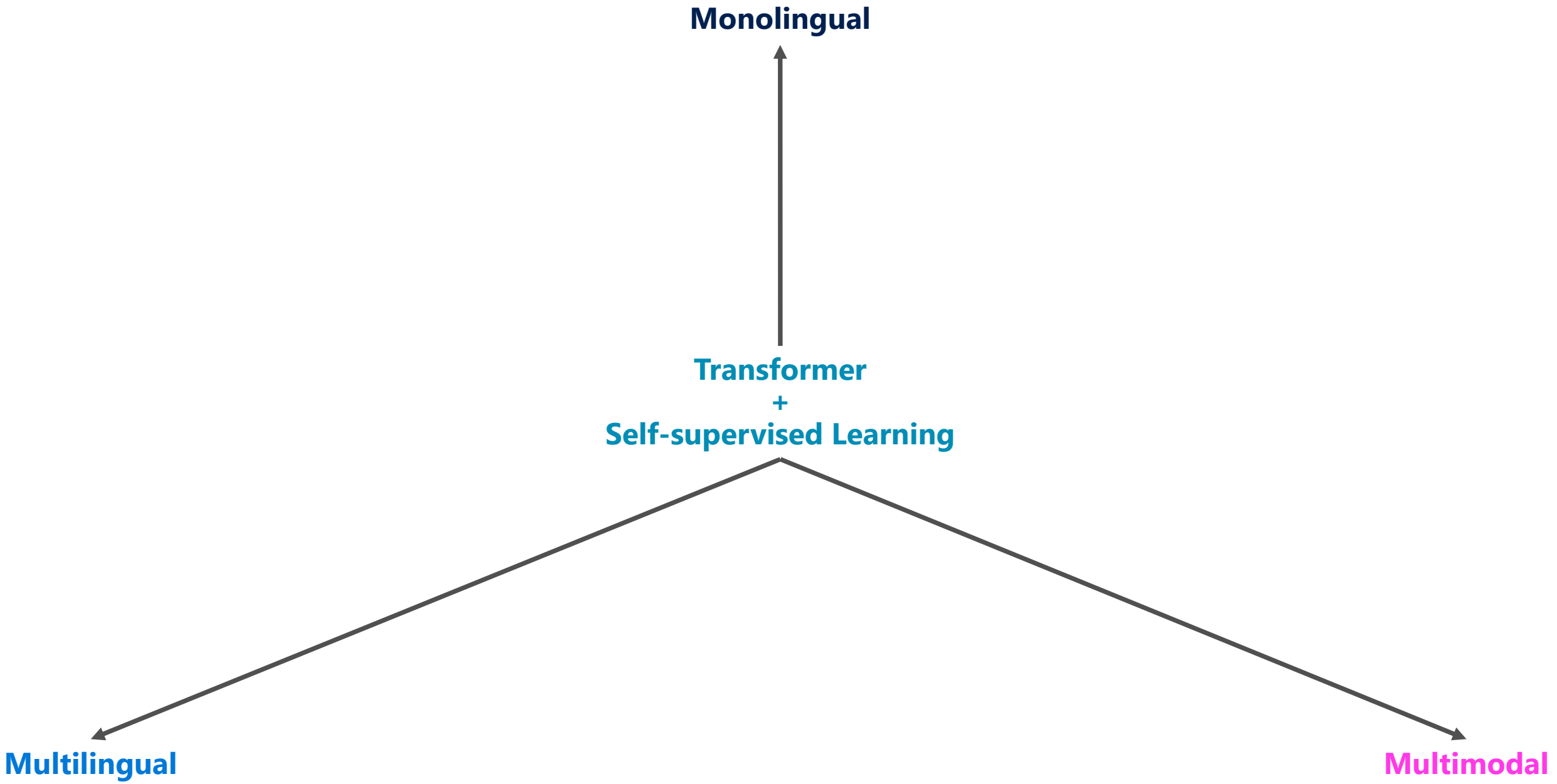
A baseball game is played. →

A girl on the voice kids talks to the judges. →

Digit 9 is moving down then up. →

Digit 7 moves right then left while digit 3 moves down then up. →

3 Key Scenarios, and Multimodal is the Trend.



Goal of Pre-training

Learn universal representations

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

Self-Supervised Learning

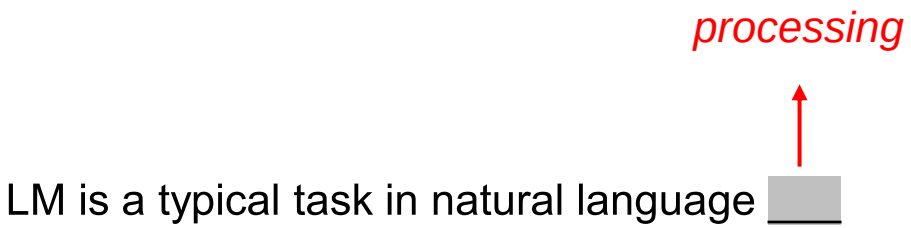
A form of unsupervised learning where the data itself provides the supervision.

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

1. Auto-regressive

Maximize the likelihood under the forward auto-regressive factorization.

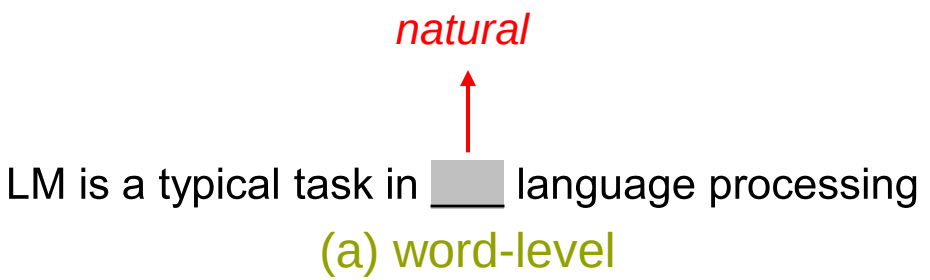
$$\max_{\theta} \mathbb{E}_{\mathbf{w}\sim D} \sum_{t=1}^{|\mathbf{w}|} \log p_\theta(w_t | w_{<t})$$



2. Denoising Auto-encoding

Reconstruct masked words from corrupted inputs.

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



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

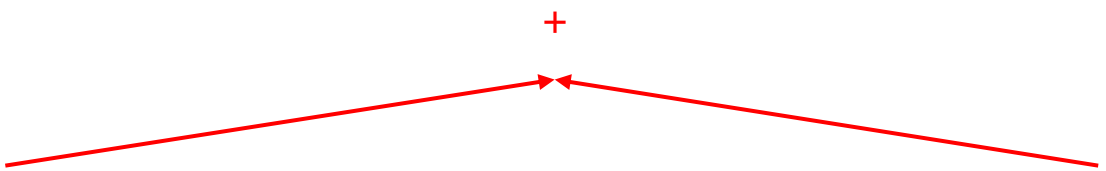
LM is a ~~typical~~ task in processing ___ language

(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。

So many good work, but we will focus on these ones this time ☺

Models:

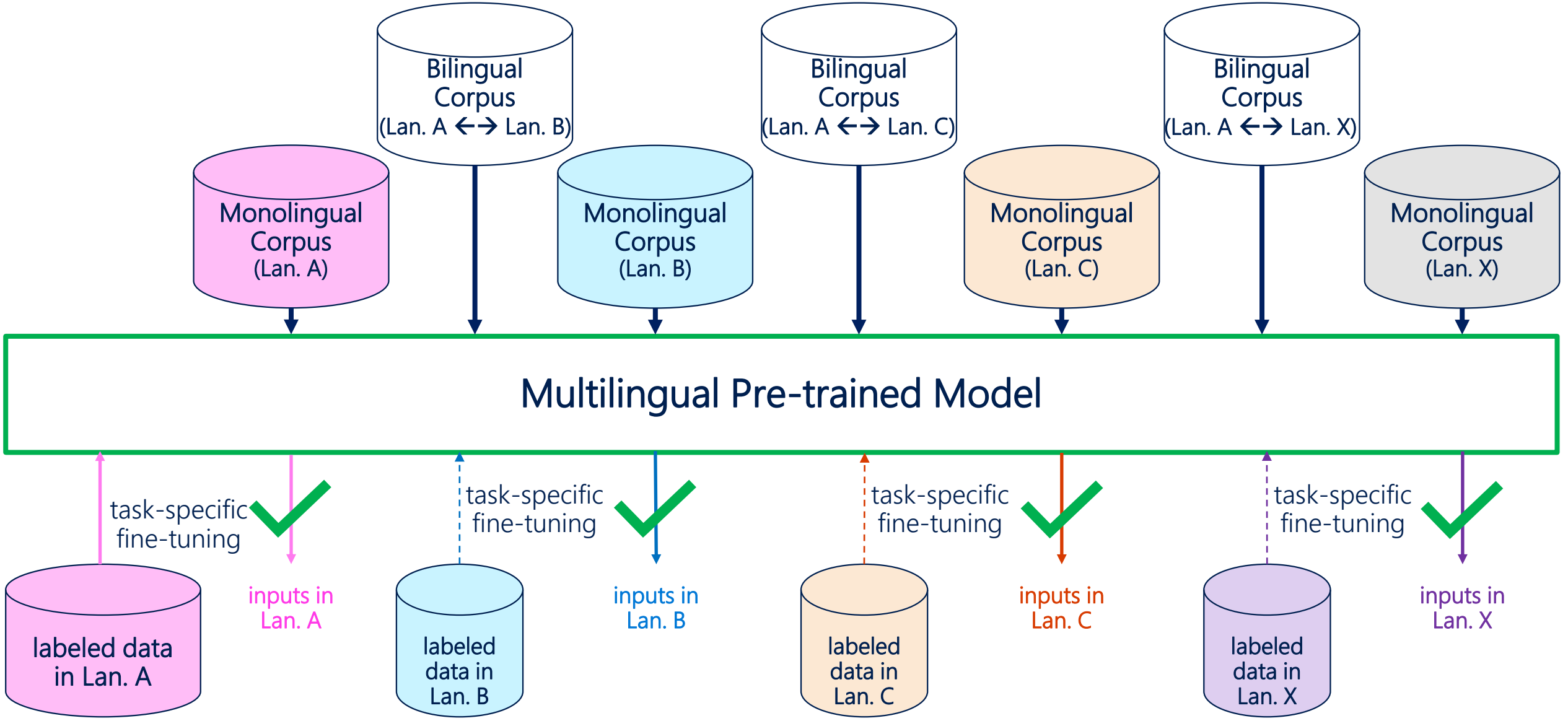
- Unicoder, Unicoder-VL/M3P, GODIVA, CodeBERT, ProphetNet, PoolingFormer, BANG, etc.

Benchmarks:

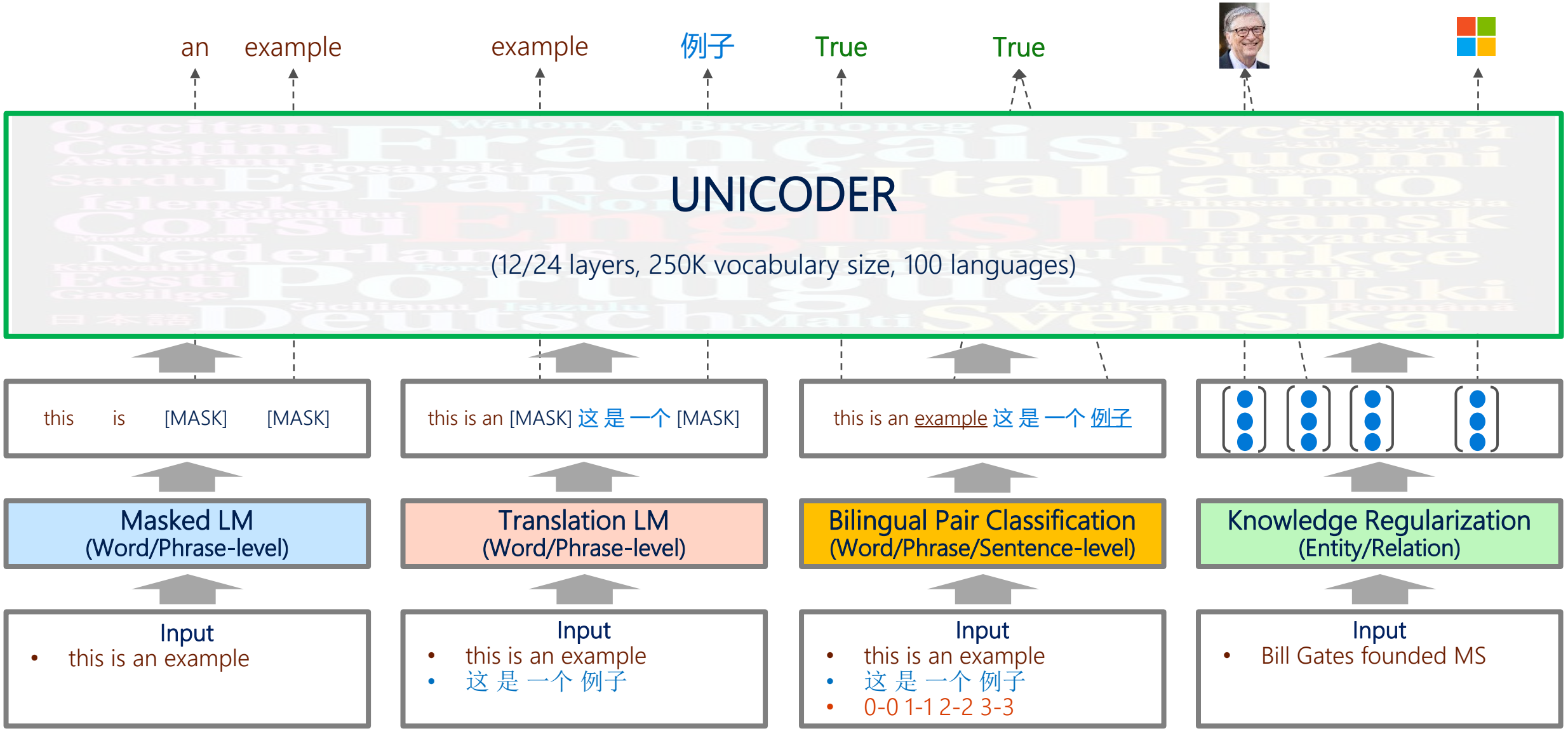
- XGLUE, CodeXGLUE, GEM, GLGE, etc.

2). Pre-trained Model & Benchmark for Multiple Languages

Why is Multilingual Pre-training Important

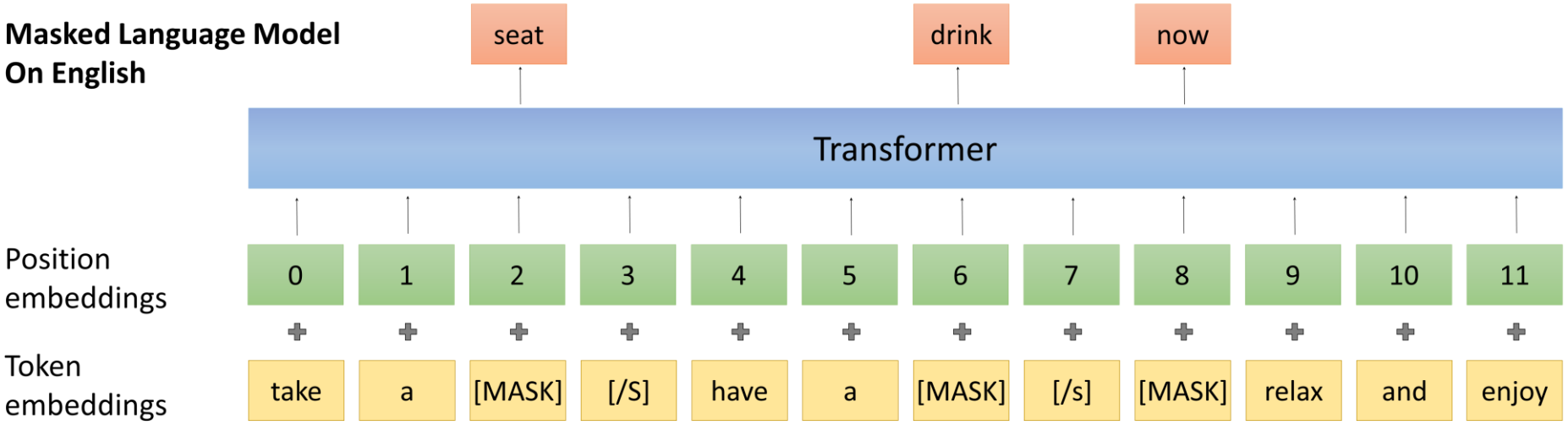


UNICODER: a UNIversal enCODER for multiple languages

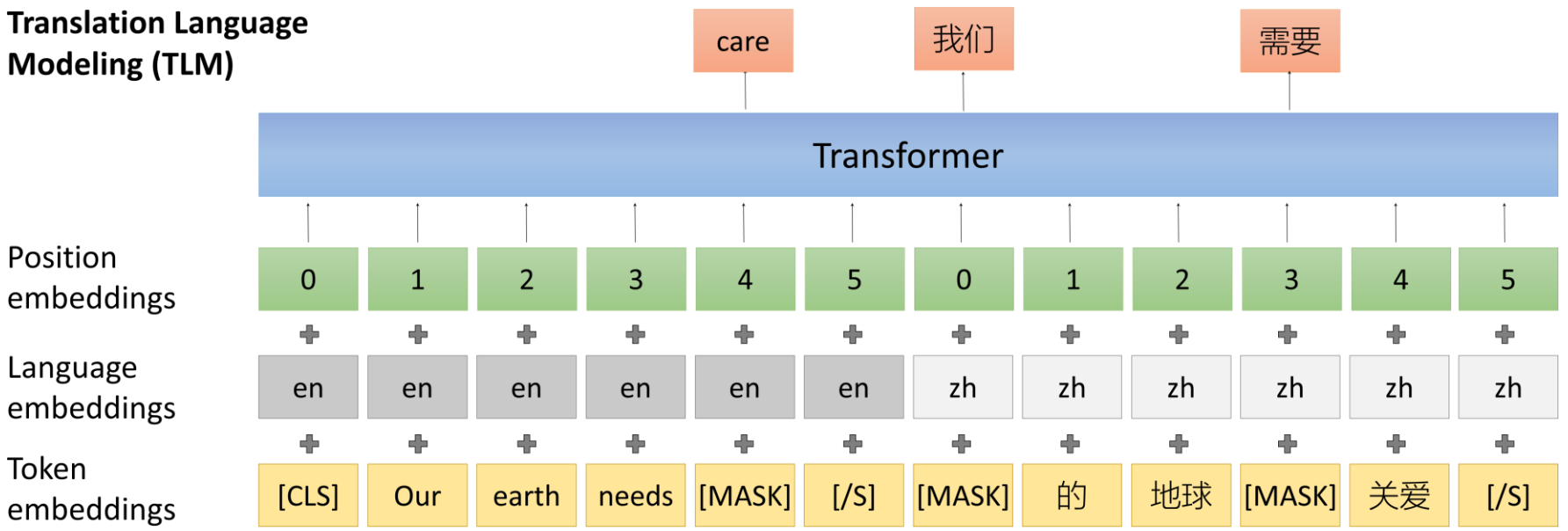


Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. **Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks**. EMNLP, 2019.

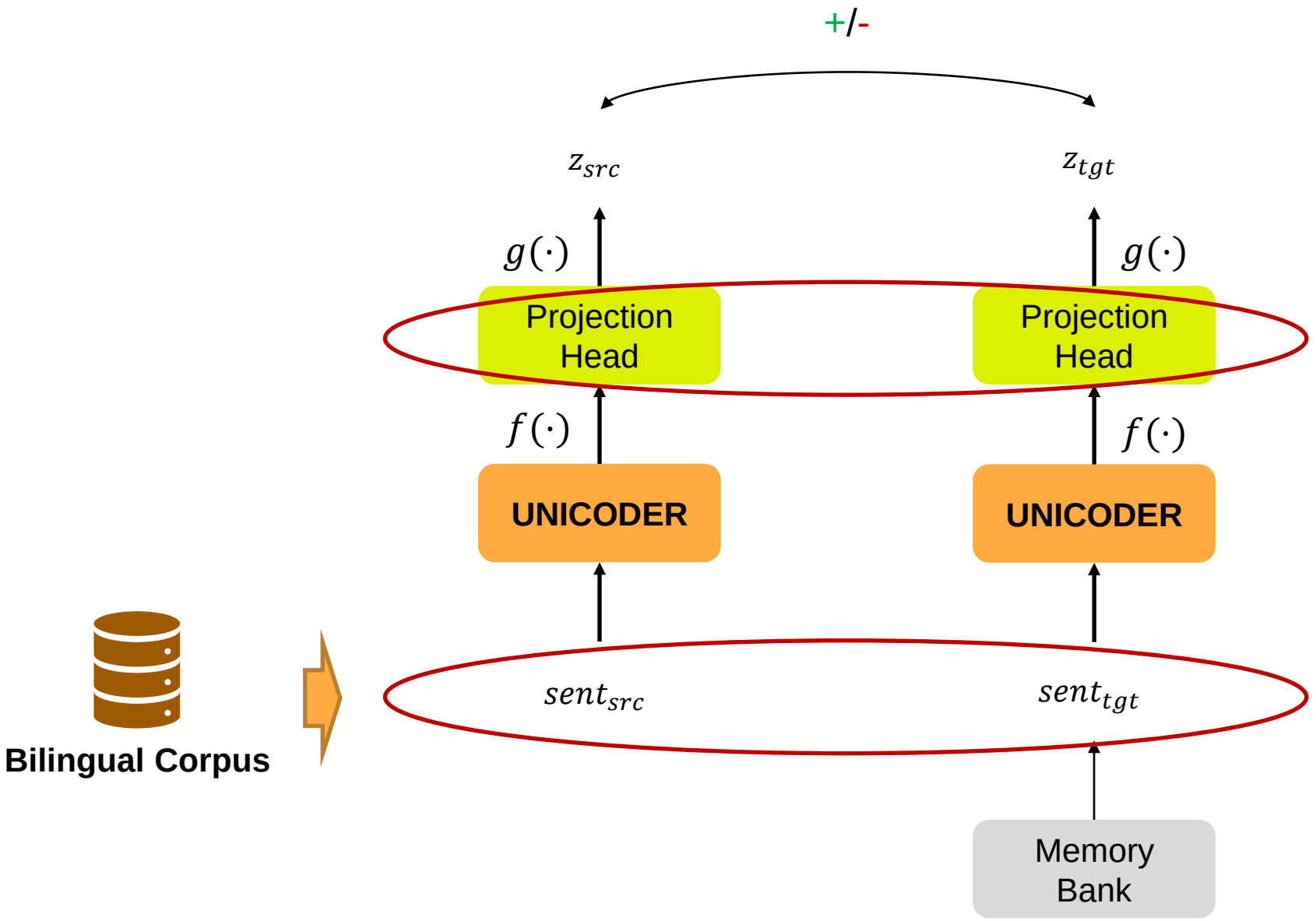
Pre-training Task (1): Masked Language Model



Pre-training Task (2): Translation Language Model

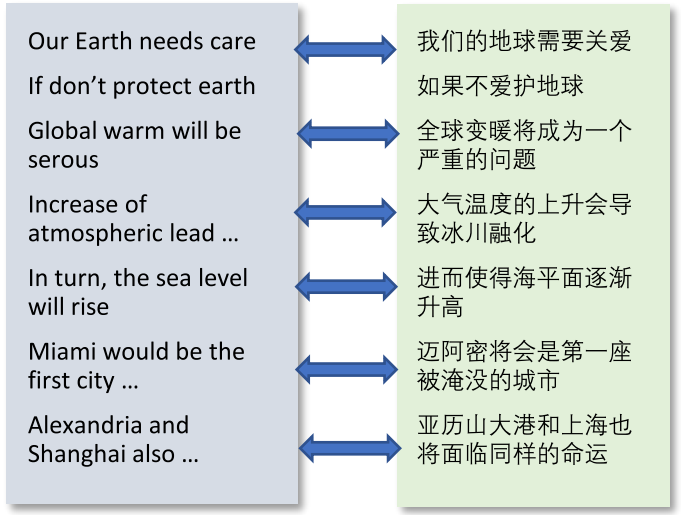


Pre-training Task (3): Bilingual Pair Classification

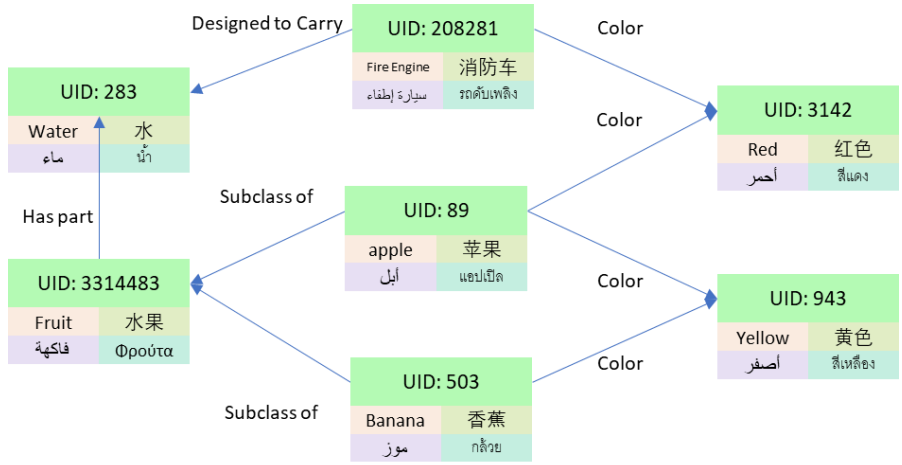


Enrich Multilingual Pre-training with Multilingual Knowledge Base

Monolingual Data

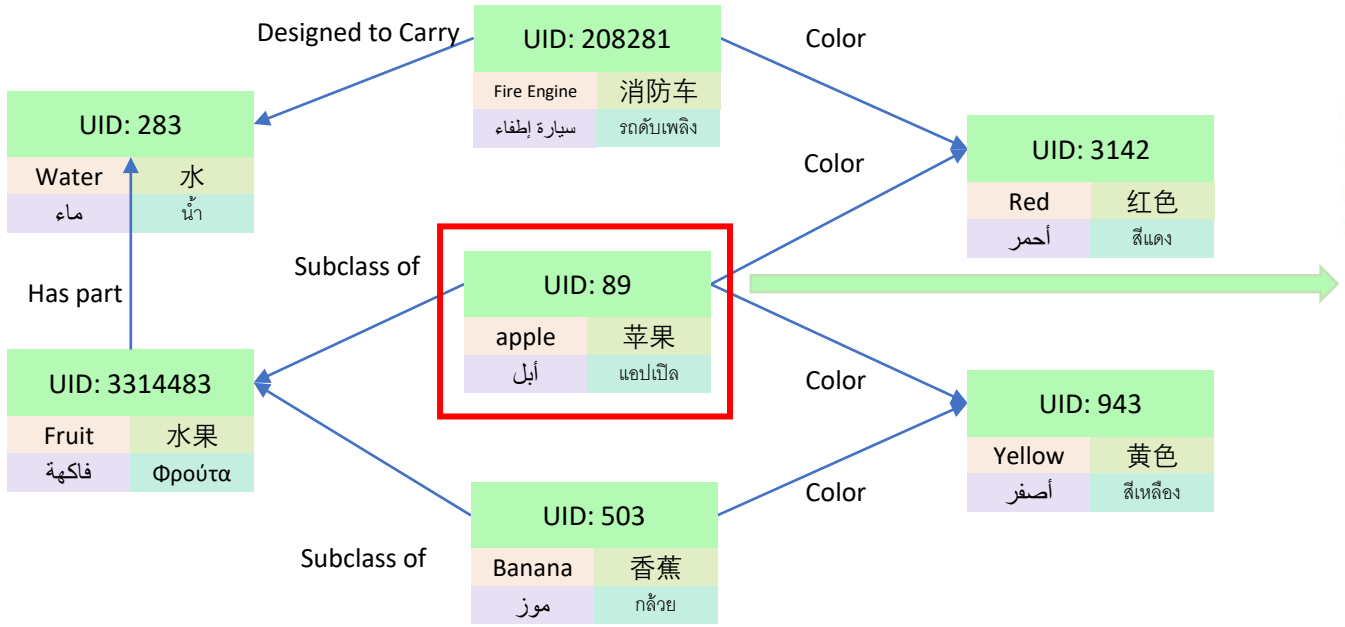


Bilingual Data



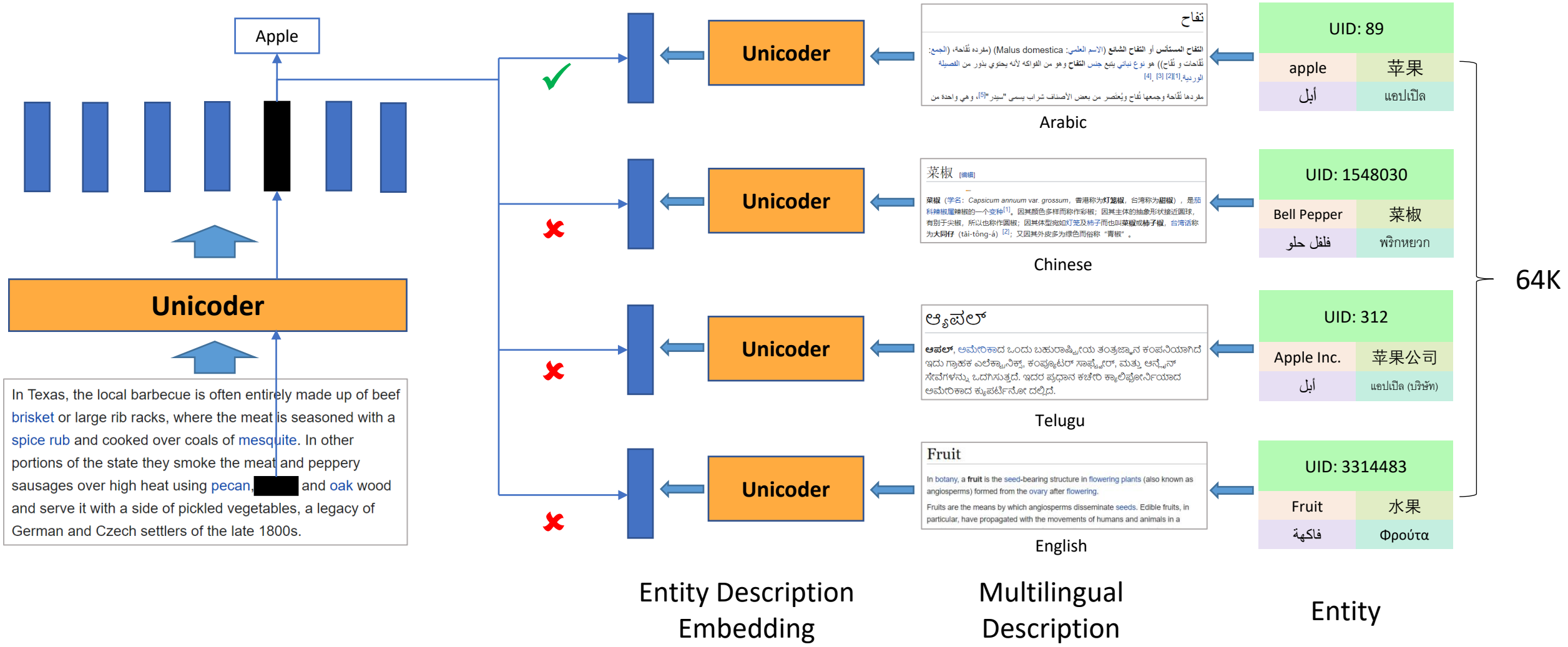
Multilingual Knowledge Base

Multilingual Knowledge Base

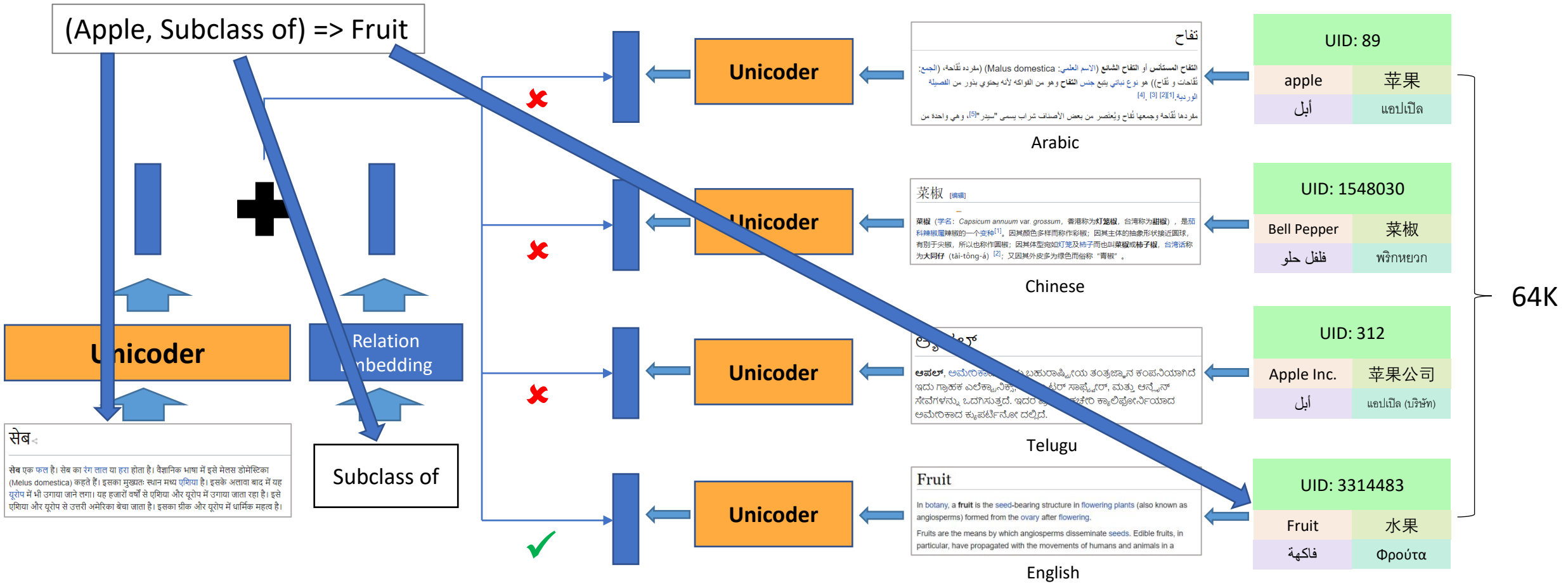


- Multilingual KB is a set of <Subject, Relation, Object> triplets
  - <Apple, Subclass of, Fruit>
  - <Apple, Color, Red>
  - …
- Each entity has names and descriptions in different languages

Pre-training Task (4): Knowledge Regularization (entity-based)



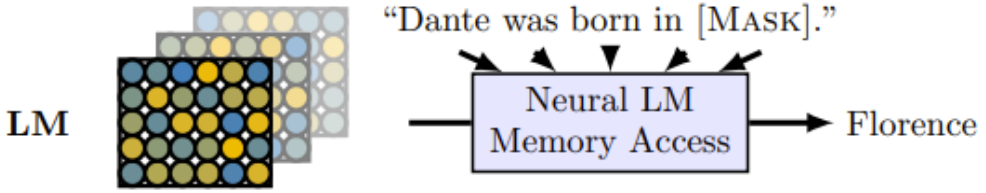
Pre-training Task (4): Knowledge Regularization (triple-based)



Evaluation: XNLI/NER/MLQA

Model	MLQA	NER	XNLI
XLM-R	65.1	79.0	74.5
XLM-R + entity-based pre-training	66.4	79.6	74.6
XLM-R + triple-based pre-training	66.7	**79.9**	74.6
XLM-R + both	**67.1** (**+2.0**)	79.8 (**+0.8**)	**74.8** (**+0.3**)

Evaluation: Knowledge Probing



Probe the factual and commonsense knowledge in language models.

Dataset	Data Source	Relation Number	Original Sample	Cloze Sample
Google-RE	Wikipedia	3	(Einstein, born in, German)	Einstein was born in _.
T-REx	Wikipedia	22	(iPod Touch, produced by, Apple)	iPod Touch is produced by _.
SQuAD	Wikipedia	-	Who developed the theory of relativity? Passage: Einstein is known for developing the theory of relativity…	The theory of relativity was developed by _.

Dataset	XLM-R	XLM-R + entity-based pre-training	XLM-R + triple-based pre-training	XLM-R + Both
Google-RE	7.4	8.0	9.9	**11.2**
T-REx	21.7	27.9	23.4	**29.7**
SQuAD	5.5	10.1	9.7	**11.5**

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

Evaluation: XTREME Benchmark for Multilingual NLU

Leaderboard results

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	Unicoder + ZCode	MSRA + Cognition	Microsoft	Apr 26, 2021	81.6	88.4	76.2	72.5	93.7
2	VECO	DAMO NLP Team	Alibaba	Mar 22, 2021	81.4	88.9	75.6	72.9	92.7
3	Polyglot	MLNLC	ByteDance	Feb 25, 2021	81.3	88.1	80.6	71.8	89.4
4	ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
5	HiCTL	DAMO MT Team	Alibaba	Mar 21, 2021	80.8	89.0	74.4	71.9	92.6
6	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
7	Anonymous3	Anonymous3	Anonymous3	Jan 3, 2021	79.9	88.2	74.6	71.7	89.0
8	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
9	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
	XLM-R	XTREME	Alphabet,						

Participate in Competition

Multilingual NLP Benchmarks

XNLI

The Cross-Lingual NLI Corpus (XNLI)

Alexis Conneau
Guillaume Lample
Ruty Rinott
Holger Schwenk
Ves Stoyanov
Facebook AI

Adina Williams
Sam Bowman
NYU

Introduction

The Cross-lingual Natural Language Inference (XNLI) corpus is a crowd-sourced collection of 5,000 test and 2,500 dev pairs for the MultiNLI corpus. The pairs are annotated with textual entailment and translated into 14 languages: *French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu*. This results in 112.5k annotated pairs. Each premise can be associated with the corresponding hypothesis in the 15 languages, summing up to more than 1.5M combinations. The corpus is made to evaluate how to perform inference in any language (including low-resources ones like Swahili or Urdu) when only English NLI data is available at training time. One solution is cross-lingual sentence encoding, for which XNLI is an evaluation benchmark.

Examples

Language	Premise	Label	Hypothesis
Face-to-face conversation			
English	There's so much you could talk about on that I'll just skip that.	*contradictory*	I want to tell you everything I know about

Google AI Blog

The latest news from Google AI

XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization

Monday, April 13, 2020

Posted by Melvin Johnson, Senior Software Engineer, Google Research and Sebastian Ruder, Research Scientist, DeepMind

One of the key challenges in natural language processing (NLP) is building systems that not only work in English but in *all* of the world's ~6,900 languages. Luckily, while most of the world's languages are *data sparse* and do not have enough data available to train robust models on their own, many languages do share a considerable amount of underlying structure. On the vocabulary level, languages often have words that stem from the same origin — for instance, "*desk*" in English and "*Tisch*" in German both come from the Latin "*discus*". Similarly, many languages also mark semantic roles in similar ways, such as the use of postpositions to mark temporal and spatial relations in both Chinese and Turkish.

In NLP, there are a number of methods that leverage the shared structure of multiple languages in training in order to overcome the data sparsity problem. Historically, most of these methods focused on performing a specific task in multiple languages. Over the last few years, driven by advances in deep learning, there has been an increase in the number of approaches that attempt to learn *general-purpose multilingual representations* (e.g., mBERT, XLM, XLM-R), which aim to capture knowledge that is shared across languages and that is useful for many tasks. In practice, however, the evaluation of such methods has mostly focused on a small set of tasks and for linguistically similar languages.

To encourage more research on multilingual learning, we introduce "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization", which covers 40 typologically diverse languages (spanning 12 language families) and includes nine tasks that collectively require reasoning about different levels of syntax or semantics. The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. Among these are many under-studied languages, such as the Dravidian languages Tamil (spoken in southern India, Sri Lanka, and Singapore), Telugu and Malayalam (spoken mainly in southern India), and the Niger-Congo languages Swahili and Yoruba, spoken in Africa. The code and data, including examples for running various baselines, is available here.

XTREME Tasks and Languages
The tasks included in XTREME cover a range of paradigms, including sentence classification,

XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation

Yaobo Liang, ✉Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, Ming Zhou
{yalia,nanduan,yegong,t-niwu,v-fengu,v-weqi,migon,lisho,djiang,gucao,xiafan,bzhang,rahul.agrawal, edwac,sinwei,tbharti,jiuche,winniew,shuguanl,fanyang, ranganm, mingzhou}@microsoft.com

Abstract

In this paper, we introduce **XGLUE**, a new benchmark dataset to train large-scale cross-lingual pre-trained models using multilingual and bilingual corpora, and evaluate their performance across a diverse set of cross-lingual tasks. Comparing to GLUE (Wang et al., 2019), which is labeled in English and includes natural language understanding tasks only, XGLUE has three main advantages: (1) it provides two corpora with different sizes for cross-lingual pre-training; (2) it provides 11 diversified tasks that cover both natural language understanding and generation scenarios; (3) for each task, it provides labeled data in multiple languages. We extend a recent cross-lingual pre-trained model Unicoder (Huang et al., 2019) to cover both understanding and generation tasks, which is evaluated on XGLUE as a strong baseline. We also evaluate the base versions (12-layer) of Multilingual BERT, XLM and XLM-R for comparison.

1 Introduction

Pre-training + Fine-tuning has become a new NLP paradigm, where the general knowledge are firstly learnt from large-scale corpus by self-supervised learning and then transferred to downstream tasks by task-specific fine-tuning. Three different types of pre-trained models are explored recently, including *monolingual pre-trained models* (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019b; Dong et al., 2019; Lewis et al., 2019a), *multilingual and cross-lingual pre-trained models* (Devlin et al., 2019; Conneau and Lample, 2019; Huang et al., 2019; Conneau et al., 2019) and *multimodal pre-trained models* (Lu et al., 2019; Li et al., 2020; Chen et al., 2019; Zhou et al., 2020). In this paper, we focus on the cross-lingual pre-trained models, due to their importance to alleviating the low-resource issue among languages, where an NLP task often has rich training data in one language (such as English) but has few or no training data in other languages (such as French and German). In order to further advance the development of cross-lingual pre-trained models for various downstream tasks in different languages, this paper introduces **XGLUE**, a new benchmark dataset that can be used to: (i) train large-scale cross-lingual pre-trained models using multilingual and bilingual corpora, (ii) evaluate generalization capabilities of the cross-lingual pre-trained models across a diverse set of cross-lingual tasks.

The contribution of XGLUE is two-fold. First, it provides 11 diversified cross-lingual tasks covering both understanding and generation scenarios, which, to the best of our knowledge, is the first attempt in the cross-lingual dataset construction efforts. XTREME (Hu et al., 2020) is a concurrent work of XGLUE. But it includes cross-lingual understanding tasks only. Second, an extended version of Unicoder (Huang et al., 2019) is described and evaluated as a strong cross-lingual pre-trained model baseline on XGLUE for both understanding and generation tasks. We also evaluate the base versions (12-layer) of Multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2019) for comparison. We conduct comprehensive experiments on XGLUE, which not only show interesting findings, but also point out several ways to further improve the cross-lingual pre-trained models.

2 XGLUE Benchmark[1]

2.1 Pre-training Corpus

We collect two corpora, *Small Corpus* and *Large Corpus*, with different sizes for cross-lingual pre-training: the former can be used to evaluate new ideas effectively and the latter can be used to train large-scale models. Table 1 lists the data statistics.

[1]https://to-be-released.

XNLI (Facebook, 2018) **XTREME (Google, 2020)** **XGLUE (Microsoft, 2020)**

XGLUE Benchmark 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

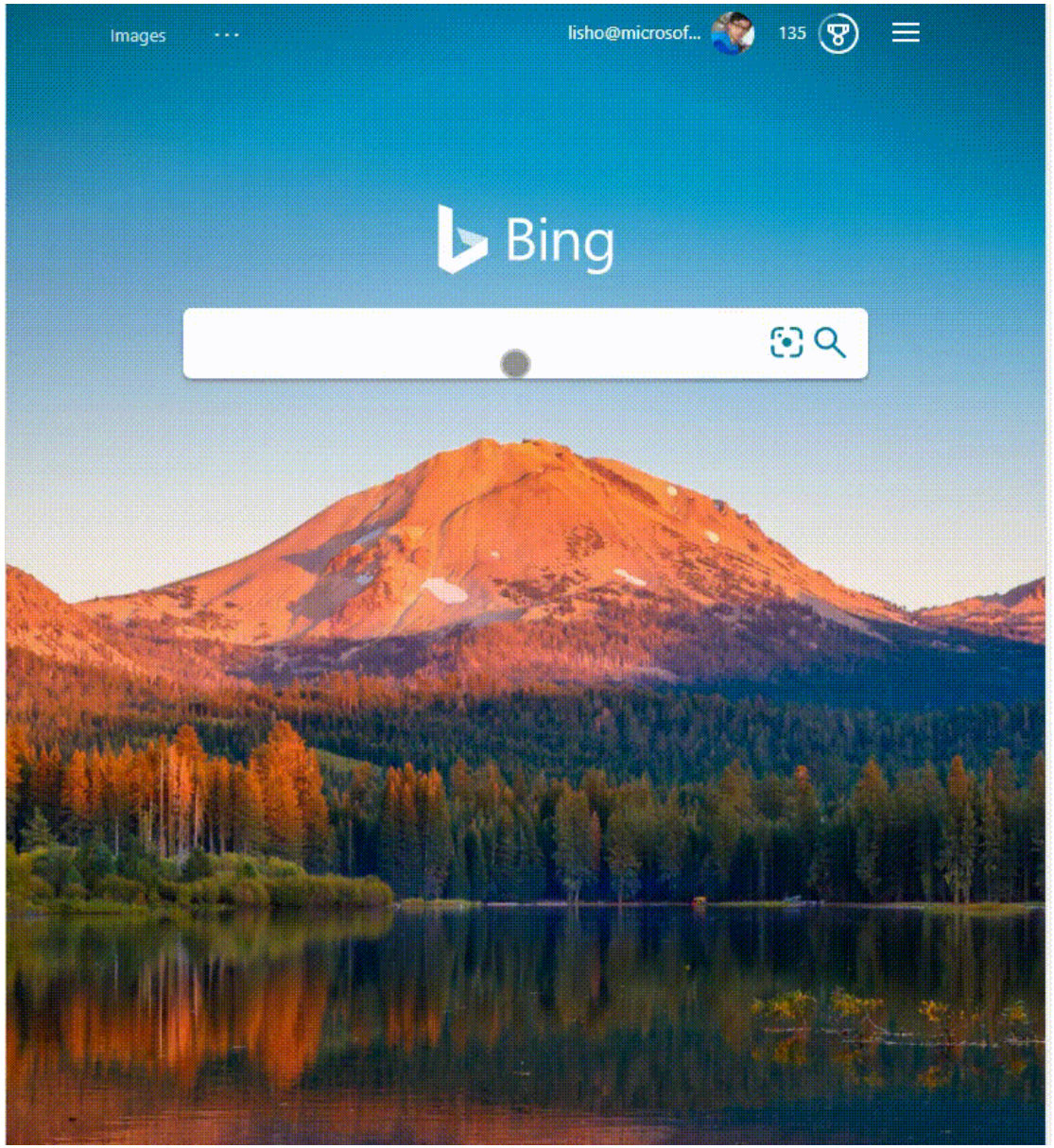
OCTOBER 1 2020

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):



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.

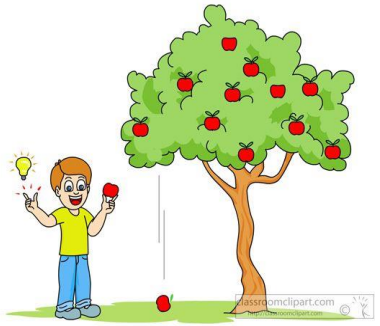
MSRA released XGLUE (https://microsoft.github.io/XGLUE/) as a new benchmark for multilingual NLP.

3). Pre-trained Model & Benchmark for Language + Vision

Our Intuition

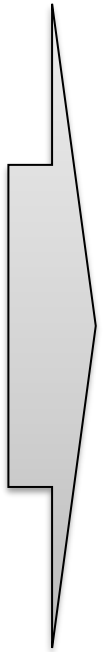
reading

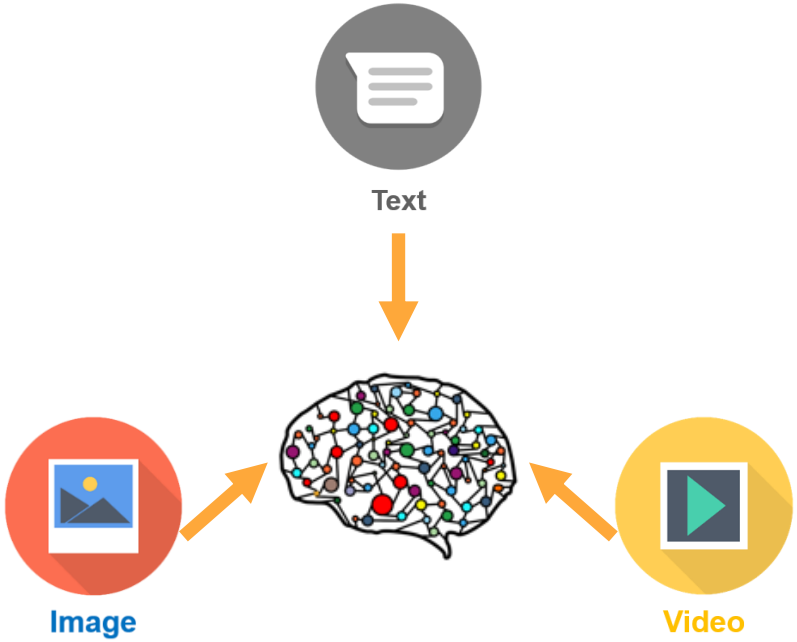
seeing



...

...





Humans learn from multiple senses.

Models learn from multimodal data.

Our Roadmap

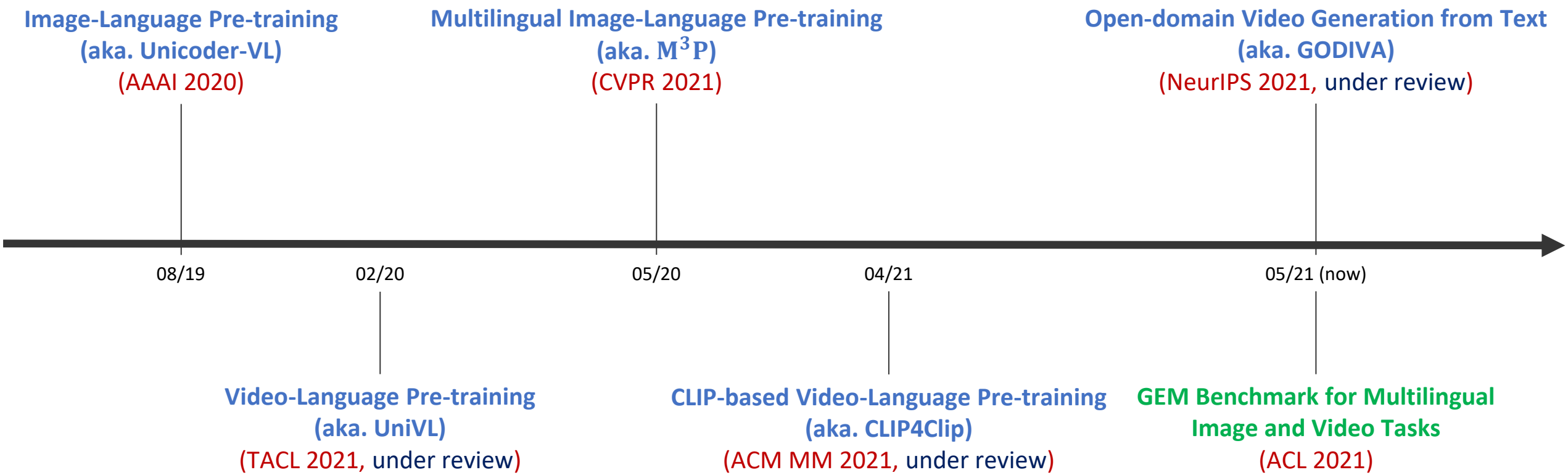
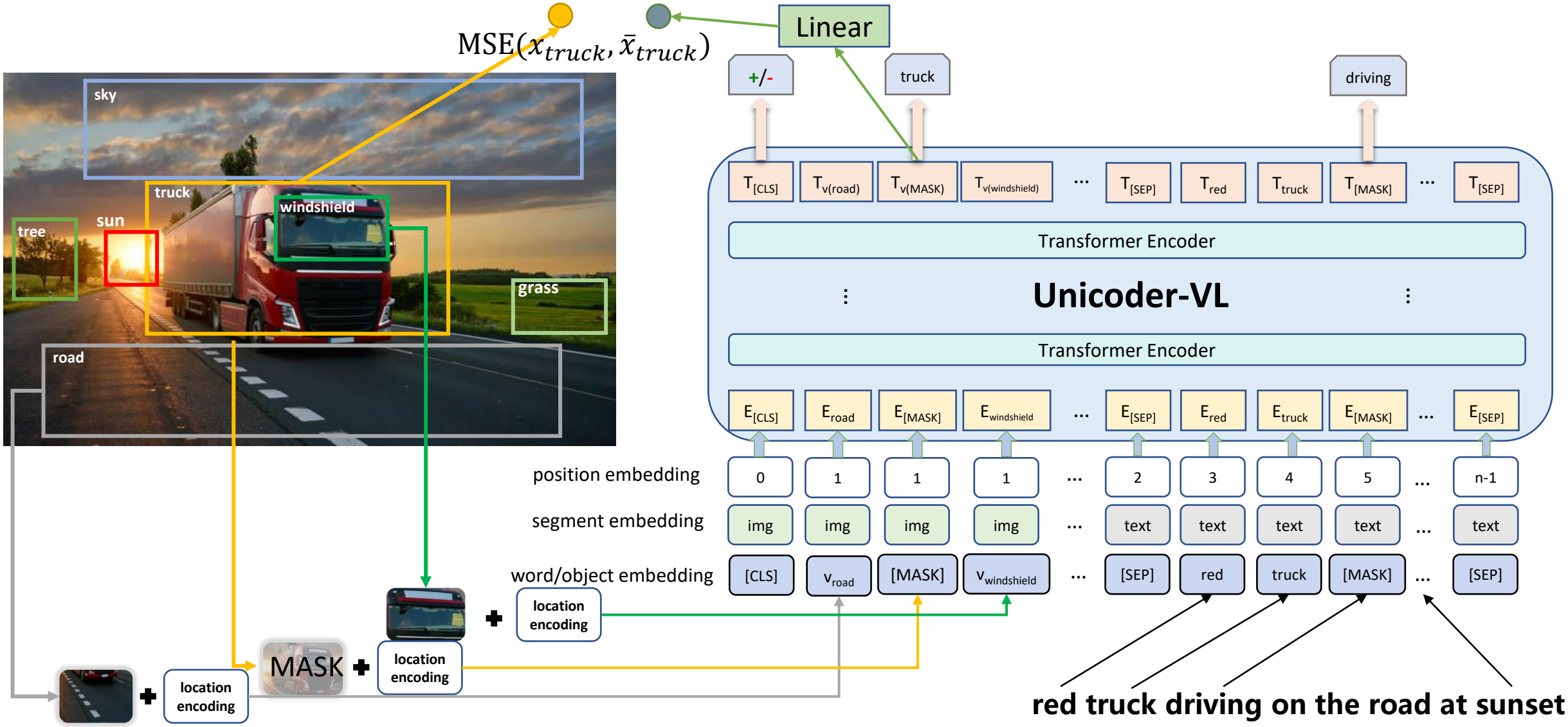


Image-Language Pre-training
(aka. Unicoder-VL)
(AAAI 2020)
Multilingual Image-Language Pre-training
(aka. M³P)
(CVPR 2021)
Open-domain Video Generation from Text
(aka. GODIVA)
(NeurIPS 2021, under review)
08/19
02/20
05/20
04/21
05/21 (now)
Video-Language Pre-training
(aka. UniVL)
(TACL 2021, under review)
CLIP-based Video-Language Pre-training
(aka. CLIP4Clip)
(ACM MM 2021, under review)
GEM Benchmark for Multilingual
Image and Video Tasks
(ACL 2021)

Unicoder-VL for Image-Language Tasks



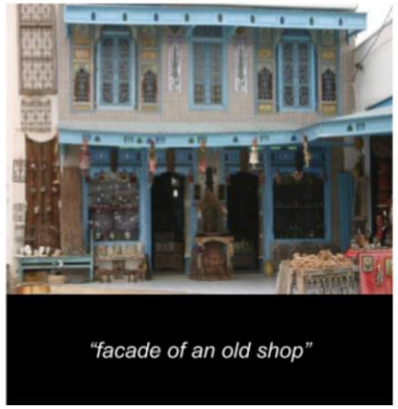
Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, Ming Zhou. **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training**. AAAI 2020.

Evaluation 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)

GQA Real-World Visual Reasoning Challenge

★ 68

Organized by: Stanford
Starts on: Feb 9, 2017 4:00:00 AM
Ends on: Mar 2, 2099 4:00:00 AM

Overview | Evaluation | Phases | Participate | Leaderboard | Discuss

Leaderboard

Phase: test2019, Split: Test

B - Baseline submission * - Private submission

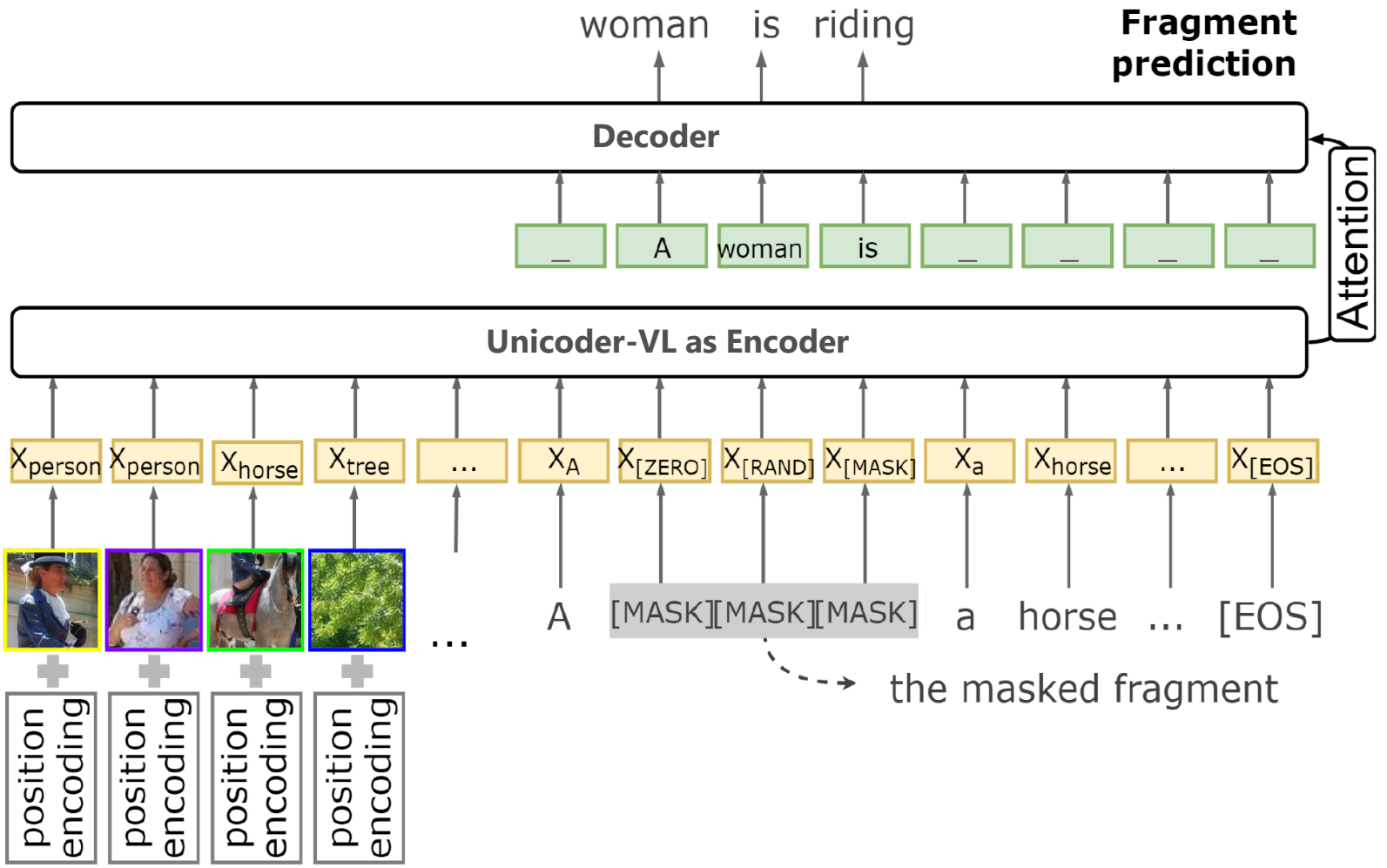
Rank	Participant team	Binary	Open	Consistency	Plausibility	Validity	Distribution	Accuracy	Last submission at
1	Human Performance (human)	91.20	87.40	98.40	97.20	98.90	0.00	89.30	2 years ago
2	DREAM+Unicoder-VL (MSRA)	84.46	68.60	91.47	83.75	96.42	3.68	76.04	2 years ago
3	TRRNet (Ensemble)	82.12	66.89	89.00	83.58	96.76	1.29	74.03	1 year ago
4	MIL-nbgao	80.80	67.64	91.76	83.90	96.73	1.70	73.81	6 months ago
5	Kakao Brain	79.68	67.73	77.02	83.70	96.36	2.46	73.33	2 years ago
6	AIOZ (Coarse-to-Fine Reasoning, Sing)	81.16	64.19	90.96	84.81	96.77	2.39	72.14	1 year ago
7	270	77.50	63.82	86.94	83.77	96.65	1.49	70.23	2 years ago
8	NSM ensemble (updated)	80.45	56.16	93.83	84.16	96.53	2.78	67.55	2 years ago
9	VinVL (Single Model)	82.63	48.77	94.35	84.98	96.62	4.72	64.65	4 months ago
10	TRRNet (Single)	77.91	50.22	89.84	85.15	96.47	5.25	63.20	1 year ago
11	NSM single (updated)	78.94	49.25	93.25	84.28	96.41	3.71	63.17	2 years ago
12	LXMERT (LXR955, Ensemble)	79.79	47.64	93.10	85.21	96.36	6.42	62.71	2 years ago

1/112

DREAM + Unicoder-VL (CVPR 2019)

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

Extend Unicoder-VL to Image Captioning

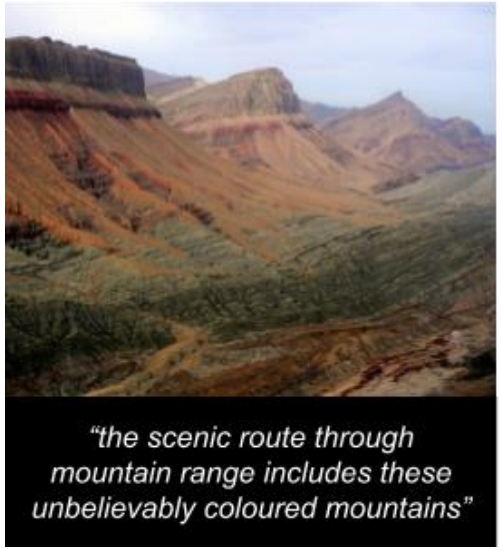


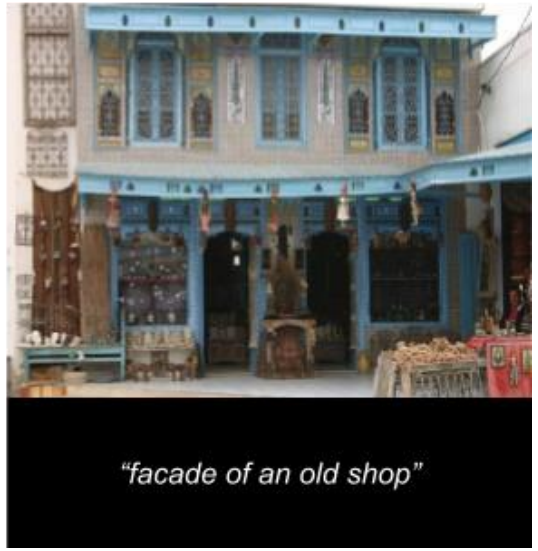
Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, Ming Zhou. **XGPT: Cross-modal Generative Pre-Training for Image Captioning**. arXiv 2020.

Evaluation Results: Image Captioning

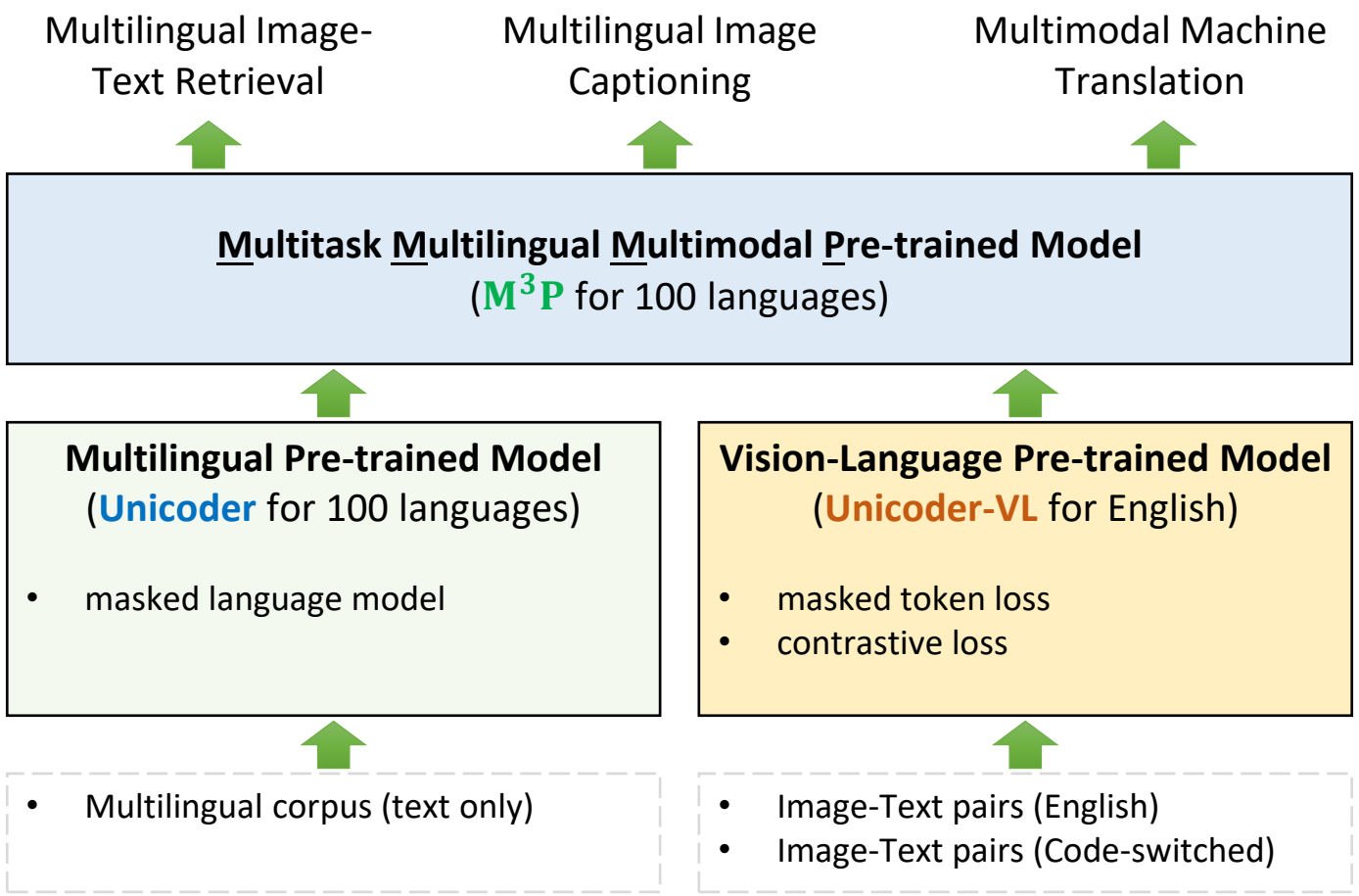
Pre-trained with Conceptual Captions dataset
~3.3M images annotated with captions
Evaluated on MSCOCO dataset

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)

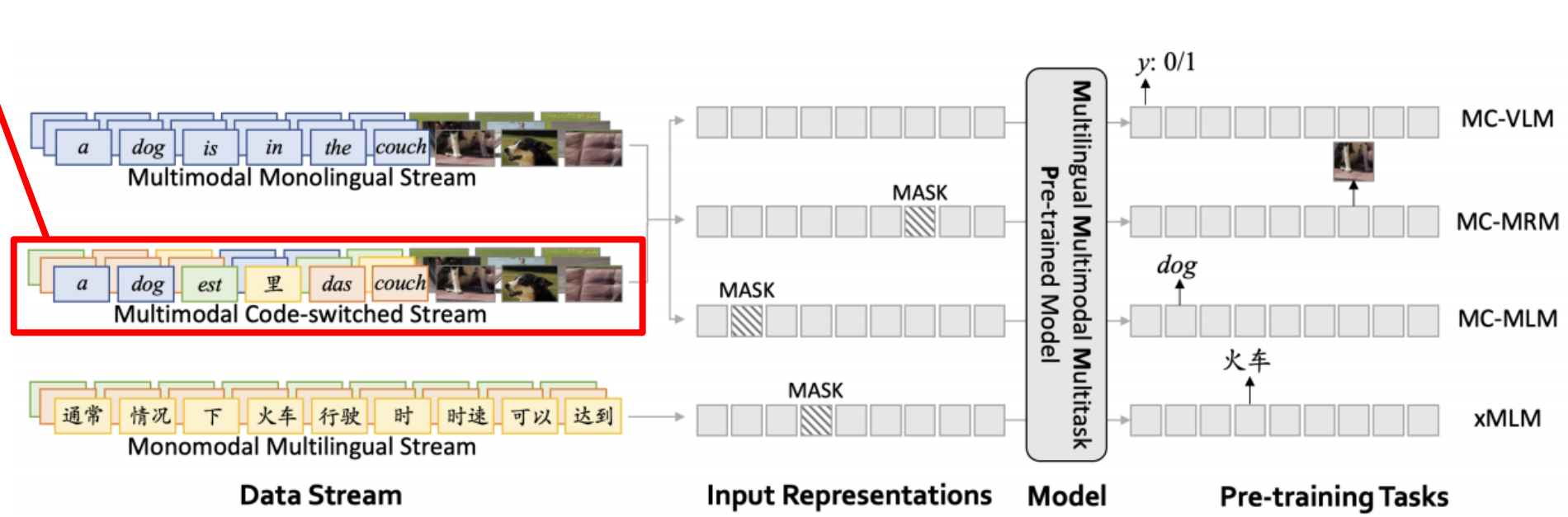


Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, Nan Duan. **M3P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training**. CVPR 2021.

M^{3}P with Code-switched Data Stream

Code-switched Data Steam Generation

- For each word in an English sentence, replace it with a translated word based on Panlex (https://panlex.org) with a probability of β.
- If a word has multiple translations, choose a random one.



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)**

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.

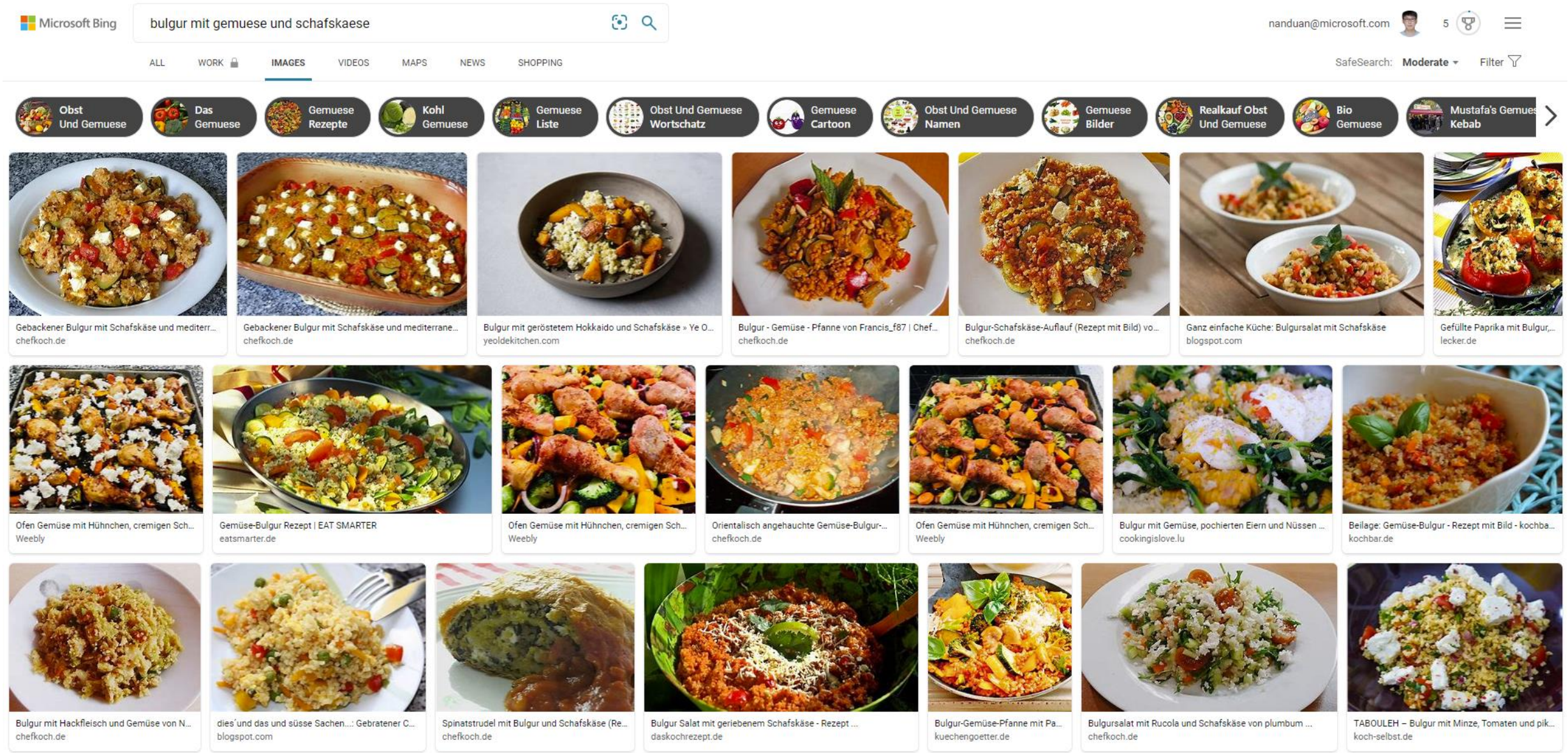
image caption output (**zh**): 一辆载着人和纸糊的房子的卡车行驶在街道上
(translation: a truck carrying people and paper houses travels down the street)

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.)

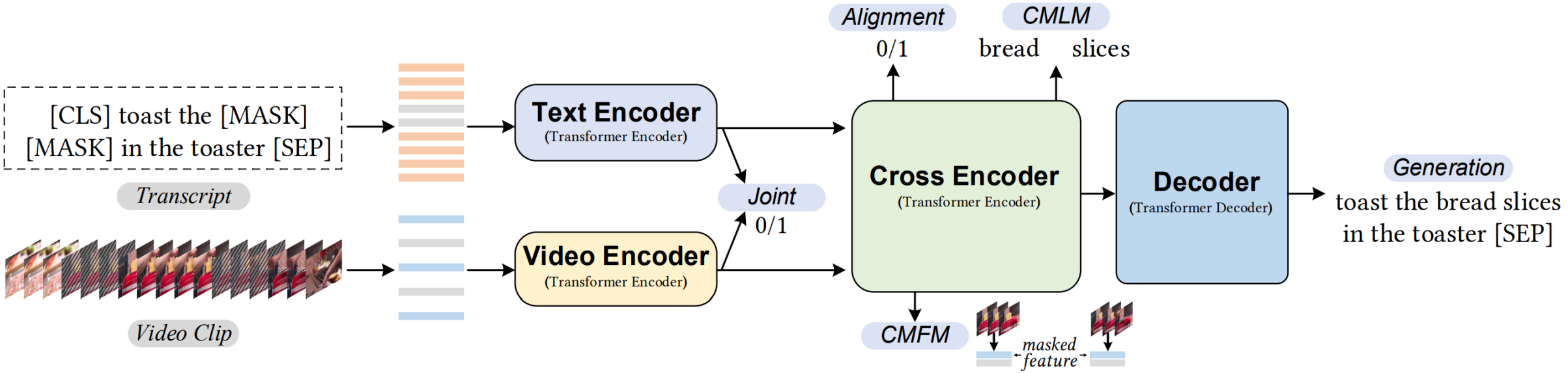
caption translation output (**de**): Ein Hund läuft auf grünem Rasen vor einem weißen Zaun.
(translation: A dog runs on green grass in front of a white fence.)

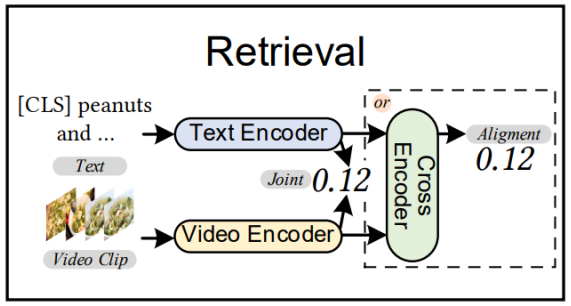
Unicoder-VL scaled Bing image search to 8 top-tier languages and 17 markets.

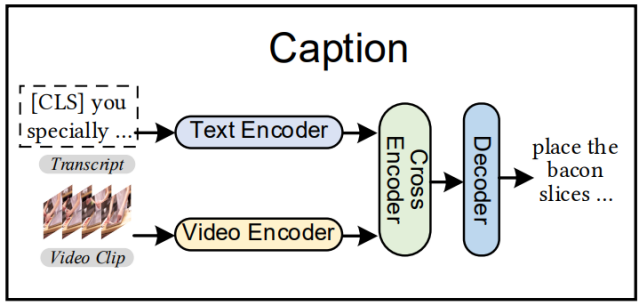


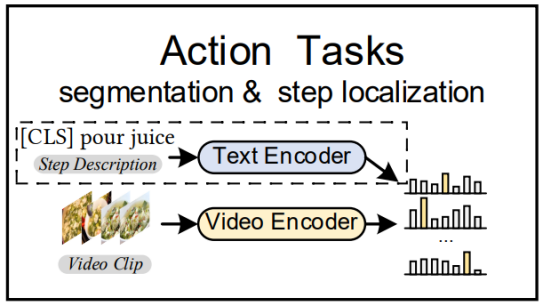
Unicoder-VL for Video-Language Tasks

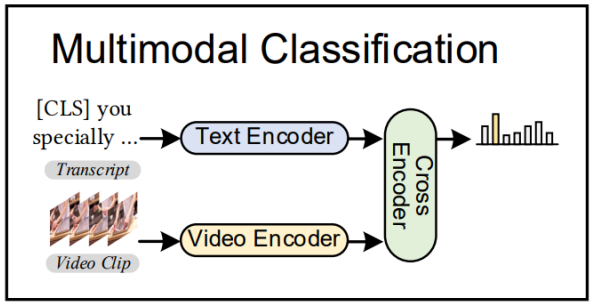
1. Video-Text Joint Embedding **2.** Video-Text Alignment **3.** Masked Frame Model **4.** Masked Language Model **5.** Caption Generation







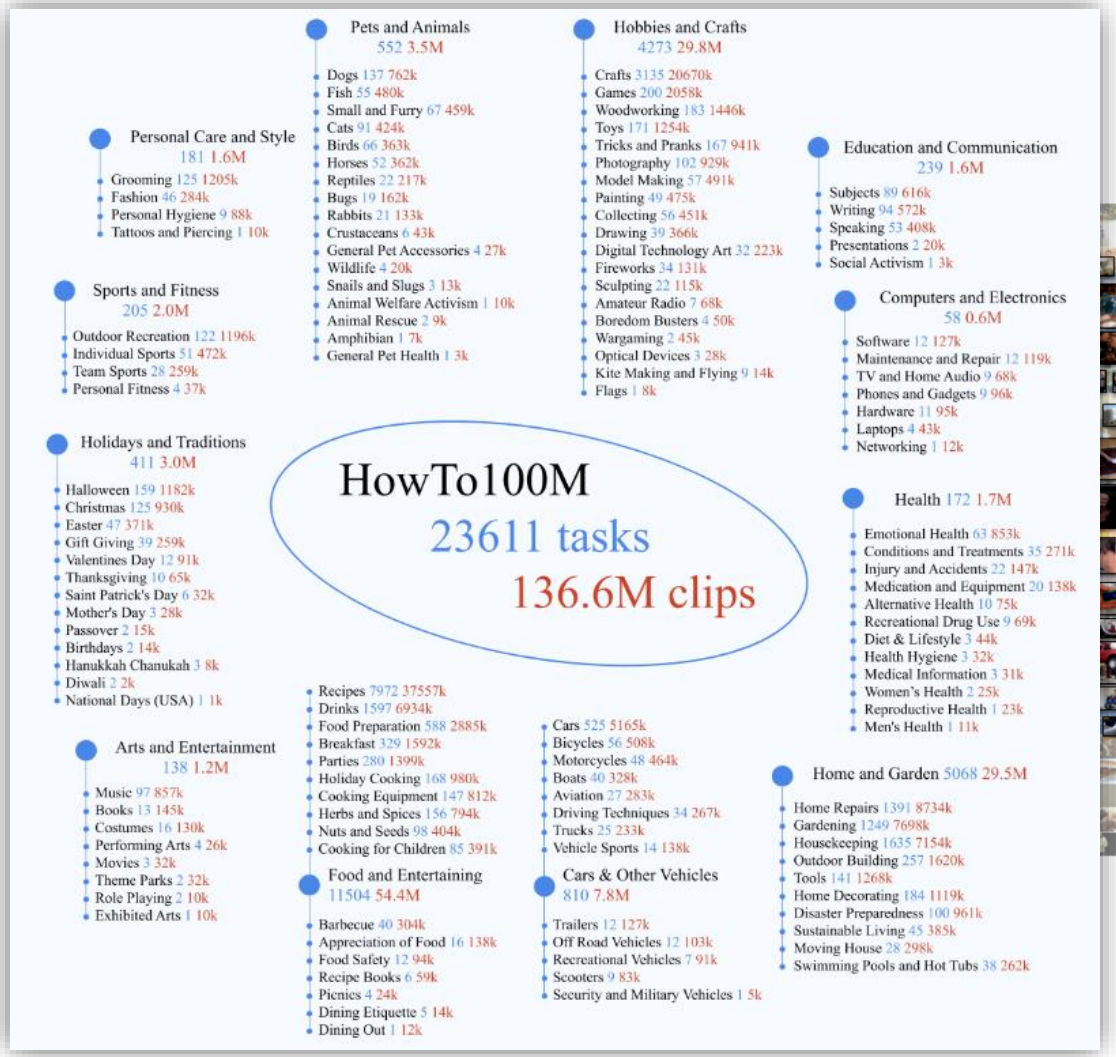




Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, Ming Zhou. **UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation**. arXiv 2020.

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**

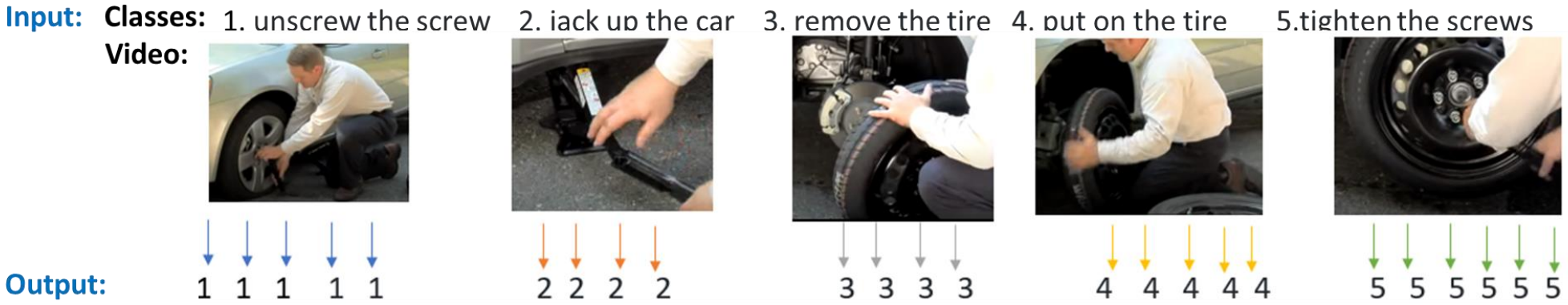
Table 3: The multimodal video captioning results on Youcook2 dataset. 'V' means video and 'T' means Transcript.

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

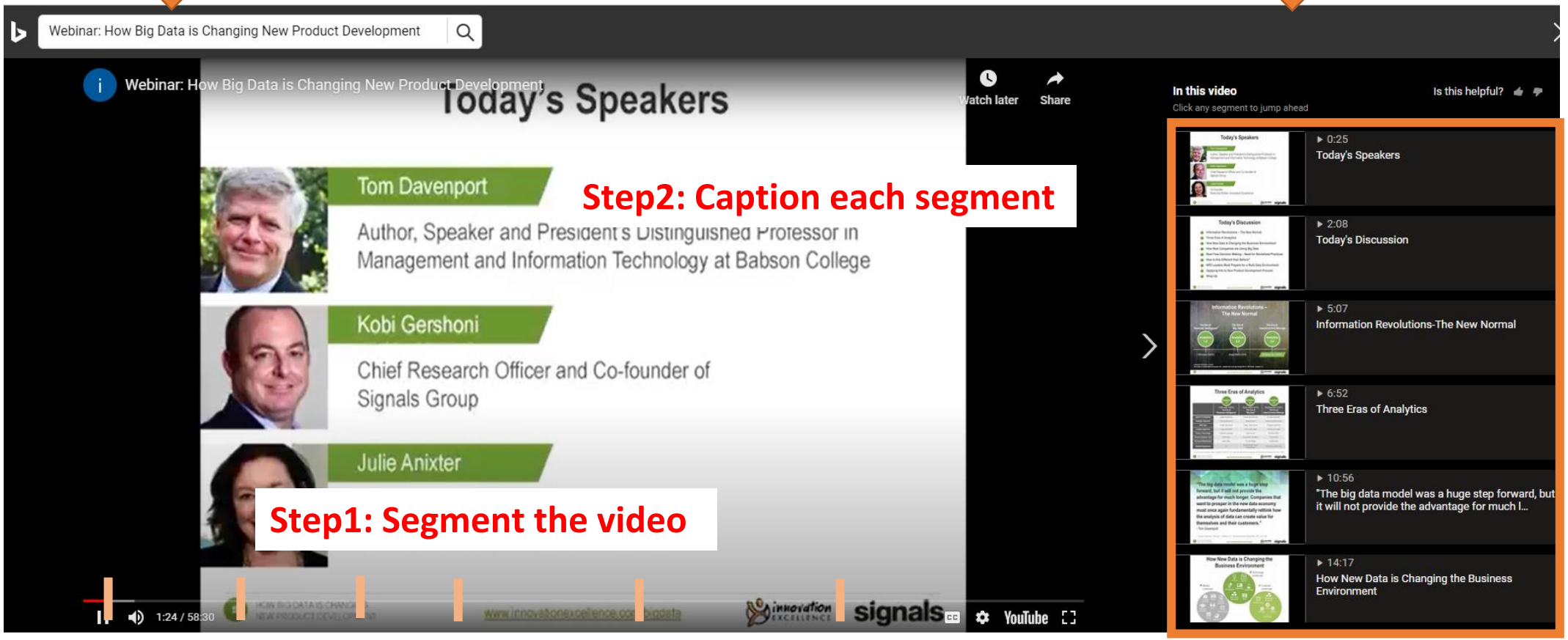
Methods	BA	F1	MAE	Corr
MV-LSTM (Rajagopalan et al., 2016)	73.9/-	74.0/-	1.019	0.601
TFN (Zadeh et al., 2017)	73.9/	73.4/-	1.040	0.633
MARN (Zadeh et al., 2018b)	77.1/	77.0/-	0.968	0.625
MFN (Zadeh et al., 2018a)	77.4/	77.3/-	0.965	0.632
RMFN (Liang et al., 2018)	78.4/	78.0/-	0.922	0.681
RAVEN (Wang et 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

Limitation of Unicoder-VL

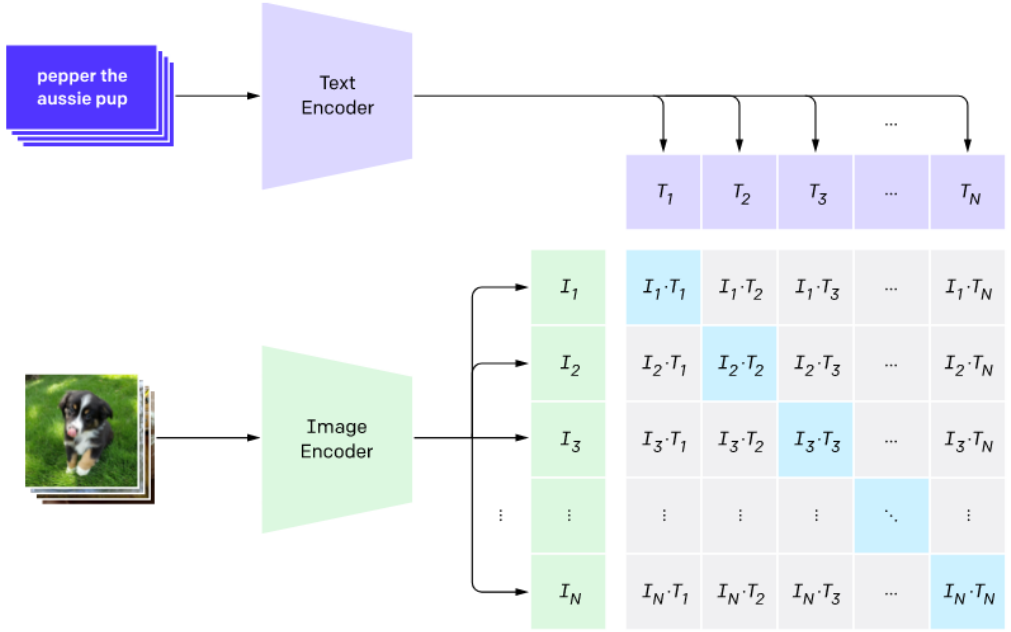
1) Fixed feature extraction VS. End-to-end learning.

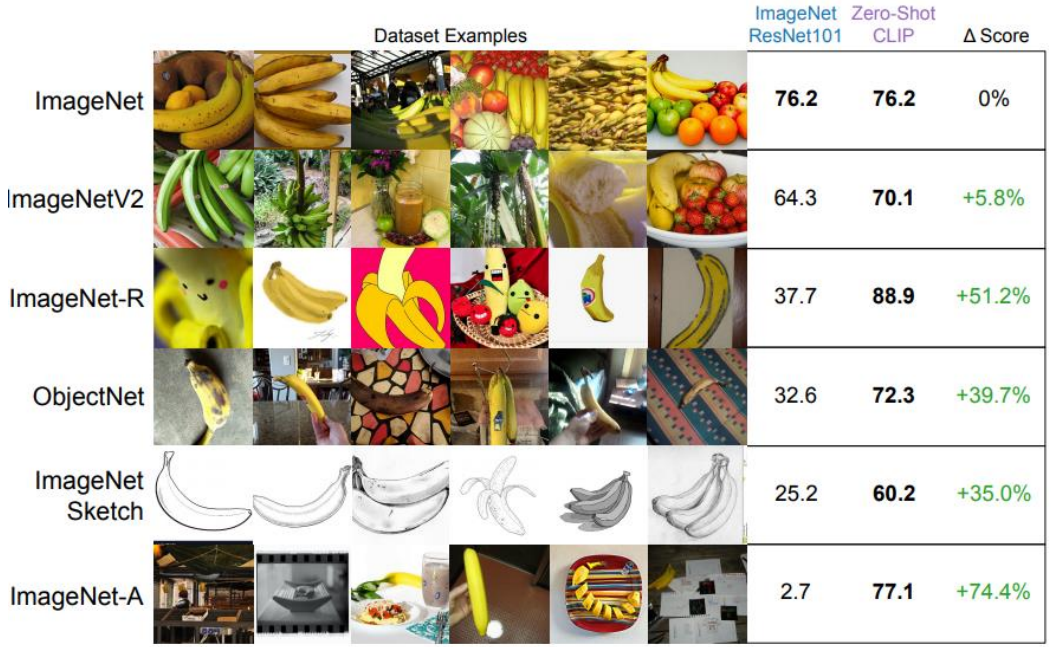
2) Visual backbone: Convolution VS. Transformer.

CLIP: Image Models From Language Supervision

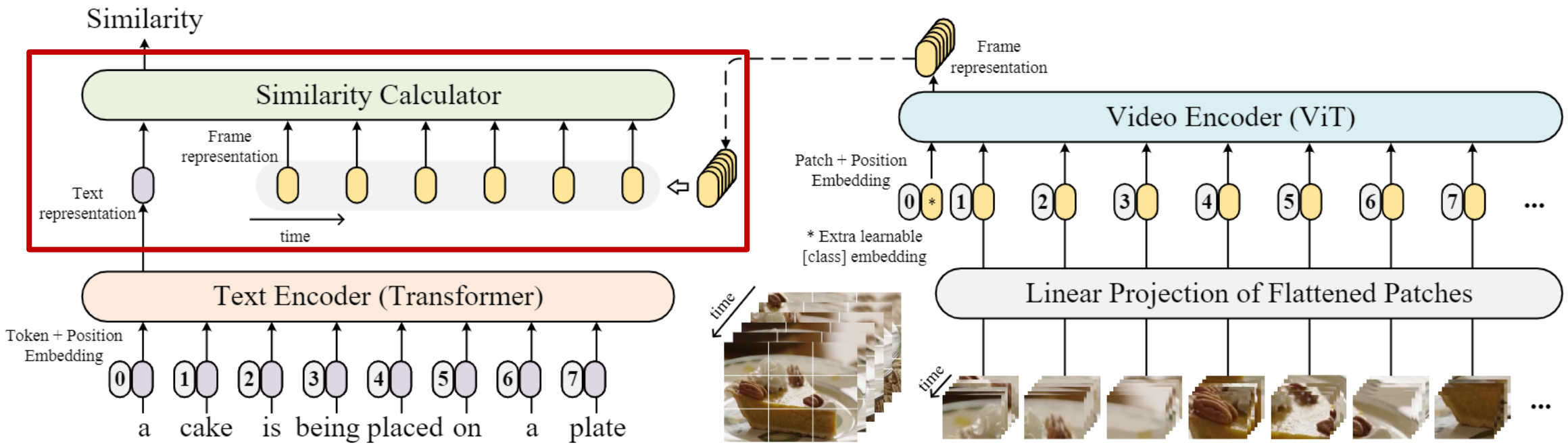
- An efficient and scalable way to learn image models from language.

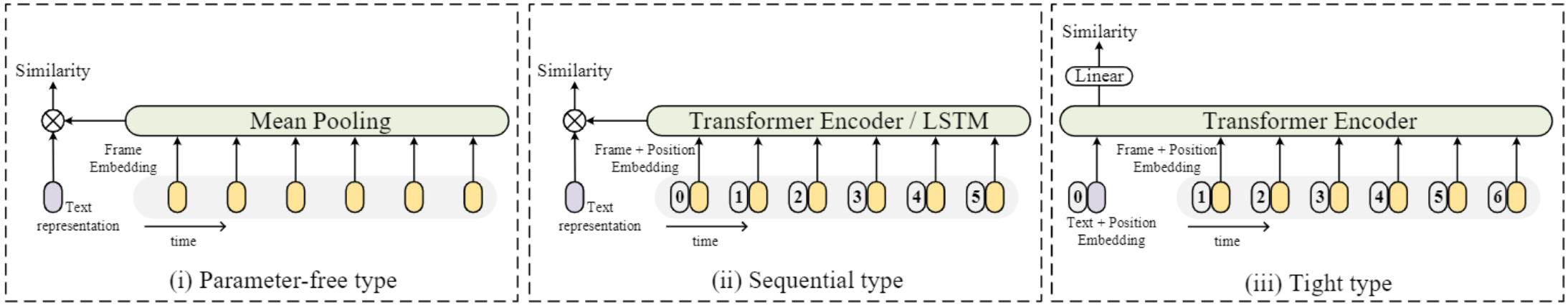
- Benchmarking on over 30 different datasets including tasks such as OCR, geo-localization, and fine-grained object classification.





Unicoder-VL with CLIP: End-to-End Video-Language Learning





Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, Tianrui Li. **CLIP4Clip: An Empirical Study of CLIP for End-to-End Video Clip Retrieval**. arXiv 2021.

Evaluation Results: Text-based Video Retrieval

Video Retrieval
(eval on MSR-VTT)

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
C+LSTM+SA[a]	M	✓	4.2	12.9	19.9	55	-
VSE[b]	M	✓	3.8	12.7	17.1	66	-
SNUVL[c]	M	✓	3.5	15.9	23.8	44	-
Kaufman et al.[d]	M	✓	4.7	16.6	24.1	41	-
CT-SAN[e]	M	✓	4.4	16.6	22.3	35	-
JSFusion[f]	M	✓	10.2	31.2	43.2	13	-
HowTo100M[g]	H+M	✓	14.9	40.2	52.8	9	-
ActBERT[h]	H+M		8.6	23.4	33.1	36	-
NoiseE[i]	H+M		17.4	41.6	53.6	8	-
UniVL[j]	H+M		21.2	49.6	63.1	6	-
HERO[k]	H+M		16.8	43.4	57.7	-	-
ClipBERT[l]	C+G+M	✓	22.0	46.8	59.9	6	-
(Ours)-meanP	W+M	✓	**42.1**	**71.9**	**81.4**	**2**	**15.7**
(Ours)-seqLSTM	W+M	✓	41.7	68.8	78.7	**2**	16.6
(Ours)-seqTransf	W+M	✓	42.0	68.6	78.7	**2**	16.2
(Ours)-tightTransf	W+M	✓	37.8	68.4	78.4	**2**	17.2

GEM: A General Evaluation Benchmark for Multimodal Tasks

Language	Train	Dev	Test	Total
English (en)	998,000	1,000	1,000	1,000,000
Spanish (es)	18,000	1,000	1,000	20,000
French (fr)	18,000	1,000	1,000	20,000
Italian (it)	18,000	1,000	1,000	20,000
Portuguese (pt)	18,000	1,000	1,000	20,000
German (de)	18,000	1,000	1,000	20,000
Korean (ko)	8,000	1,000	1,000	10,000
Polish (pl)	8,000	1,000	1,000	10,000
Catalan (ca)	2,000	1,000	1,000	4,000
Dutch (nl)	2,000	1,000	1,000	4,000
Japanese (ja)	2,000	1,000	1,000	4,000
Indonesian (id)	2,000	1,000	1,000	4,000
Vietnamese (vi)	2,000	1,000	1,000	4,000
Czech (cs)	2,000	1,000	1,000	4,000
Romanian (ro)	2,000	1,000	1,000	4,000
Turkish (tr)	0	0	1,000	1,000
Galician (gl)	0	0	1,000	1,000
Croatian (hr)	0	0	1,000	1,000
Hungarian (hu)	0	0	1,000	1,000
Malay (ms)	0	0	1,000	1,000
Total	**1,118,000**	**15,000**	**20,000**	**1,153,000**

Table 1: Language distribution and data statistics of GEM-I for multilingual image-language tasks.

Language	Train	Dev	Test	Total
German (de)	3,316	1,000	1,000	5,316
Portuguese (pt)	3,258	1,000	1,000	5,258
Dutch (nl)	2,961	1,000	1,000	4,961
Spanish (pt)	2,894	1,000	1,000	4,894
Russian (ru)	2,804	1,000	1,000	4,804
French (fr)	2,776	1,000	1,000	4,776
Italian (it)	2,589	1,000	1,000	4,589
Korean (ko)	2,452	1,000	1,000	4,452
English (en)	2,426	1,000	1,000	4,426
Japanese (ja)	2,000	1,000	1,000	4,000
Arabic (ar)	2,000	1,000	1,000	4,000
Polish (pl)	2,000	1,000	1,000	4,000
Chinese-Traditional (zh-t)	2,000	1,000	1,000	4,000
Farsi (fa)	2,000	1,000	1,000	4,000
Indonesian (id)	2,000	1,000	1,000	4,000
Turkish (tr)	2,000	1,000	1,000	4,000
Vietnamese (vi)	2,000	1,000	1,000	4,000
Hebrew (he)	1,807	1,000	1,000	3,807
Romanian (ro)	1,441	1,000	1,000	3,441
Swedish (sv)	1,419	1,000	1,000	3,419
Filipino (tl)	1,294	1,000	1,000	3,294
Malay (ms)	0	0	1,000	2,668
Norwegian (no)	0	0	1,000	1,098
Catalan (ca)	0	0	1,000	1,002
Croatian (hr)	0	0	907	907
Georgian (ka)	0	0	863	863
Chinese-Simplified (zh-s)	0	0	833	833
Hungarian (hu)	0	0	811	811
Albanian (sq)	0	0	809	809
Serbian-Latin (sr-l)	0	0	774	774
Total	**47,437**	**21,000**	**28,997**	**99,202**

Table 2: Language distribution and data statistics of GEM-V for multilingual video-language tasks.

Lin Su, Nan Duan, Edward Cui, Lei Ji, Chenfei Wu, Huaishao Luo, Yongfei Liu, Ming Zhong, Taroon Bharti, Arun Sacheti. **GEM: A General Evaluation Benchmark for Multimodal Tasks**. Findings of ACL, 2021.

GEM: A General Evaluation Benchmark for Multimodal Tasks

Figure 1: Examples in GEM-I dataset: Q: Query, T: Title.

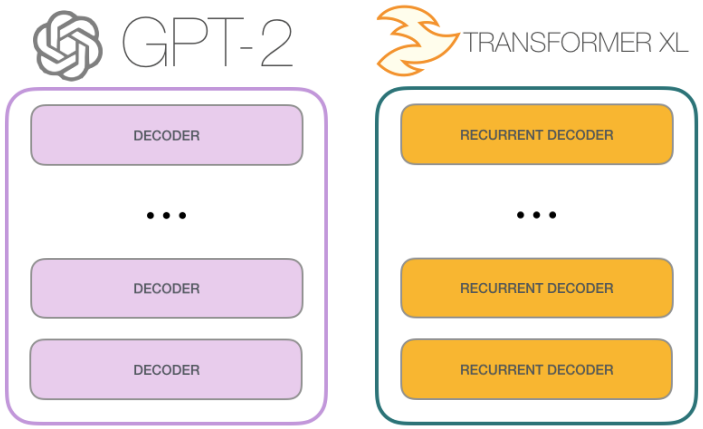
GEM: A General Evaluation Benchmark for Multimodal Tasks

Figure 2: Examples in GEM-V dataset: Q: Query, T: Title.

From Multimodal Understanding to Multimodal Generation:
GODIVA for Open Domain Video Generation from Text

From GPT to DALL-E, Large-scale Pretraining shows great potential for generating sequences.

Text-to-Text Generation



Text-to-Image Generation

DALL-E

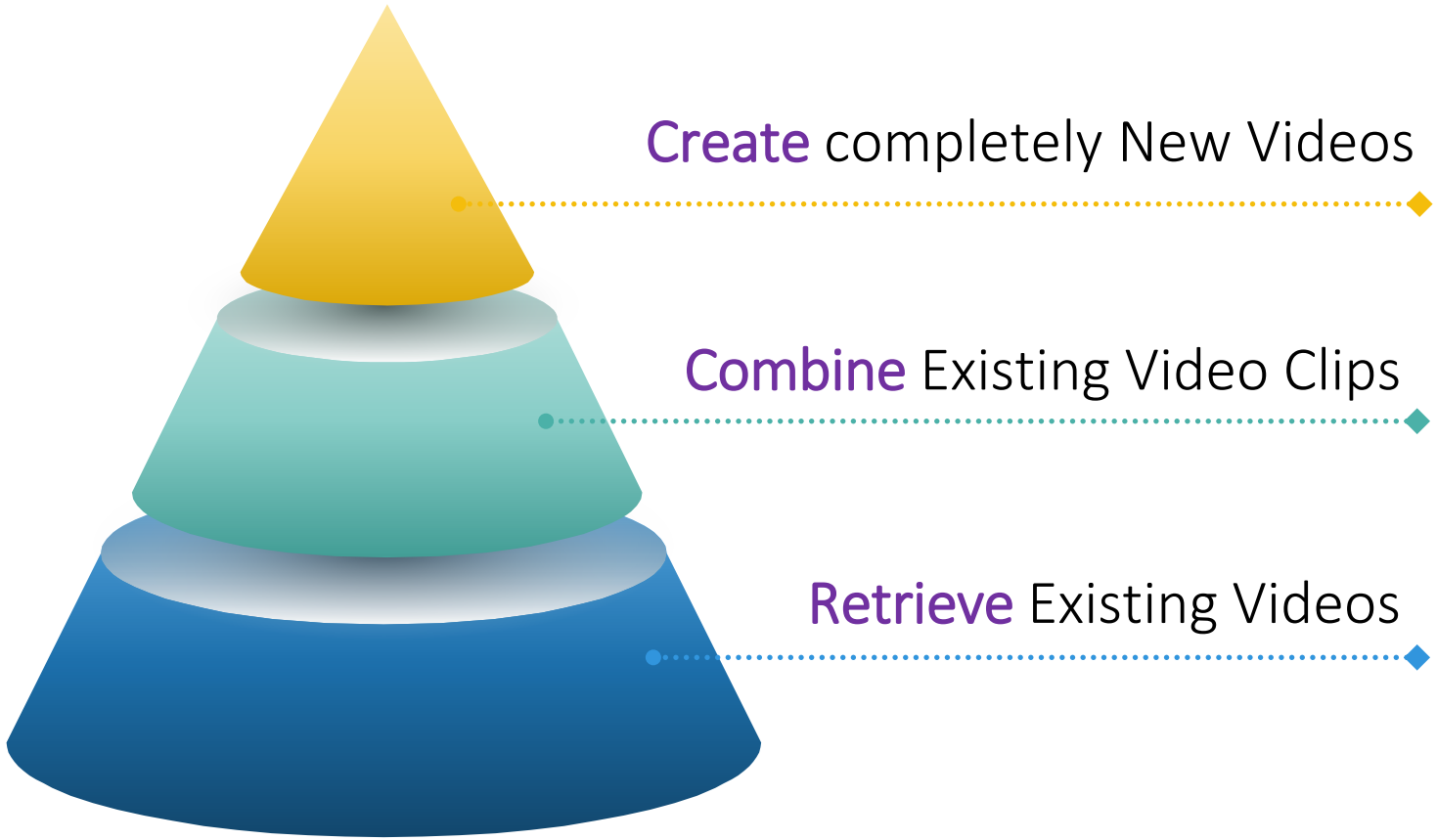
Text-to-Video Generation?

A baseball game is played.

3 Levels of Text-to-Video Generation

Given a query:

"a girl on the voice talks to the judges"



Challenges for Open-Domain Video Generation

A baseball game is played.

(1)  $= \quad 10\ frame \quad \times \quad 64 \times 64\ pixels/frame$

$= \quad 40{,}960\ pixels$

Transformer$\approx 512\ tokens$
Transformer-XL $\approx 3600\ tokens$

(2)  should be

- **semantic consistent** with the input text.
- **smooth** in and between frames.

VQ-VAE for Discrete Image Representation

64

64

$x \in \mathbb{R}^{64\times64\times3}$

Continuous pixels

VQ-VAE Encoder

31	1	41	…	19
7	3	2	…	5
12	2	1	…	2
⋮	⋮	⋮	⋱	9
93	36	22	…	73

16

16

$z \in \mathbb{R}^{16\times16}$

Discrete tokens

VQ-VAE Decoder

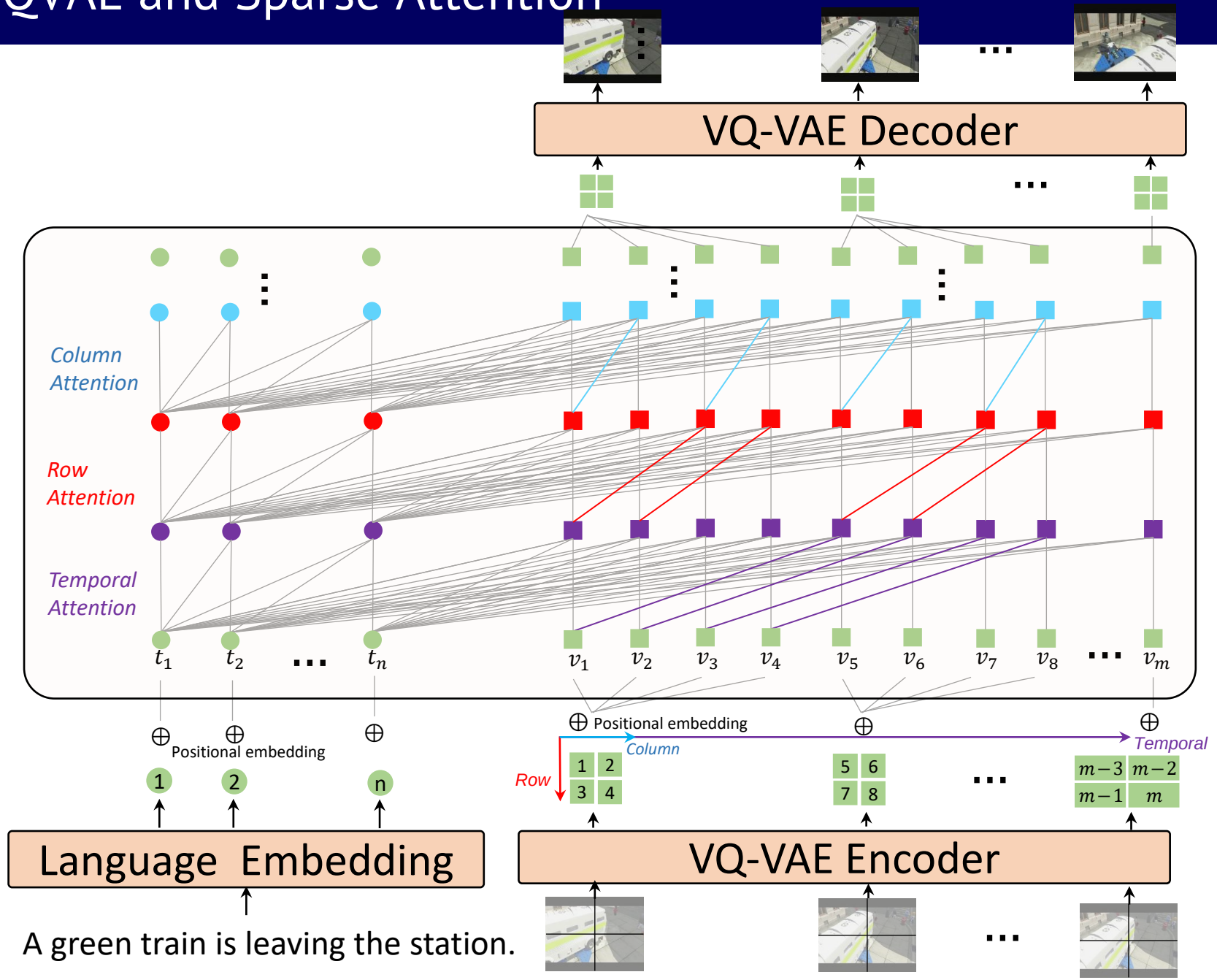
64

64

$x' \in \mathbb{R}^{64\times64\times3}$

Continuous pixels

GODIVA with VQVAE and Sparse Attention



GODIVA: Fine-tuning on MSR-VTT

Text	Grout Truth Video	Generated Video	Text	Grout Truth Video	Generated Video
A baseball game is played.			Someone is giving demo about the car.		
A lady doing make up for herself on her face.			Spongebob and squidward are talking.		
A male instructor is teaching a dance class.			There is a brown hair woman walking on the ramp.		
A man in glasses gives a lesson about how the body works.			The man in the video is showing a brief viewing of how the movie is starting.		

GODIVA: Fine-tuning on Moving MNIST

Model	Input Sentence: Digit 9 is moving down then up.	Input Sentence: Digit 7 moves right then left while digit 3 moves down then up.
VGAN[27]		
SyncDraw[14]		
TGANs[15]		
MocoGAN[25]		
IRC-GAN[6]		
GODIVA(ours)		

Summary

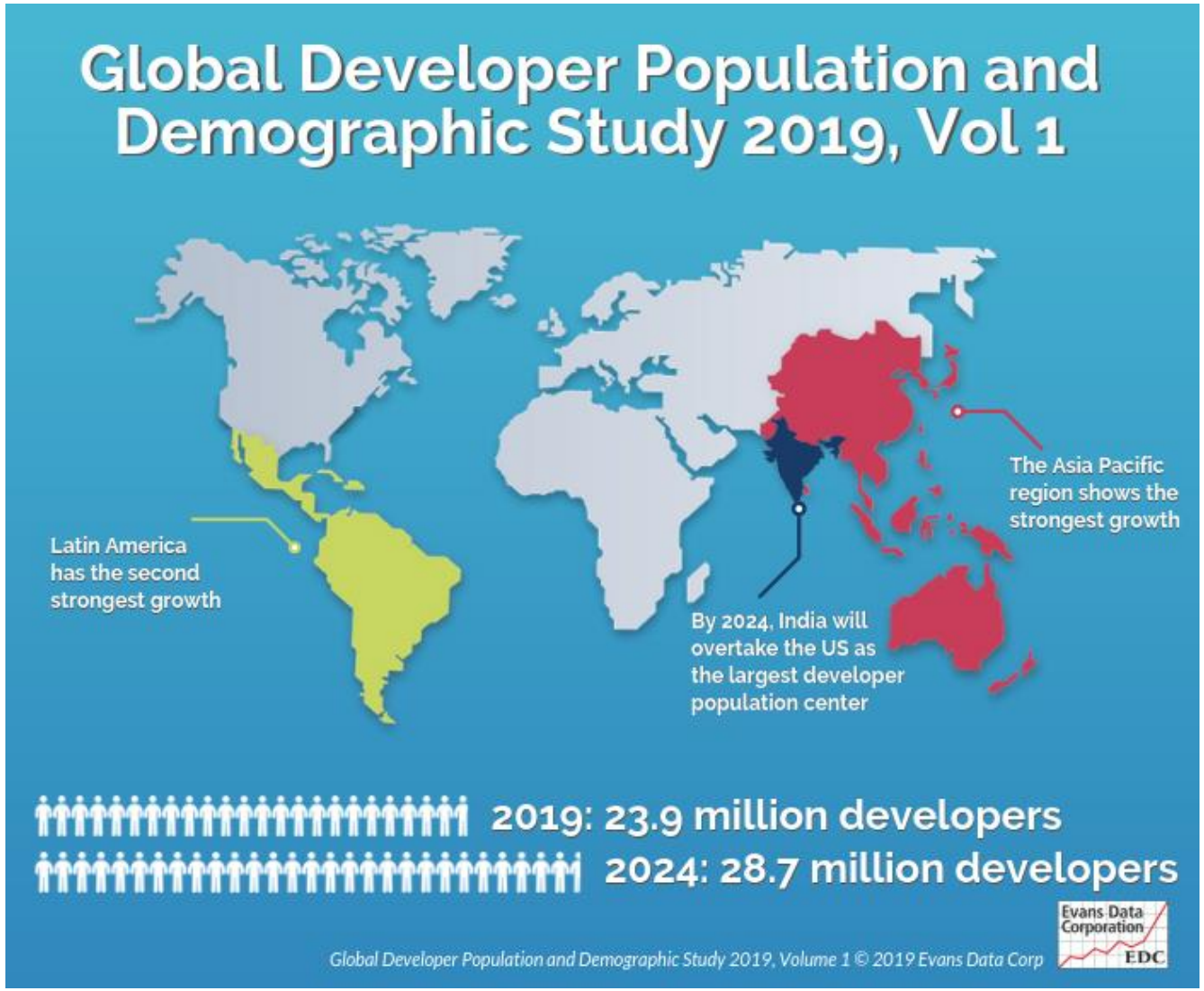
Unify vision and language contents for multimodal pre-training.

Develop an open-domain text-to-video generation model.

Propose GEM as a comprehensive benchmark dataset for language, images and videos.

4). Pre-trained Model & Benchmark for Language + Code

Why is Code Intelligence Important

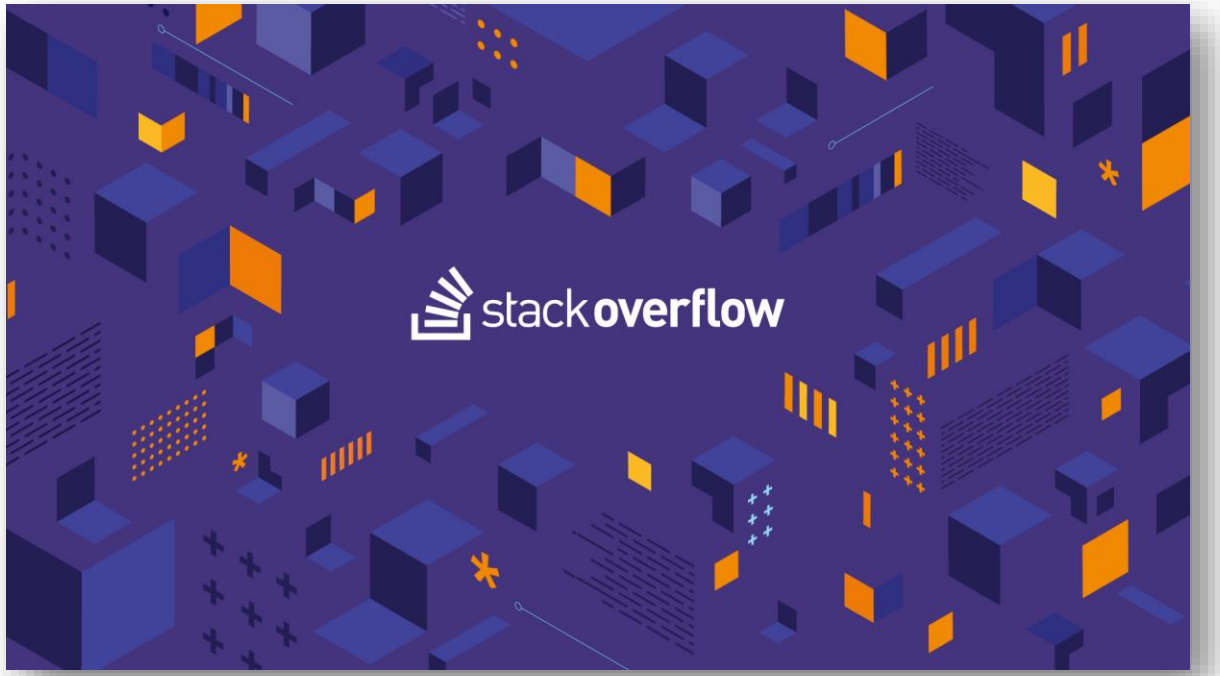


https://evansdata.com/press/viewRelease.php?pressID=278

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

Why NOW: Large-scale Code Corpus on the Web

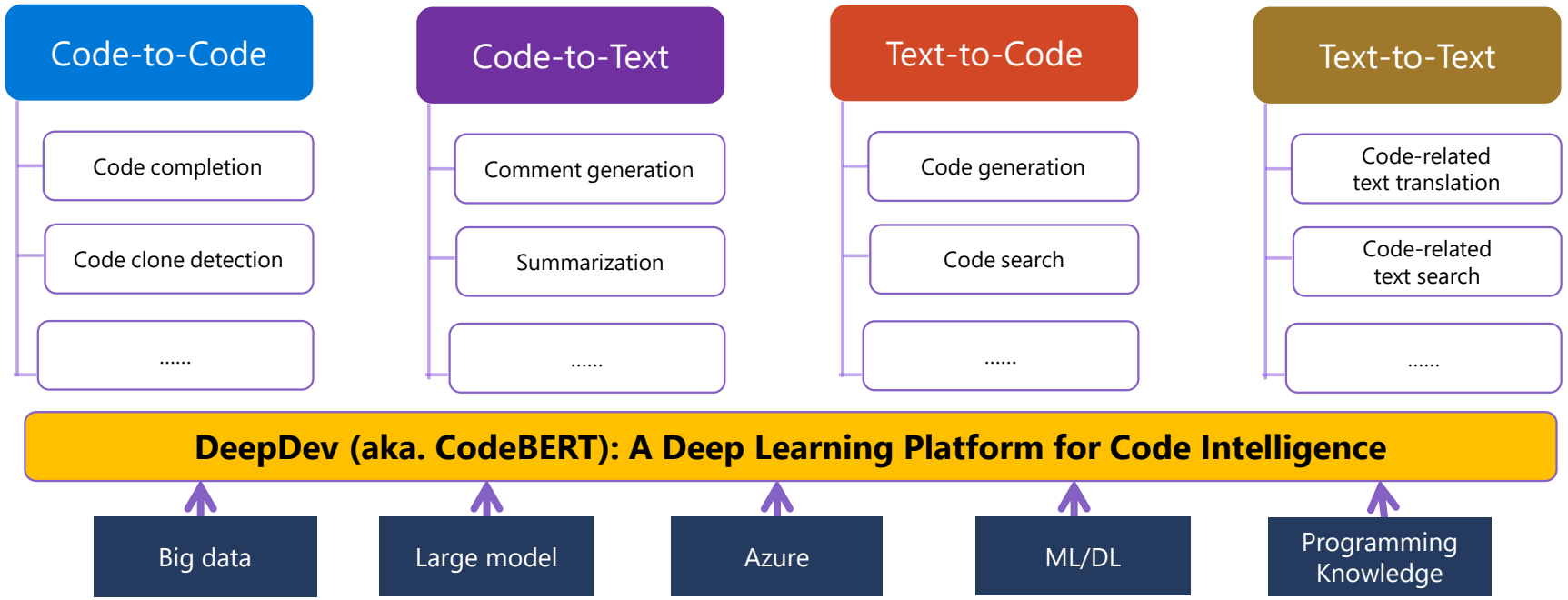
https://github.com/



https://stackoverflow.com/

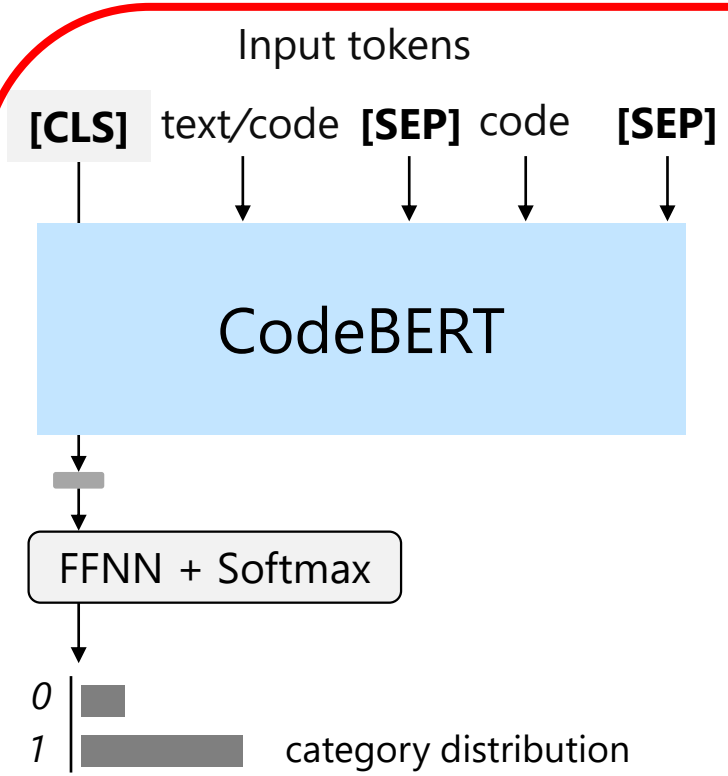
AI for Code

To build large-scale pre-trained models for code to help developers to improve their programming productivity.



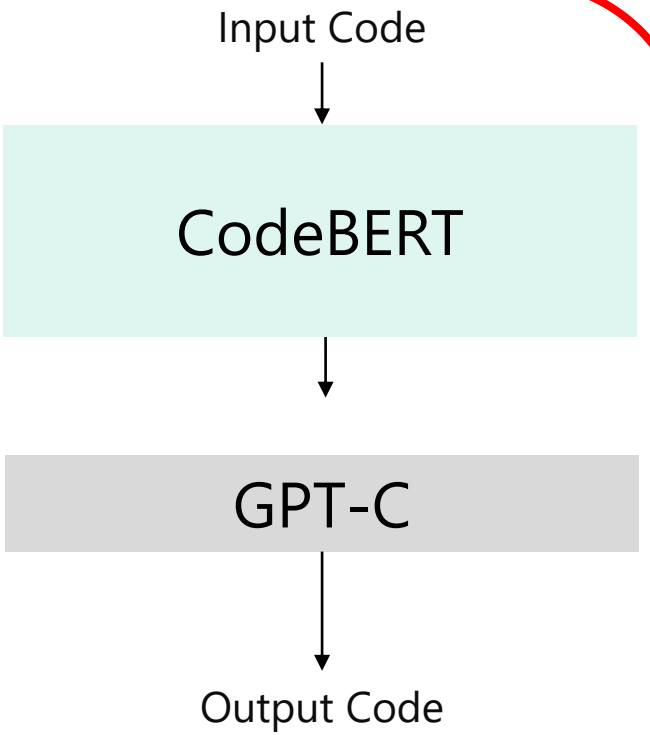
3 Models in CodeBERT Model Family

Understanding

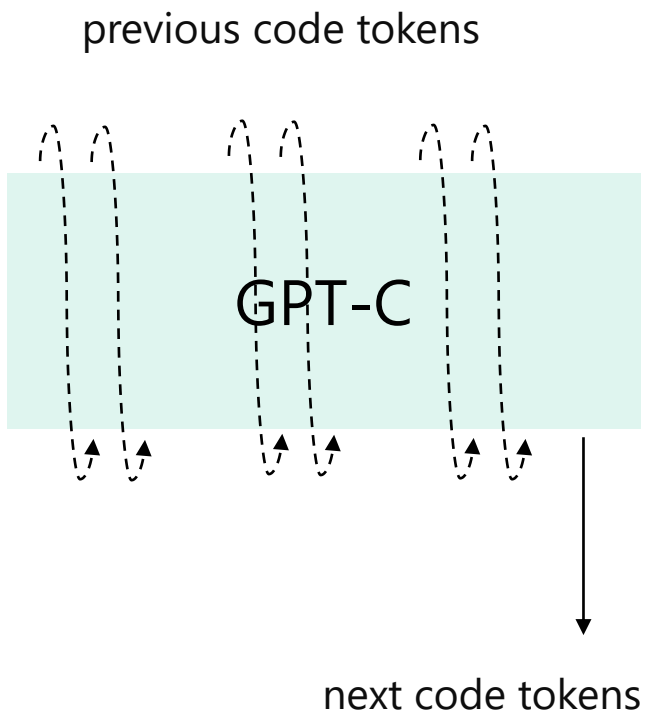


Supported tasks:
- code search
- code clone detection

Generation



Supported tasks:
- code refinement
- code translation



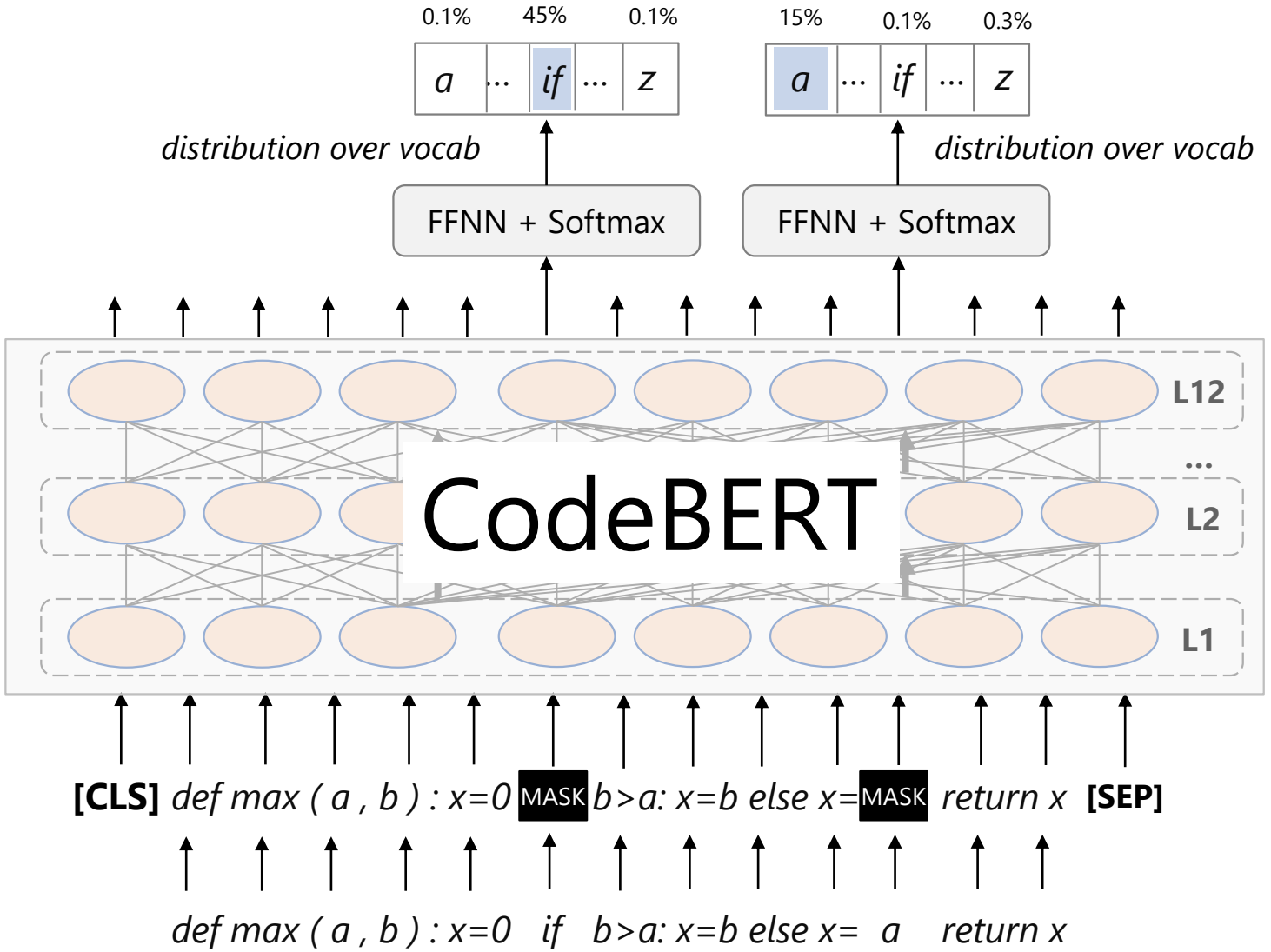
Supported tasks:
- code completion
- code generation

CodeBERT (Model-1): Pre-Train with Code

Predict the masked code token with the output of CodeBERT

Source code

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



Output

Transformer

Randomly mask 15% of tokens

Input

https://arxiv.org/abs/2002.08155

CodeBERT (Model-2): Pre-Train with Code+Text

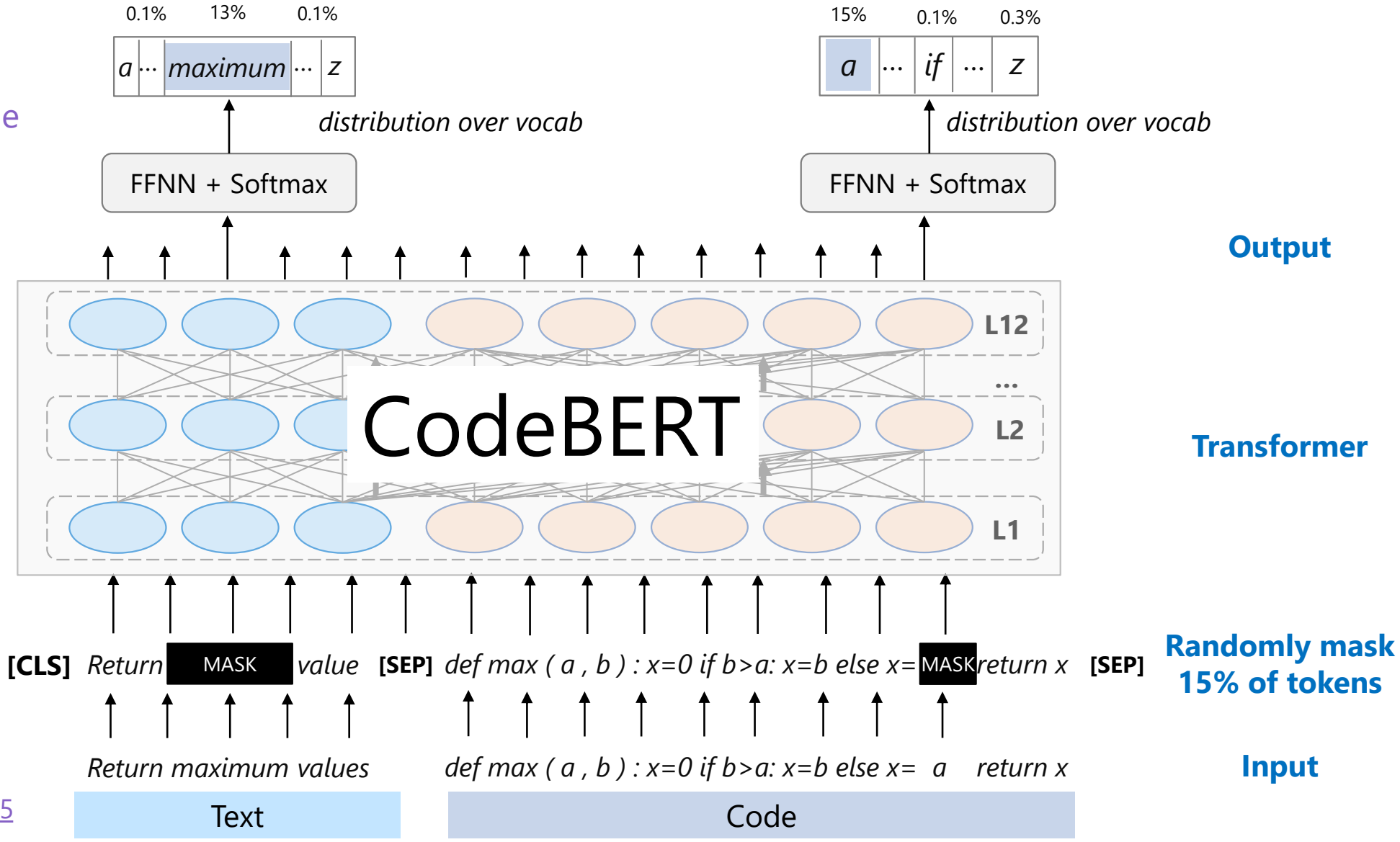
Predict the masked code/text tokens with the output of CodeBERT

Source code

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

Comment

Return maximum value



https://arxiv.org/abs/2002.08155

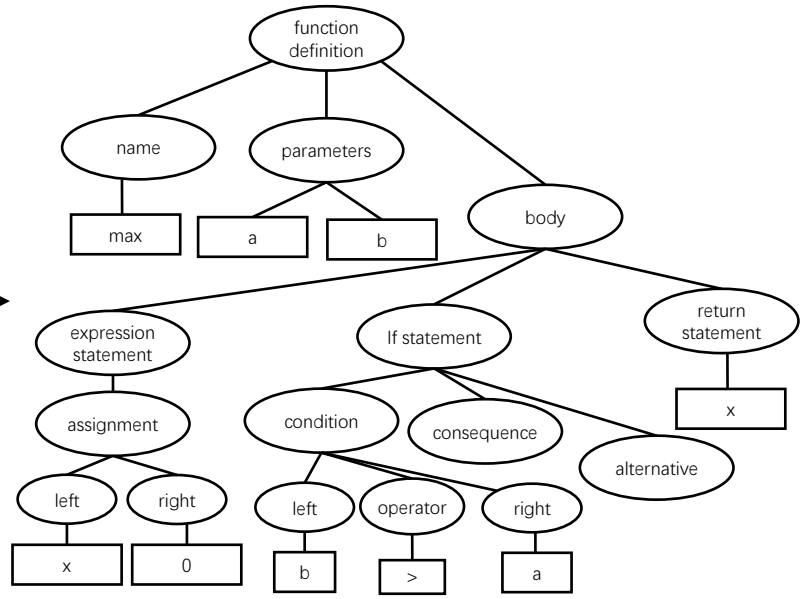
Code Structure

Source code

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

TreeSitter (public tool)

Parse into AST



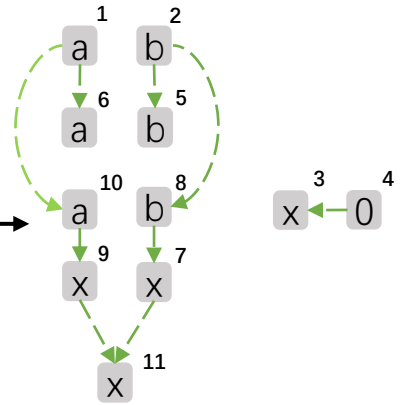
Identify variable sequence in AST

Identify variable sequence

```
def max(a[1], b[2]):
    x[3]=0[4]
    if b[5]>a[6]:
        x[7]=b[8]
    else:
        x[9]=a[10]
    return x[11]
```

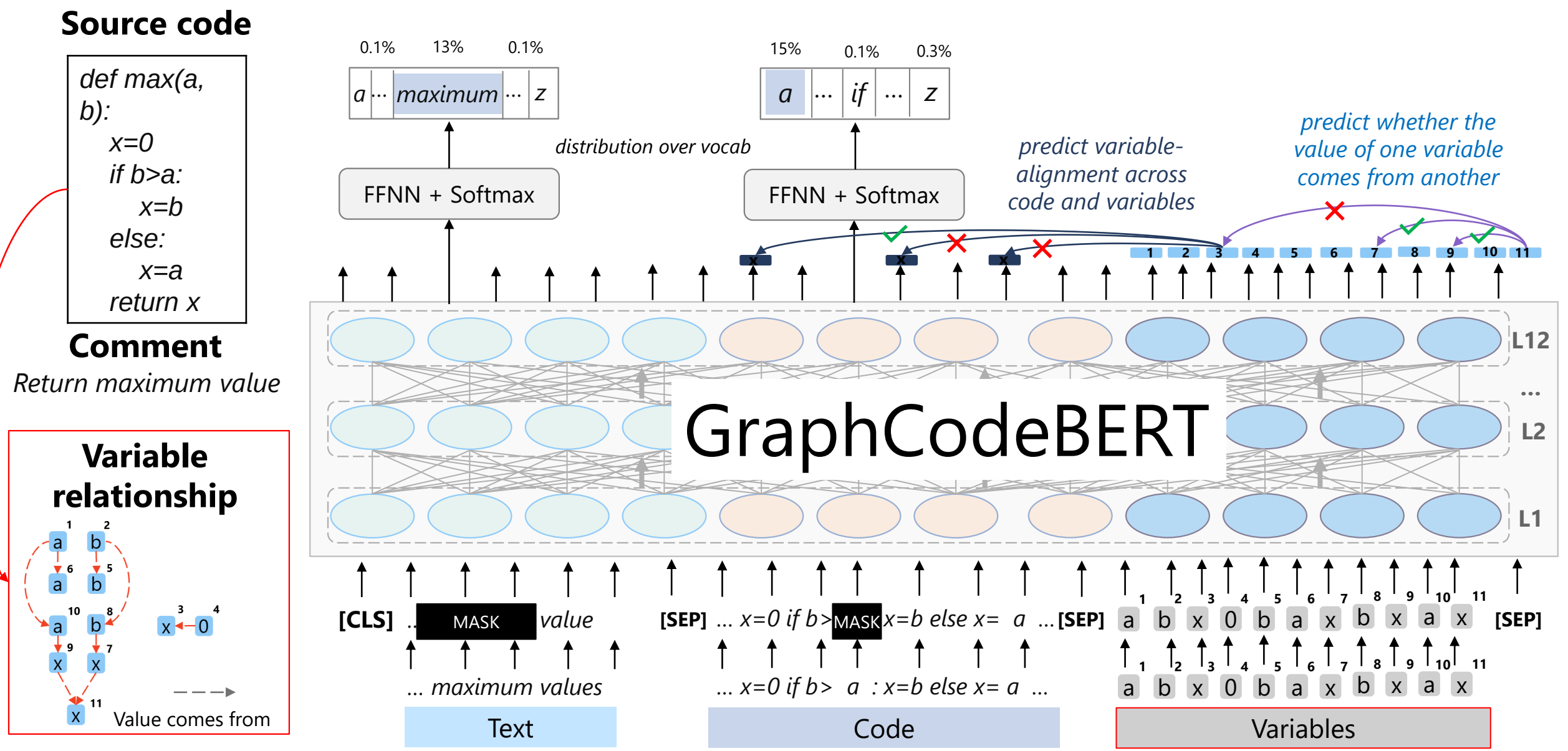
Extract variable relationship from AST according to paths between variables

Variable relationship



10k functions per minute with one CPU

CodeBERT (Model-3): Pre-Train with Code+Text+Structure



Understanding Results

Results on code search.

Model	Ruby	JavaScript	Go	Python	Java	PHP	Overall
BiRNN	0.213	0.193	0.688	0.290	0.304	0.338	0.338
RoBERTa	0.587	0.517	0.850	0.587	0.599	0.560	0.617
RoBERTa (code)	0.628	0.562	0.859	0.610	0.620	0.579	0.643
CodeBERT	0.679	0.620	0.882	0.672	0.676	0.628	0.693
GraphCodeBERT	**0.703**	**0.644**	**0.897**	**0.692**	**0.691**	**0.649**	**0.713**

Input: *How to read text file in Python?*

Output:

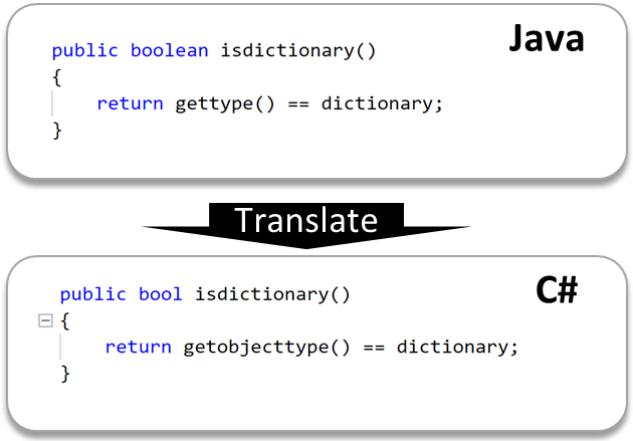
```python
def read_text_file(filename, encoding="utf-8"):
    """
    Reads a file under python3 with encoding (default UTF-8).
    Also works under python2, without encoding.
    Uses the EAFP (https://docs.python.org/2/glossary.html#term-eafp)
    principle.
    """
    try:
        with open(filename, 'r', encoding) as f:
            r = f.read()
    except TypeError:
        with open(filename, 'r') as f:
            r = f.read()
    return r
```

code source from GitHub

Generation Results

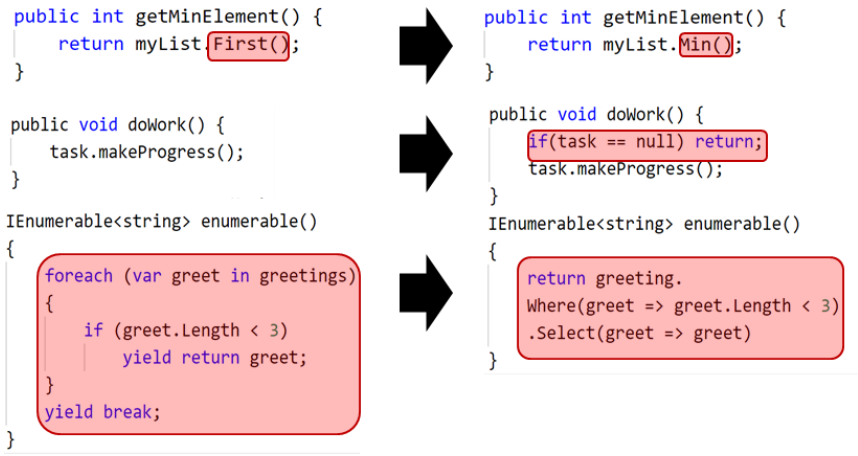
Results on code translation.

Model	Java->C#		C#->Java	
	BLEU	Accuracy	BLEU	Accuracy
Naïve	18.54	0.0	18.69	0.0
PBSMT	43.53	12.5	40.06	16.1
RoBERTa (code)	77.46	56.1	71.99	57.9
CodeBERT	79.92	59.0	72.14	58.0
GraphCodeBERT	**80.58**	**59.4**	**72.64**	**58.8**



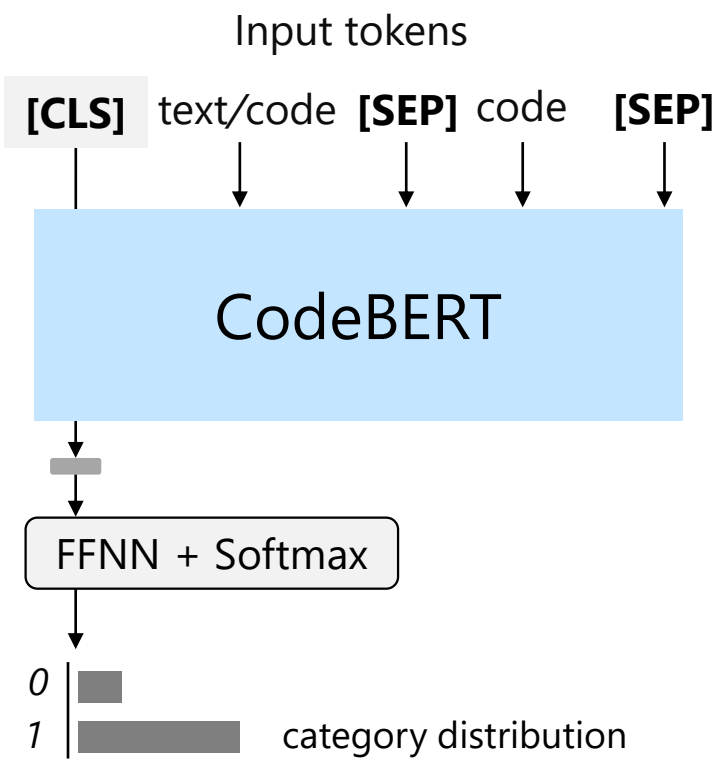
Results on code refinement.

Model	Small		Medium	
	BLEU	Accuracy	BLEU	Accuracy
Naïve	78.06	0.0	90.91	0.0
LSTM	76.76	10.0	72.08	2.5
RoBERTa (code)	77.30	15.9	90.07	4.1
CodeBERT	77.42	16.4	90.07	5.2
GraphCodeBERT	**80.02**	**17.3**	**91.31**	**9.1**



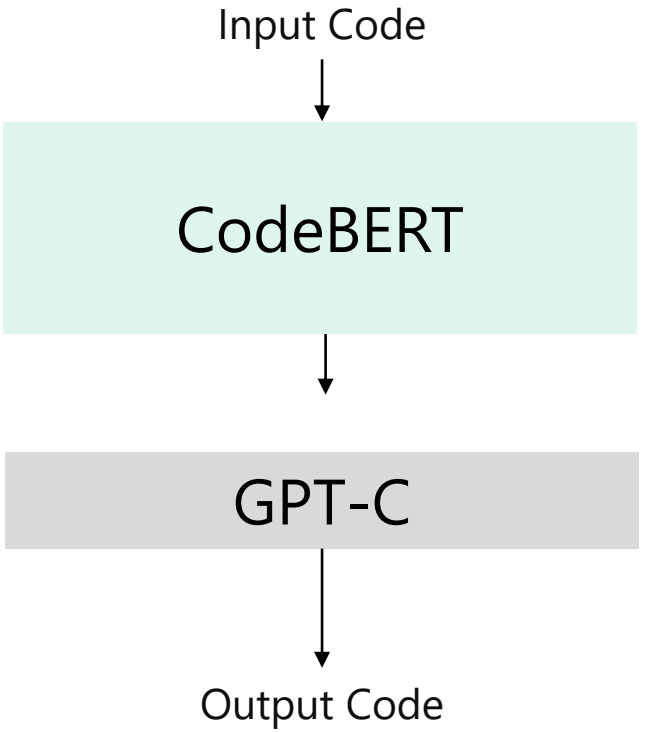
3 Models in CodeBERT Model Family

Understanding

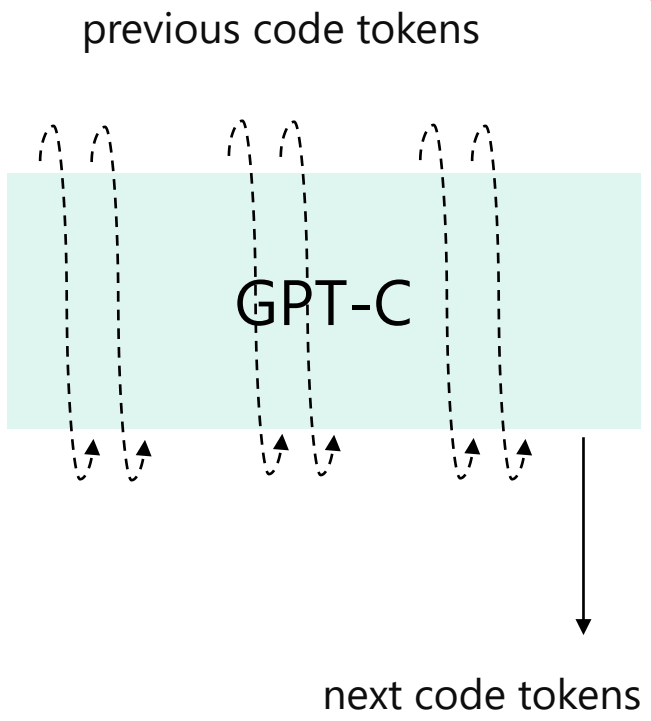


Supported tasks:
- code search
- code clone detection

Generation



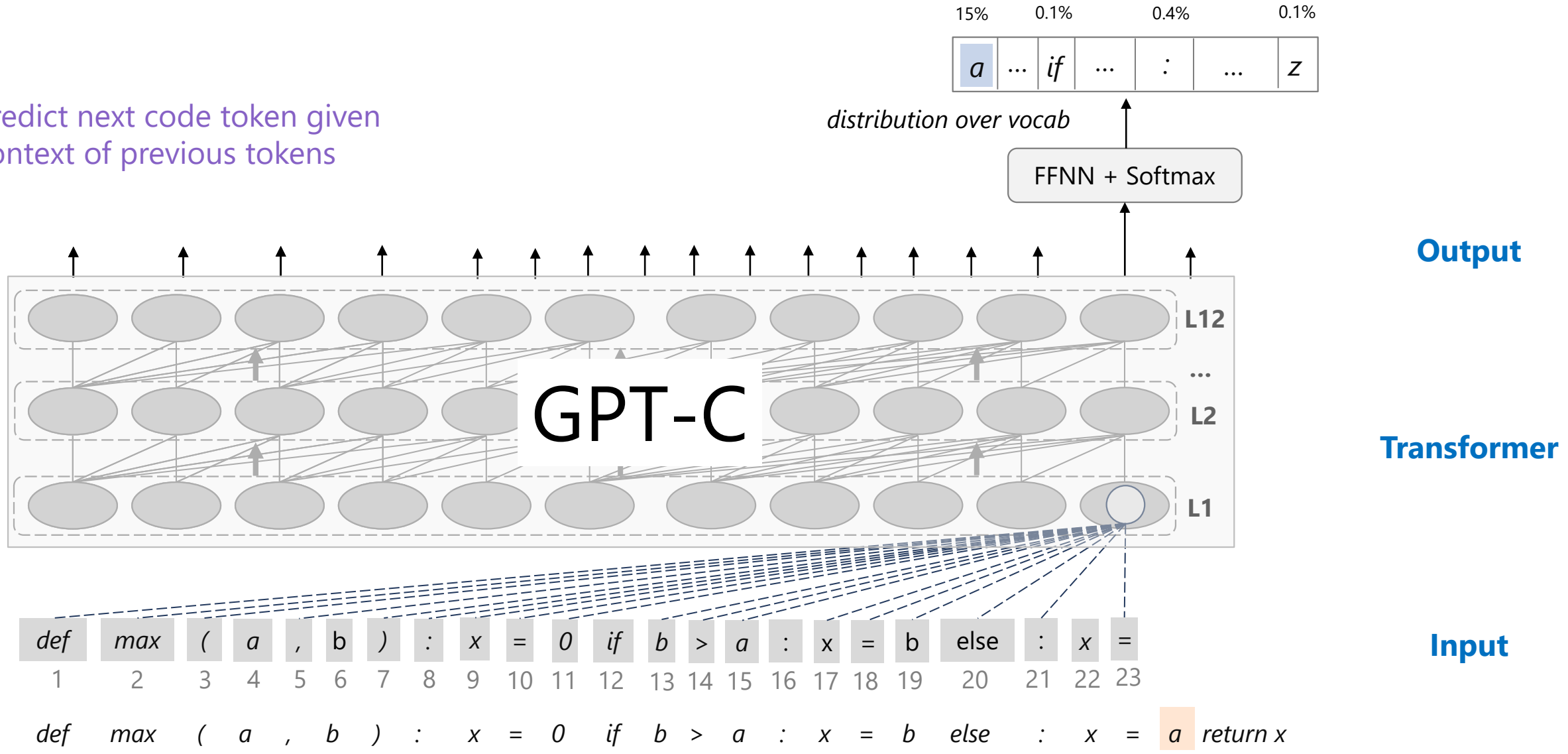
Supported tasks:
- code refinement
- code translation



Supported tasks:
- code completion
- code generation

GPT-C (v1): Multilingual Pre-trained GPT for Code

Predict next code token given context of previous tokens



Trained for 10 PLs: JavaScript, C, Java, Go, PHP, Python, C++, C#, Ruby, TypeScript

Evaluation on Code Completion

- Corpus
  - 10 programming langauges
  - 354K code projects
  - 18B code lines
  - Split 8:1:1 as Train/Dev/Test

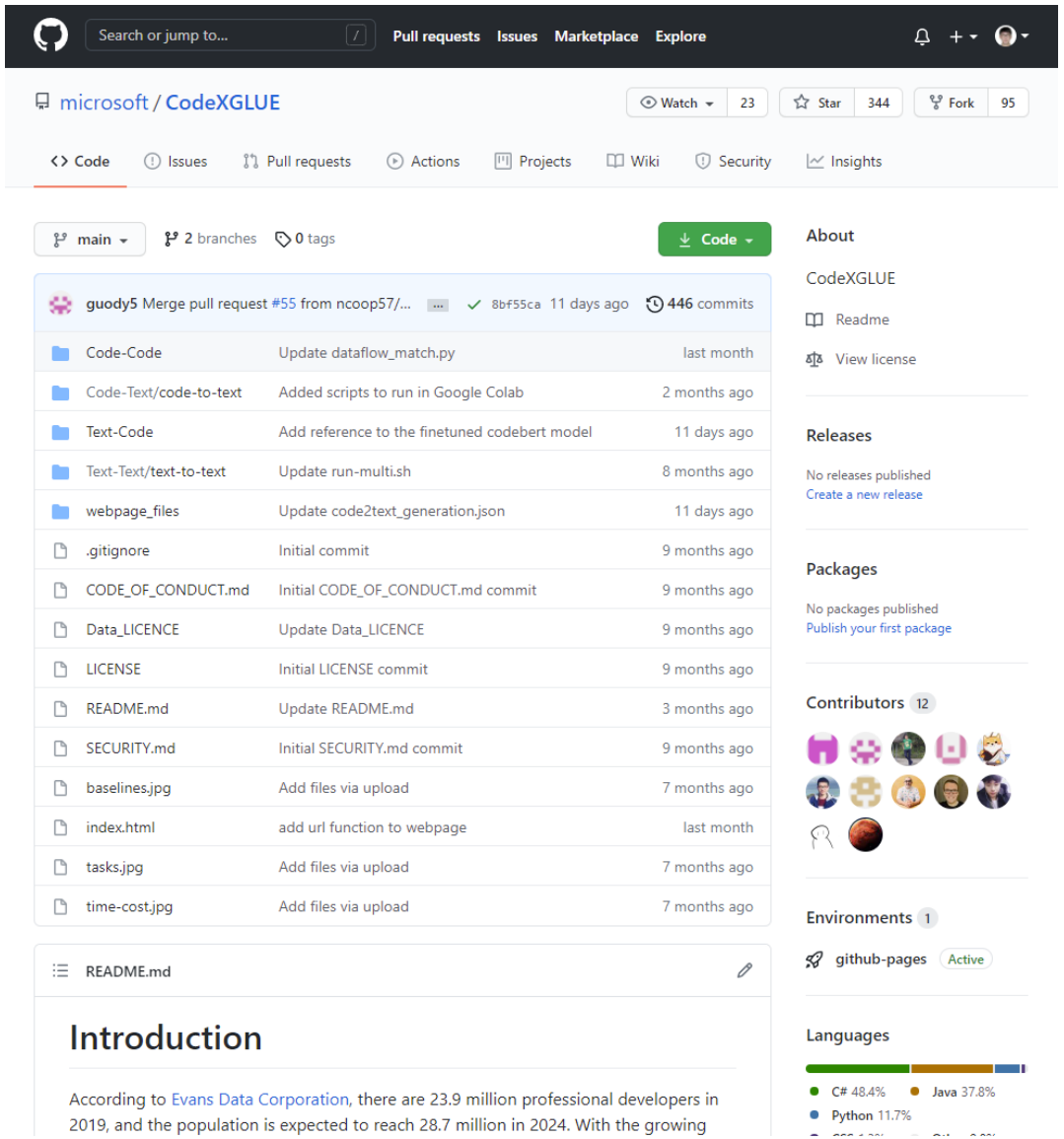
Programming language	#Projects	#Files (↓)	#Lines
JavaScript	113,890	15,330,706	4,226,121,235
C	19,900	13,462,890	7,253,471,852
Java	46,921	10,385,540	1,491,132,997
Go	17,922	5,720,219	1,997,845,604
PHP	24,625	4,691,140	653,891,761
Python	71,343	4,465,808	854,503,198
C++	20,958	4,293,413	1,400,309,370
C#	17,387	3,765,835	550,267,681
Ruby	17,804	1,663,262	137,558,948
TypeScript	3,801	466,924	64,671,728

Model	Test language	ROUGE-L			Average edit similarity (Levenshtein, %)
		Precision	Recall	F1	
GPT-C (12L)	Python	0.72	0.80	0.76	83.0
	C#	0.56	0.75	0.64	77.5
	JavaScript	0.71	0.81	0.76	92.2
	Go	0.65	0.72	0.68	81.0
	Scala **(zero-shot)**	0.41	0.54	0.47	64.2

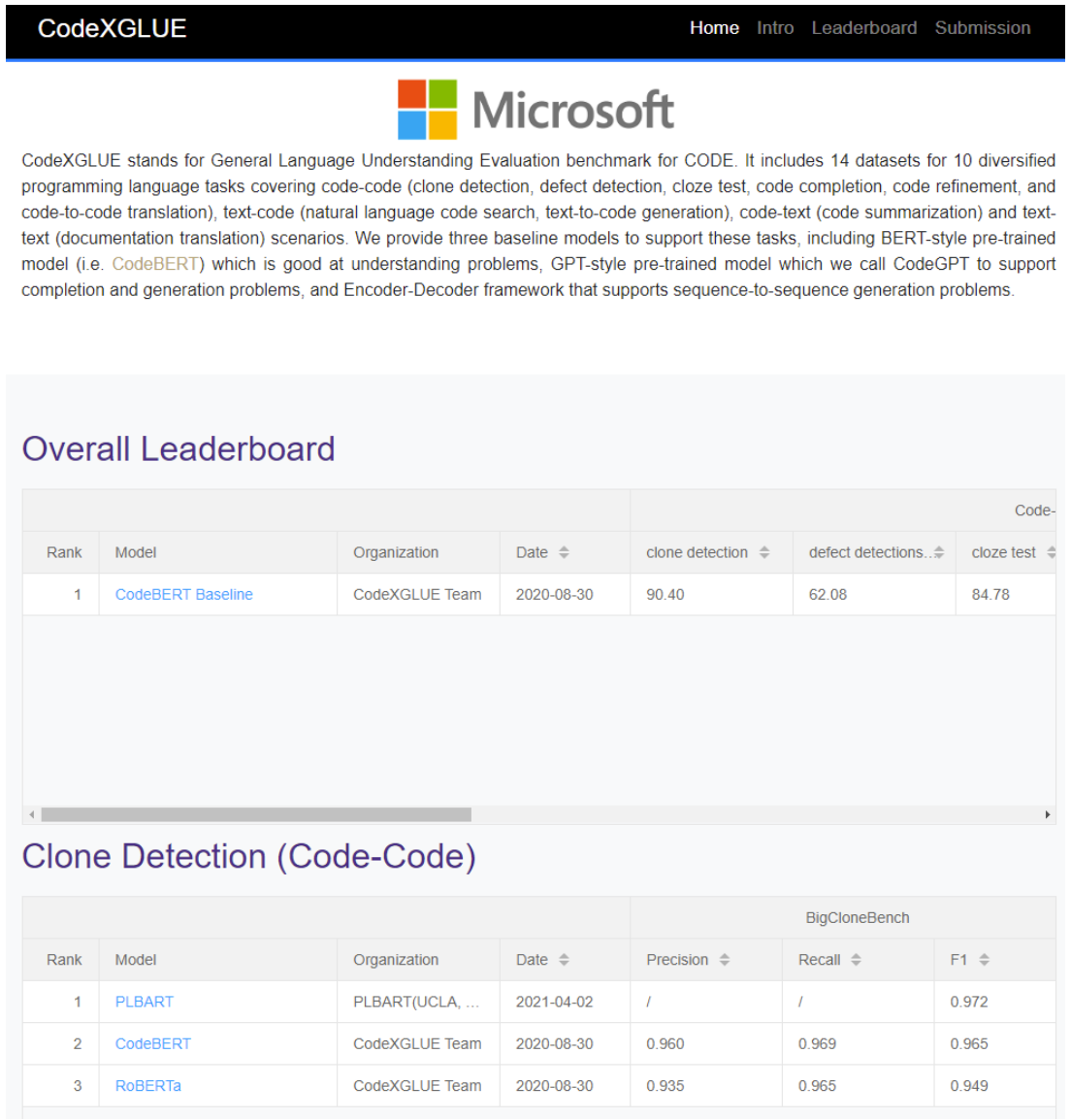
CodeXGLUE: 14 datasets for 10 tasks

Category	Task	Dataset Name	Language	Train/Dev/Test Size	Baselines	Dataset Provider	Task definition
Code-Code	Clone Detection	BigCloneBench	Java	900K/416K/416K	CodeBERT	Univ. of Saskatchewan	Predict semantic equivalence for a pair of codes.
		POJ-104	C/C++	32K/8K/12K		Peking Univ	Retrieve semantically similar codes.
	Defect Detection	Defects4J	C	21k/2.7k/2.7k		Univ. of Washington	Identify whether a function is vulnerable.
	Cloze Testing	CT-all	Python, Java, PHP, JavaScript, Ruby, Go	-/-/176k		Created by MSRA based on CodeSearchNet	Tokens to be predicted come from the entire vocab.
		CT-max/min	Python, Java, PHP, JavaScript, Ruby, Go	-/-/2.6k		Created by MSRA based on CodeSearchNet	Tokens to be predicted come from {max, min}.
	Code Completion	PY150	Python	100k/5k/50k	CodeGPT	ETH Zurich, line-level data added by MSRA	Predict following tokens given contexts of codes.
		GitHub Java Corpus	Java	13k/7k/8k		Univ. of Edinburgh, line-level data added by MSRA	
	Code Refinement	Bugs2Fix	Java	98K/12K/12K	Encoder-Decoder	The College of William and Mary	Automatically refine codes by fixing bugs.
	Code Translation	CodeTrans	Java-C#	10K/0.5K/1K		MSRA	Translate the codes from one programming language to another programming language.
Text-Code	NL Code Search	CodeSearchnet, AdvTest	Python	251K/9.6K/19K	CodeBERT	GitHub + MSR Cambridge, test provided by MSRA	Given a natural language query as input, find semantically similar codes.
		StacQC, WebQueryTest	Python	2.9k/0.9k/1.9k		The Ohio State Univ, test provided by MSRA	Given a pair of natural language and code, predict whether they are relevant or not.
	Text-to-Code Generation	CONCODE	Java	100K/2K/2K	CodeGPT	Univ. of Washington	Given a natural language docstring/comment as input, generate a code.
Code-Text	Code Summarization	CodeSearchNet*	Python, Java, PHP, JavaScript, Ruby, Go	908K/45K/53K	Encoder-Decoder	Filtered based on CodeSearchNet data	Given a code, generate its natural language docstring/comment.
Text-Text	Documentation Translation	Microsoft Docs	English-Latvian/Danish/Norwegian/Chinese	156K/4K/4K		MSRA	Translate code documentation between human languages (e.g. En-Zh), intended to test low-resource multi-lingual translation.

CodeXGLUE: 14 datasets for 10 tasks



https://github.com/microsoft/CodeXGLUE



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

Submission Status & Call for Participation

There have been some research teams submitting their results to CodeXGLUE

- **IBM Research** (Defect Detection)
- **UCLA** (Clone Detection, Defect Detection, Code Refinement, Code Translation, Code Generation, Code Summarization)
- **Columbia University** (Clone Detection, Defect Detection, Code Refinement, Code Translation, Code Generation, Code Summarization)
- **Case Western Reserve University** (Defect Detection, Code Refinement, Code Generation, Code Summarization)
- **USTC+MSRA** (Code Summarization)

Defect Detection (Code-Code)

Rank	Model	Organization	Date	Accuracy
1	CoTexT	Case Western R...	2021-04-23	66.62
2	C-BERT	AI4VA (IBM Res...	2021-03-19	65.45
3	PLBART	PLBART(UCLA, ...	2021-04-02	63.18
4	CodeBERT	CodeXGLUE Team	2020-08-30	62.08
5	RoBERTa	CodeXGLUE Team	2020-08-30	61.05
6	TextCNN	CodeXGLUE Team	2020-08-30	60.69
7	BiLSTM	CodeXGLUE Team	2020-08-30	59.37

Code Generation (Text-Code)

				Text2Code Generation		
Rank	Model	Organization	Date	EM	BLEU	CodeBLEU
1	CoTexT	Case Western R...	2021-04-23	20.1	37.4	40.14
2	PLBART	PLBART(UCLA, ...	2021-04-02	18.75	36.69	38.52
3	CodeGPT-adapted	CodeXGLUE Team	2020-08-30	20.1	32.79	35.98
4	CodeGPT	CodeXGLUE Team	2020-08-30	18.25	28.69	32.71
5	GPT-2(12L)	CodeXGLUE Team	2020-08-30	17.35	25.37	29.69
6	Iyer-Simp+200 idoms	CodeXGLUE Team	2020-08-30	12.20	26.60	/

Summary

Pre-training techniques can be applied to programming langauges.

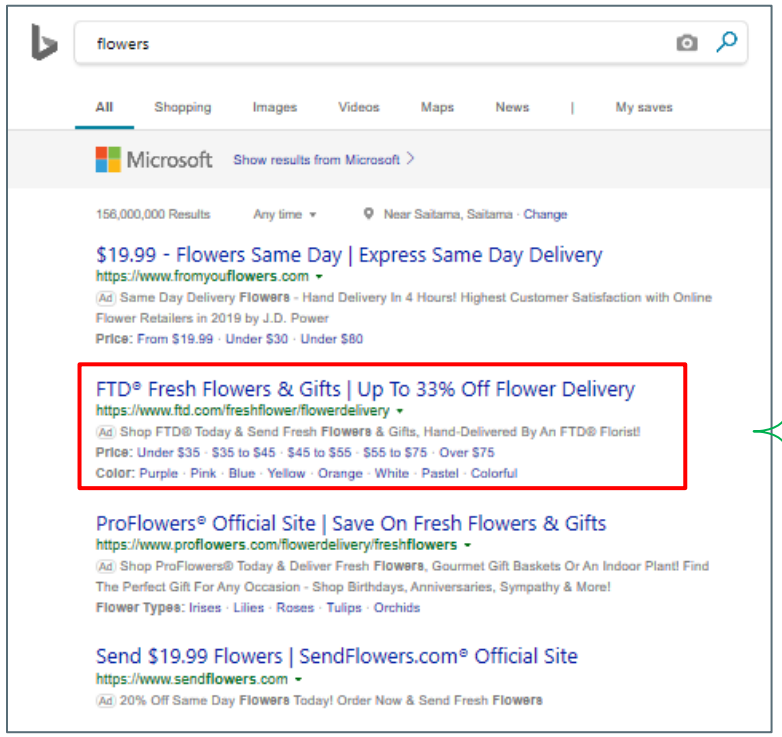
Code structures can lead to better pre-trained models for programming languages.

MSRA released CodeXGLUE (https://microsoft.github.io/CodeXGLUE/) as a new benchmark for SE community.

5). Pre-trained Model & Benchmark for Language Generation

ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *EMNLP-Findings 2020*

BACKGROUND



flower s

gift

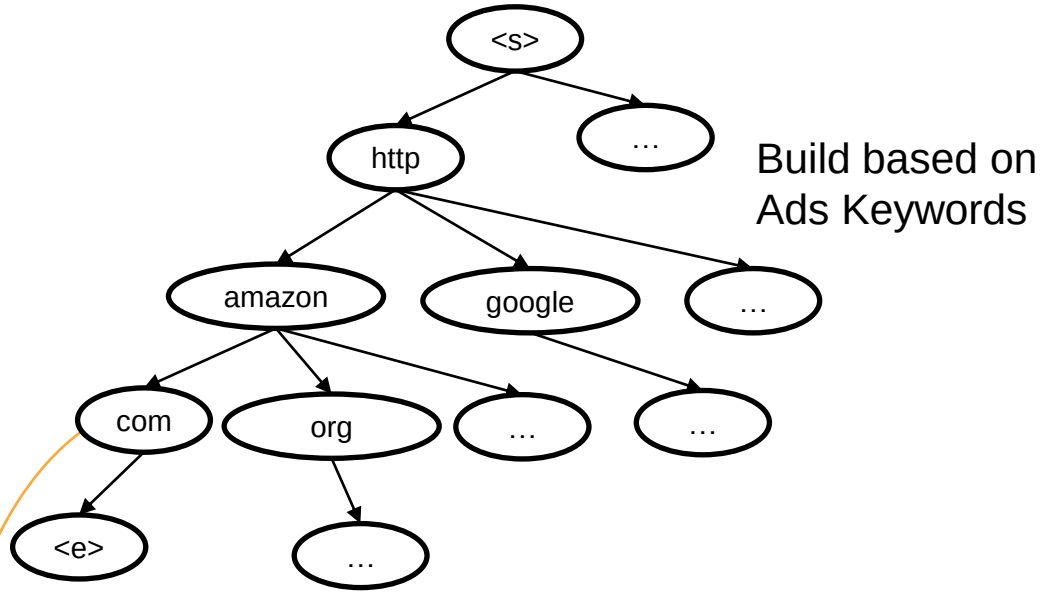
...

https://www.bing.com/

Expanded Exact Match

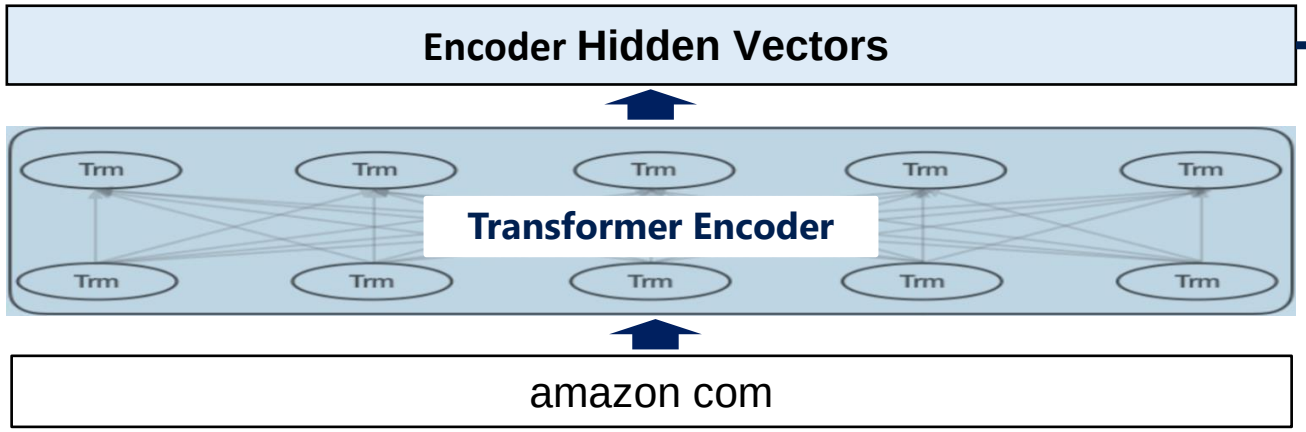
Query	Keywords
flowers	flower s
flowers	com flower
flowers	com flowers
flowers	flowered
flowers	are flowers
flowers	floweres
flowers	dot flower
flowers	s flower
flowers	http www flowers com
flowers	are flower
flowers	flowere
flowers	flowers net
flowers	flowers com
flowers	www flowers

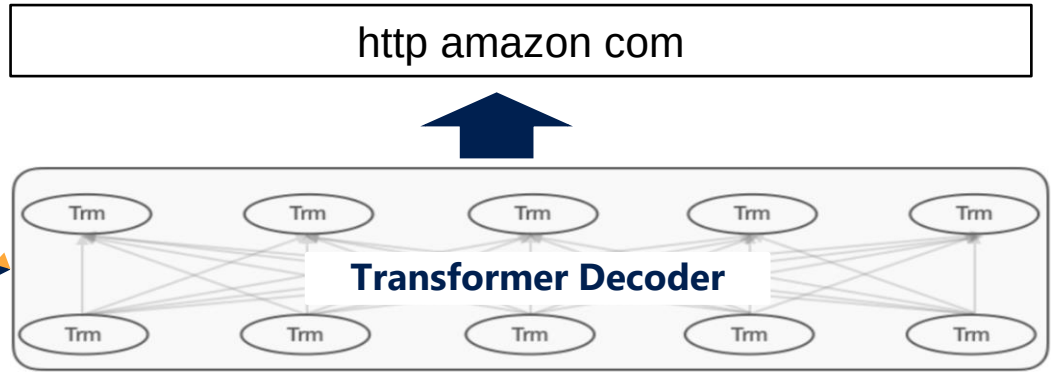
NLG with Trie



Build based on
Ads Keywords

We build a Trie for
constraint decoding

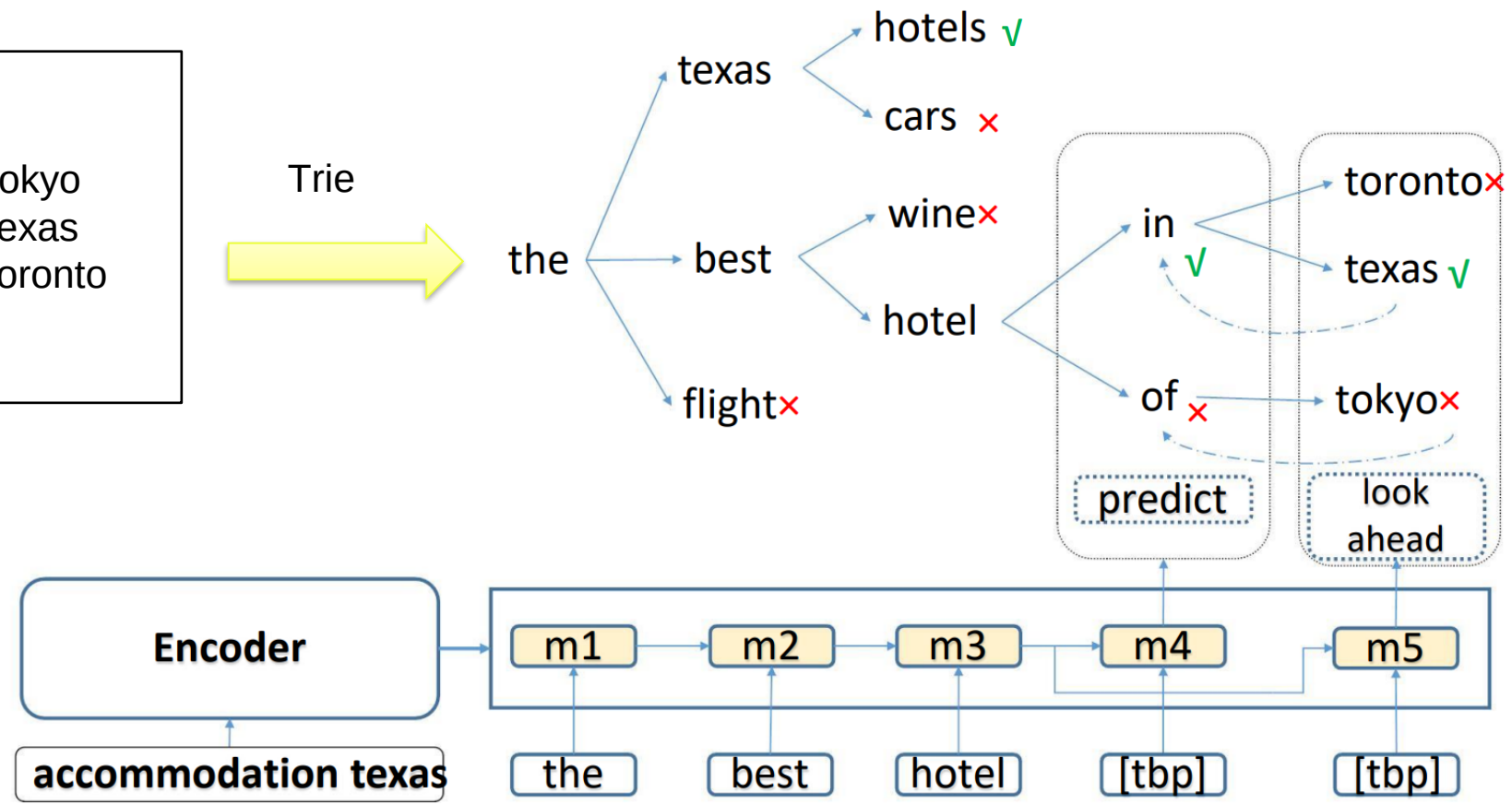




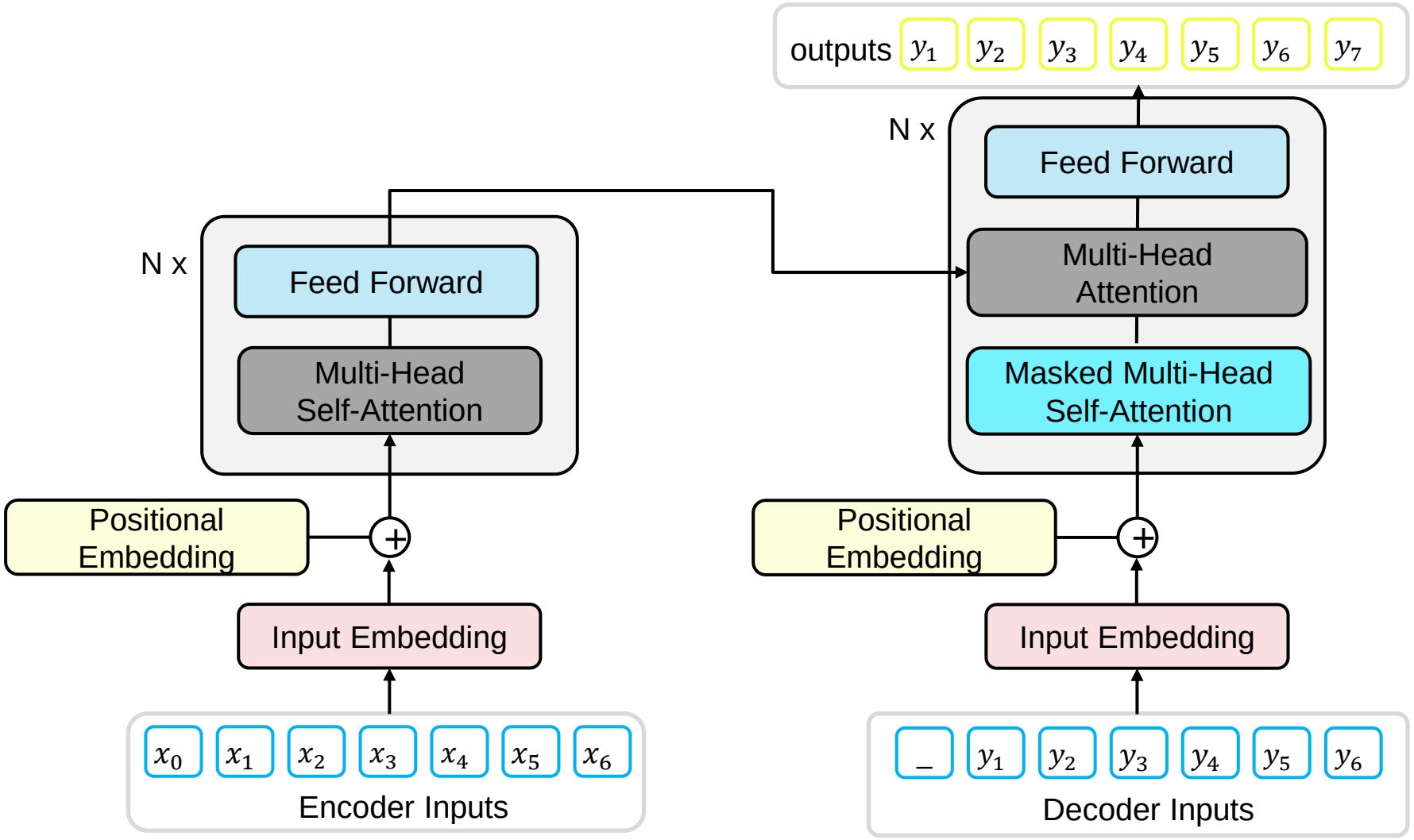
Motivation of ProphetNet

the flight
the best wine
the best hotel of tokyo
the best hotel in texas
the best hotel in toronto
the texas hotels
the texas cars

Trie

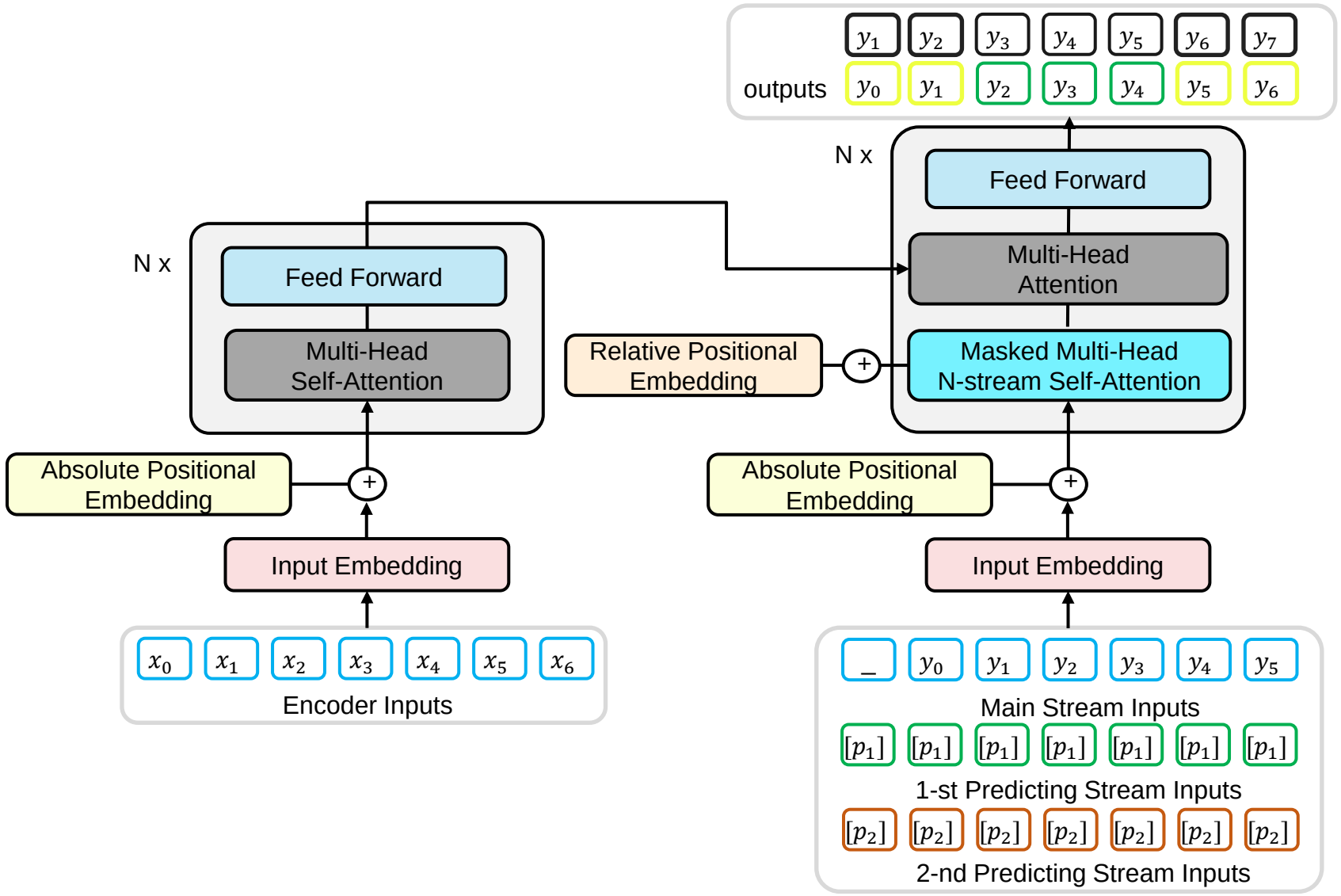


Architecture of Transformer



Attention Is All You Need, NIPS2017

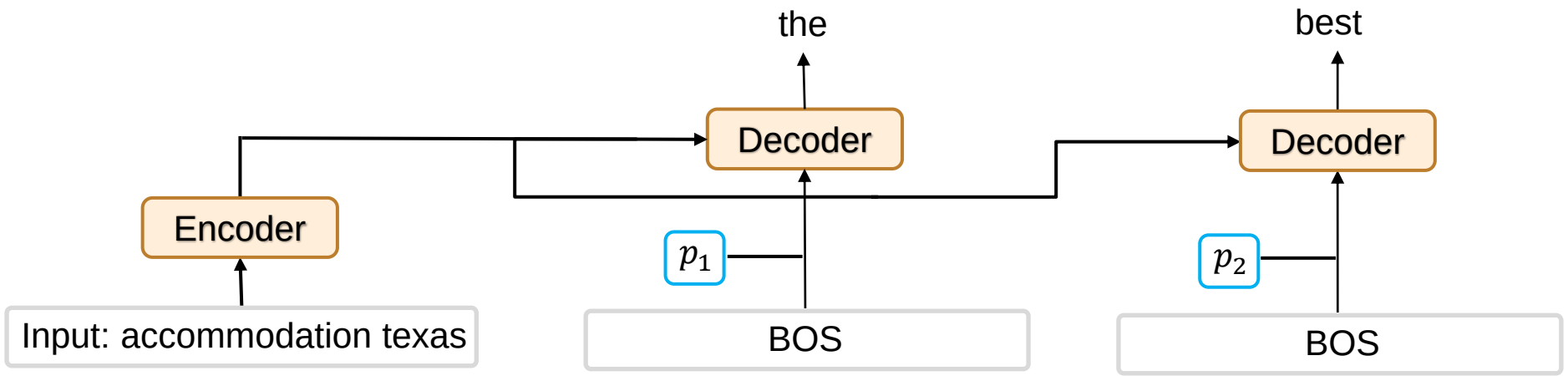
Architecture of ProphetNet



ProphetNet Training

e.g.
Input: accommodation texas
Output: the best hotel in texas

$$L = \alpha_1 \cdot L_{the} + \alpha_2 \cdot L_{best}$$

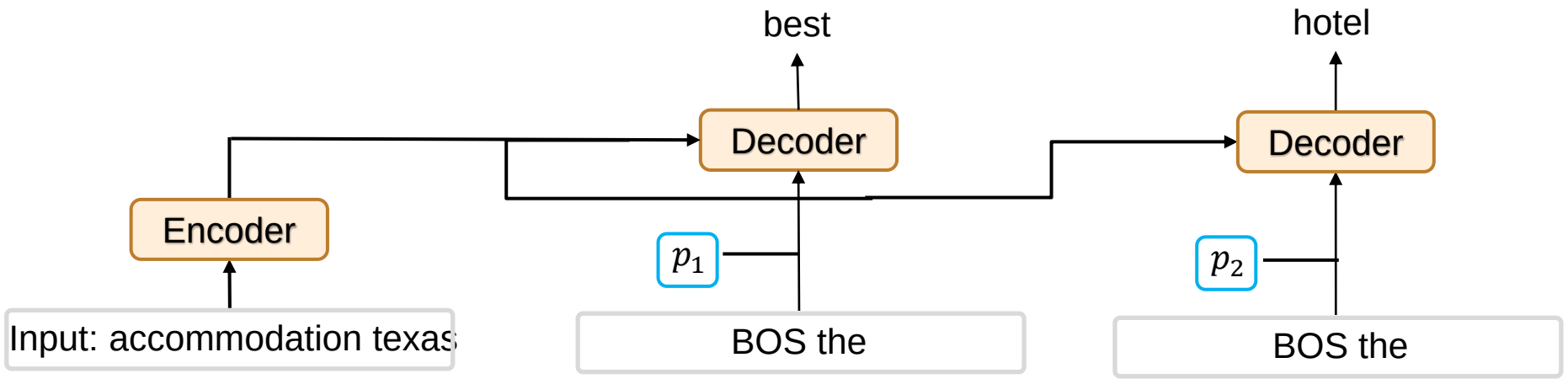


ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *EMNLP-Findings 2020*
Yeyun Gong, Yu Yan*, Weizhen Qi*, Dayiheng Liu*, Nan Duan, Jiusheng Chen, Bruce Zhang, Ming Zhou.*

ProphetNet Training

e.g.
Input: accommodation texas
Output: the best hotel in texas

$$L = \alpha_1 \cdot L_{best} + \alpha_2 \cdot L_{hotel}$$

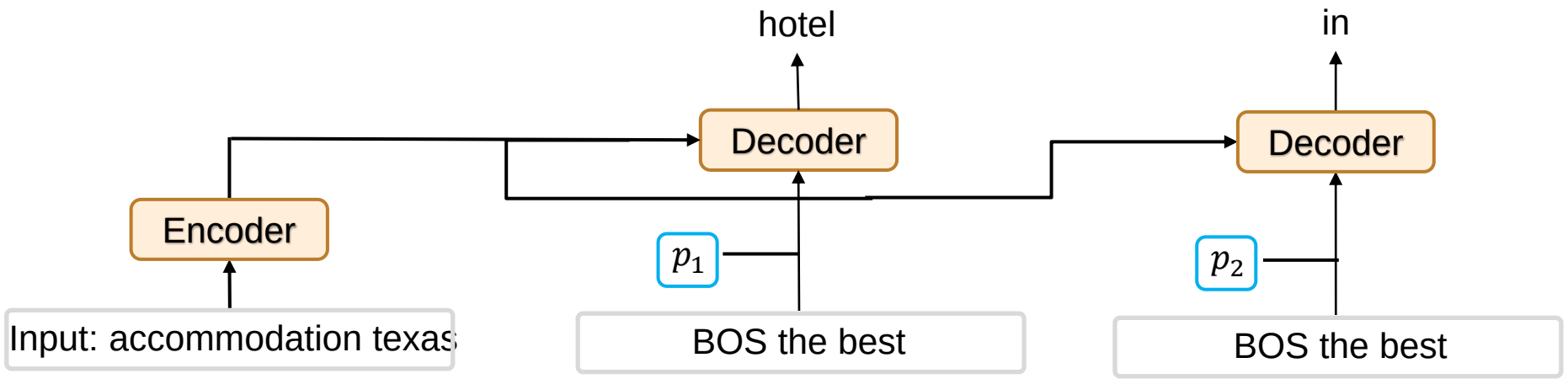


ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *EMNLP-Findings 2020*
Yeyun Gong, Yu Yan*, Weizhen Qi*, Dayiheng Liu*, Nan Duan, Jiusheng Chen, Bruce Zhang, Ming Zhou.*

ProphetNet Training

e.g.
Input: accommodation texas
Output: the best hotel in texas

$$L = \alpha_1 \cdot L_{hotel} + \alpha_2 \cdot L_{in}$$

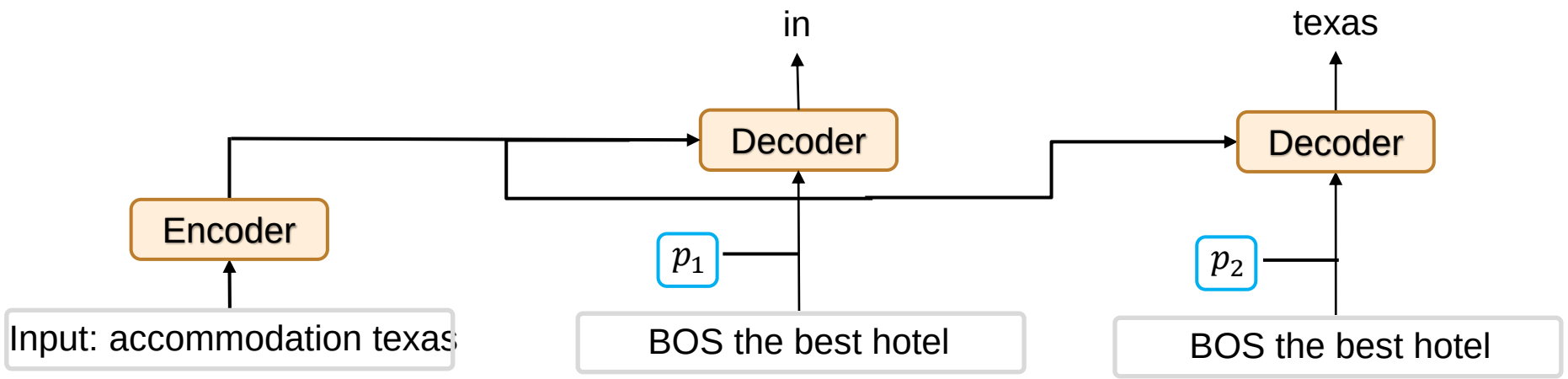


ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *EMNLP-Findings 2020*
Yeyun Gong, Yu Yan*, Weizhen Qi*, Dayiheng Liu*, Nan Duan, Jiusheng Chen, Bruce Zhang, Ming Zhou.*

ProphetNet Training

e.g.
Input: accommodation texas
Output: the best hotel in texas

$$L = \alpha_1 \cdot L_{in} + \alpha_2 \cdot L_{texas}$$

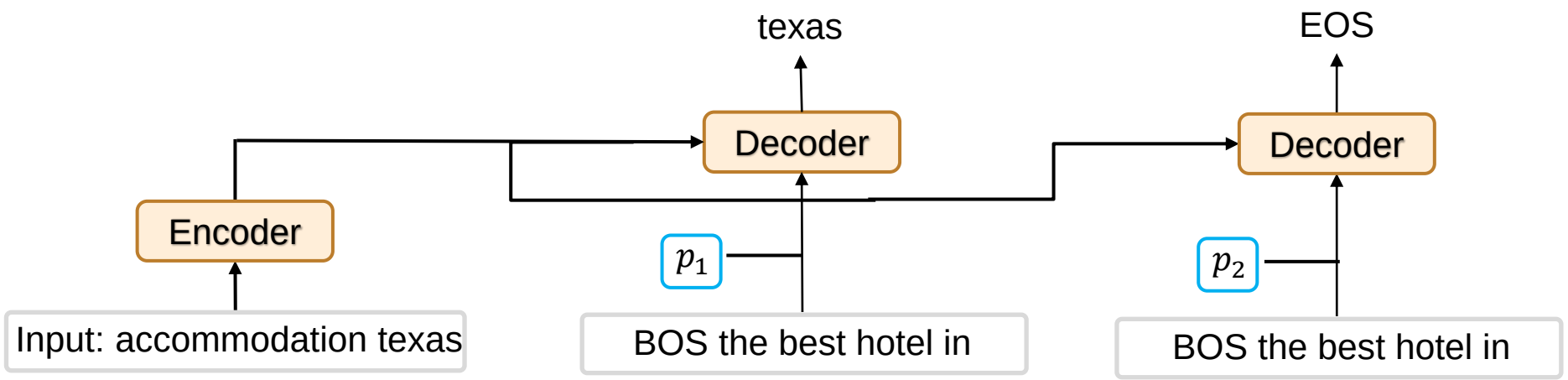


ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *EMNLP-Findings 2020*
Yeyun Gong, Yu Yan*, Weizhen Qi*, Dayiheng Liu*, Nan Duan, Jiusheng Chen, Bruce Zhang, Ming Zhou.*

ProphetNet Training

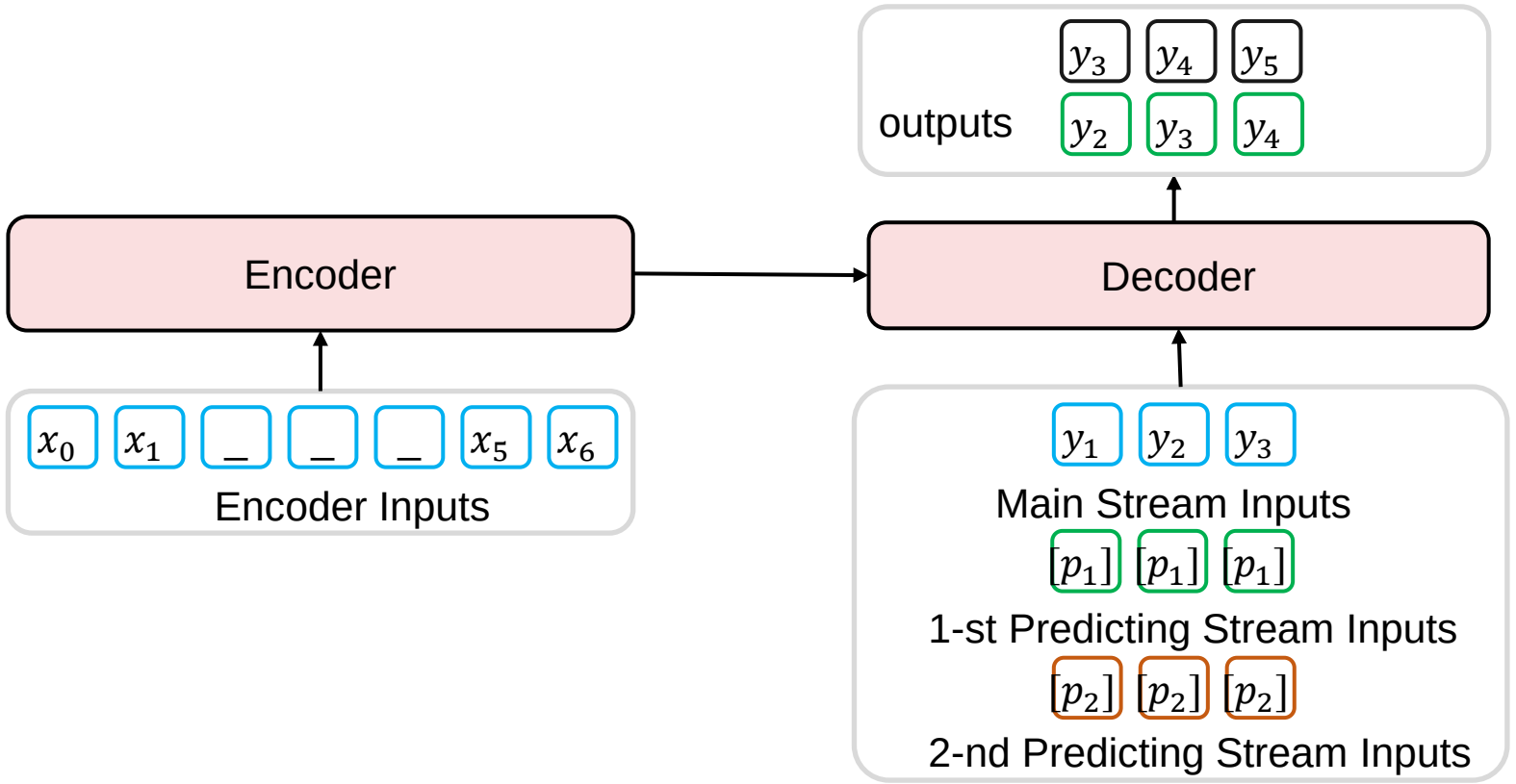
e.g.
Input: accommodation texas
Output: the best hotel in texas

$$L = \alpha_1 \cdot L_{texas} + \alpha_2 \cdot L_{EOS}$$



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

ProphetNet Pre-training



Pre-training Datasets

- **Pre-training data (Chinese Only)**

  We pre-train ProphetNet on the 160GB Chinese raw text. (Xu et al., 2020b)

- **Pre-training data (English Only)**

  Base: we use BookCorpus (Zhu et al., 2015) and English Wikipedia 16GB in total.

  Large: we pre-train ProphetNet on the 160GB English language corpora. (Lewis et al., 2019)

- **Pre-training data (Multi-lingual)**

  We pre-train ProphetNet on the 101GB Wikipedia data, and 2500GB CommonCrawl data, including 100 languages. (Liang et al., 2020)

- **Pre-training data (Dialog)**

  English dialog: using 60 million sessions Reddit open-domain dialog corpus. (Zhou et al., 2018; Galley et al., 2019)

  Chinese dialog: collected Chinese dialog corpus over 30 million sessions. (Wang et al. 2020)

https://github.com/microsoft/ProphetNet

Tasks (Chinese)

MATINF-QA & MATINF-SUMM (Xu et al., 2020)

MATINF-SUMM

婴幼保健 Infant health care	*Class*
宝宝为什么总是吐舌头啊? Why does my baby always stick his tongue out ?	*Question*
我家宝宝出生快满四个月了，这几天我忽然发现宝宝总是吐舌头，而且口水也很多，那么这到底是咋回事啊? My baby is almost four months old. In these few days, I suddenly found that my baby always stick his tongue out and has a lot of saliva. So what is this?	*Description*
正常，不要担心的，小孩子都这个样子。宝宝吐舌头也是很正常的现象，你也不用过于担心，宝宝流口水可能是要长牙齿了。 Don't worry, it's normal. Kids are like this. It is also normal for your baby to stick his tongue out. You don't have to worry too much. Your baby's drooling may be a sign of teeth growing.	*Answer*

MATINF-QA

LCSTS (Hu et al., 2015)

Short Text: 水利部水资源司司长陈明忠今日在新闻发布会上透露，根据刚刚完成的水资源管理制度的考核，有部分省接近了红线的指标，有部分省超过红线的指标。在一些超过红线的地方，将对一些取用水项目进行区域的限批，严格地进行水资源论证和取水许可的批准。

Mingzhong Chen, the Chief Secretary of the Water Devision of the Ministry of Water Resources, revealed today at a press conference, according to the just-completed assessment of water resources management system, some provinces are closed to the red line indicator, some provinces are over the red line indicator. In some places over the red line, It will enforce regional approval restrictions on some water projects, implement strictly water resources assessment and the approval of water licensing.

Summarization: 部分省超过年度用水红线指标 取水项目将被限批

Some provinces exceeds the red line indicator of annual water using, some water project will be. limited approved

Human Score: 5

Experimental Results (Chinese)

	MATINF-QA			MATINF-SUMM			LCSTS		
Method	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
TextRank	--	--	--	35.53	25.78	36.84	24.38	11.97	16.76
LexRank	--	--	--	33.08	23.31	34.96	22.15	10.14	14.65
Seq2Seq	16.62	4.53	10.37	23.05	11.44	19.55	--	--	--
Seq2Seq-Att	19.62	5.87	13.34	43.05	28.03	38.58	33.80	23.10	32.50
WEAN	--	--	--	34.63	22.56	28.92	37.80	25.60	35.20
Global Encoding	--	--	--	49.28	34.14	47.64	39.40	26.90	36.50
BertAbs	--	--	--	57.31	44.05	**55.93**	--	--	--
MTF-S2S-single	20.28	5.94	13.52	43.02	28.05	38.55	33.75	23.20	32.51
MTF-S2S-multi	21.66	**6.58**	14.26	48.59	35.69	43.28	--	--	--
ProphetNet-Zh	**24.18**	6.38	**15.47**	**58.82**	**44.96**	54.26	**42.32**	**27.33**	**37.08**

Summarization Task (English)

Multi-sentence

Original Text : The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online? People are actually buying it. For $ 89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box – enough for 10 to 15 snowballs, he says. [...], a coastal suburb north of Boston. He joked about shipping the stuff to friends and family in warmer states, and an idea was born [...]

→

Summary : A man in suburban Boston is selling snow online to customers in warmer states. For $ 89, he will ship 6 pounds of snow in an insulated Styrofoam box.

Single-sentence

Original Text : the sri lankan government on wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country .

→

Headline: sri lanka closes schools as military escalates

Question Generation Task (English)

Article: architecturally , the school has a catholic character . atop the main building 's gold dome is a golden statue of the virgin mary . immediately in front of the main building and facing it , is a copper statue of christ with arms upraised with the legend `` venite ad me omnes '' . next to the main building is the basilica of the sacred heart . immediately behind the basilica is the grotto , a marian place of prayer and reflection . it is a replica of the grotto at lourdes , france where the virgin mary reputedly appeared to saint bernadette soubirous in 1858 . at the end of the main drive -lrb- and in a direct line that connects through 3 statues and the gold dome -rrb- , is a simple , modern stone statue of mary .

Question Generation Task (English)

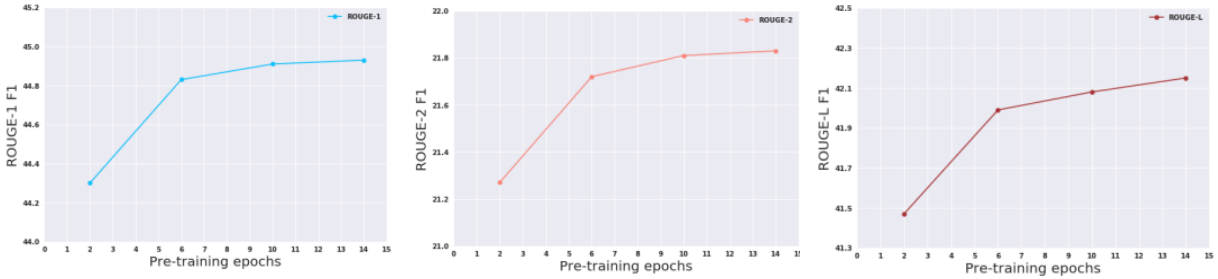
Article: architecturally , the school has a catholic character . atop the main building 's gold dome is a golden statue of the virgin mary . immediately in front of the main building and facing it , is a copper statue of christ with arms upraised with the legend `` venite ad me omnes '' . next to the main building is the basilica of the sacred heart . immediately behind the basilica is the grotto , a marian place of prayer and reflection . it is a replica of the grotto at lourdes , france where the virgin mary reputedly appeared to saint bernadette soubirous in 1858 . at the end of the main drive -lrb- and in a direct line that connects through 3 statues and the gold dome -rrb- , is a simple , modern stone statue of mary .

Question: who did the virgin mary appear to in 1858 ?

Results on CNN/DM ((Nallapati et al., 2016)

Methods	R-1	R-2	R-L
LEAD-3	40.42	17.62	36.67
PTGEN	39.53	17.28	37.98
PTGEN+Coverage	39.53	17.28	36.28
S2S-ELMo	41.56	18.94	38.47
Bottom-Up	41.22	18.68	38.34
BERTSUMABS	41.72	19.39	38.76
BERTSUMEXTABS	42.13	19.60	39.18
MASS	42.12	19.50	39.01
UniLM	43.33	20.21	40.51
ProphetNet	**43.68**	**20.64**	**40.72**

Results of Base-Scale Pre-training



Performance increase on Gigaword dev set as ProphetNet pre-trains for more epochs.

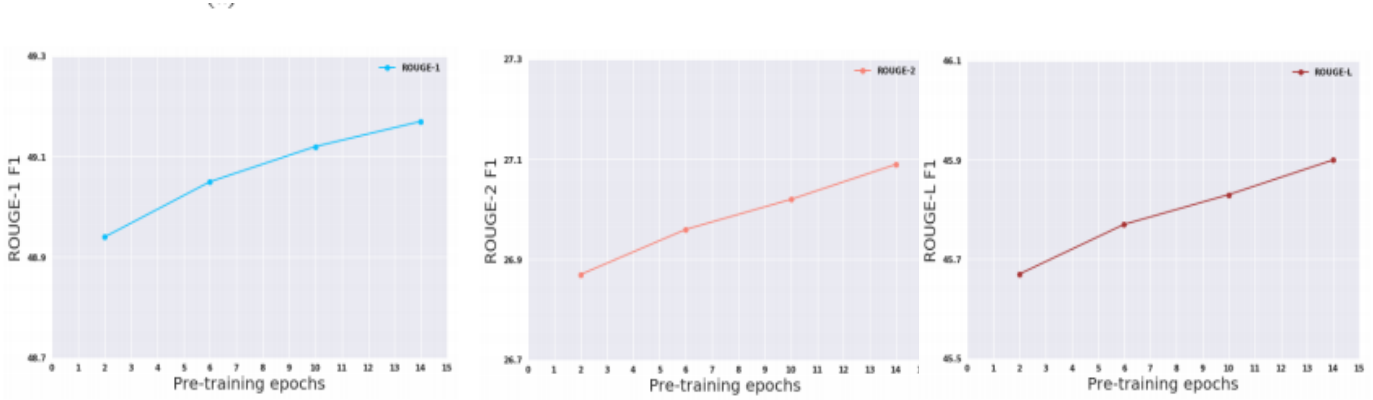
Methods	Data Size	R-1	R-2	R-L
T5	750G	43.52	**21.55**	40.69
PEGASUSLARGE	750G	43.90	21.20	40.76
PEGASUSLARGE	3800G	44.17	21.47	41.11
BART	160G	44.16	21.28	40.90
ProphetNet	160G	**44.20**	21.17	**41.30**

Results of Large-Scale Pre-training

Results on Gigaword (Rush et al., 2015)

Methods	R-1	R-2	R-L
OpenNMT	36.73	17.86	33.68
Re2Sum	37.04	19.03	34.46
MASS	37.66	18.53	34.89
UniLM	38.45	19.45	35.75
ProphetNet	**39.55**	**20.27**	**36.57**

Results of Base-Scale Pre-training



Performance increase on Gigaword dev set as ProphetNet pre-trains for more epochs.

Methods	Data Size	R-1	R-2	R-L
PEGASUSLARGE	750G	38.75	19.96	36.14
PEGASUSLARGE	3800G	39.12	19.86	36.24
ProphetNet	160G	**39.51**	**20.42**	**36.69**

Results of Large-Scale Pre-training

Results on SQuAD 1.1 (Rajpurkar et al., 2016)

- Base Scale Pre-training

Methods	BLEU-4	METOR	R-L
CorefNQG	15.16	19.12	-
SemQG	18.37	22.65	46.68
UniLM	21.63	25.04	51.09
ProphetNet	**23.91**	**26.60**	**52.26**
R-Dev-Test			
MP-GSN	16.38	20.25	44.48
SemQG	20.76	24.2	48.91
UniLM	23.08	25.57	52.03
ProphetNet	**25.80**	**27.54**	**53.65**

Methods	BLEU-4	METOR	R-L
CorefNQG	15.16	19.12	-
SemQG	18.37	22.65	46.68
UniLM	22.12	25.06	51.07
ProphetNet	**25.01**	**26.83**	**52.57**
R-Dev-Test			
MP-GSN	16.38	20.25	44.48
SemQG	20.76	24.2	48.91
UniLM	23.75	25.61	52.04
ProphetNet	**26.72**	**27.64**	**53.79**

Experimental Results (Multi-lingual)

- News Title Generation

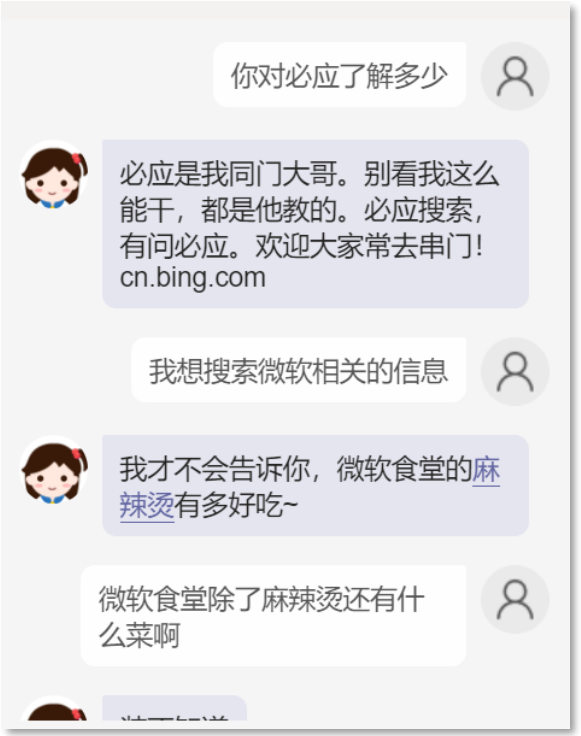
NTG	DE	FR	ES	RU	AVG
DAE-Wiki	6.8	8.7	9	7.7	8.1
xProphetNet-Wiki	7.5	9.9	11.9	8.4	9.4
xProphetNet-CC	8.4	10.9	12	7.7	9.8

- Question Generation

QG	DE	FR	ES	IT	PT	AVG
DAE-Wiki	3	4.2	12.4	15.8	8.3	8.7
xProphetNet-Wiki	3.7	4.9	14.8	17	9.5	10
xProphetNet-CC	4.2	5.7	17.4	18.9	10.7	11.4

Experimental Results (Dialog)

Setting	Win	Lose	Tie	Kappa
Ours-C vs Xiaoice-C	68%	26%	6%	0.73
Ours-C vs Xiaoice-S	76%	24%	0%	0.65
Ours-S vs Xiaoice-S	81%	19%	0%	0.67



Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
AVSD Baseline	0.629	0.485	0.383	0.309	0.215	0.487	0.746
CMU Sinbad's	0.718	0.584	0.478	0.394	0.267	0.563	1.094
PLATO	0.784	0.637	0.525	0.435	0.286	0.596	1.209
ProphetNet-Dialog-En	**0.823**	**0.688**	**0.578**	**0.482**	**0.309**	**0.631**	**1.354**

Experimental Results on DSTC7-AVSD (Alamri et al., 2019)

Ablation Study

- Without Pretraining

Methods	size	R-1	R-2	R-L
T5 Transformer	12L-768	39.19	17.6	36.69
ProphetNet	12L-768	40.66	18.05	37.79

Results on CNN/Daily Mail dev set without Pretraining

- **N**-gram comparison

Setting	R-1	R-2	R-L
$\text{MASS}_{\text{base}}$	42.12	19.50	39.01
$\text{ProphetNet}_{\text{base}}$-1gram	42.21	19.54	39.06
$\text{ProphetNet}_{\text{base}}$-2gram	42.52	19.78	39.59
$\text{ProphetNet}_{\text{base}}$-3gram	**42.61**	**19.83**	**39.67**

N-gram comparison results on CNN/DailyMail test set

Experimental Results on Product

We build our dataset on Bing keywords tables.

- 260 million keywords searching space
- 1 million training query-keyword pair
- 10k testing query-keyword pair

Model	R@5	R@10	R@15	R@20	MAP@5	MAP@10	MAP@15	MAP@20
BM25	27.86	33.40	37.30	39.13	0.2051	0.2125	0.2156	0.2166
LSTM	62.47	71.81	75.63	77.76	0.5716	0.6267	0.6442	0.6534
Bi-LSTM	63.28	72.28	76.21	78.13	0.5770	0.6292	0.6479	0.6563
Bi-LSTM+Copy	67.37	76.12	79.40	83.37	0.6114	0.6616	0.6755	0.6811
Uni-gram ProphetNet	75.00	82.50	84.90	86.50	0.6929	0.7362	0.7461	0.7526
Tri-gram ProphetNet	75.48	83.08	85.45	86.68	0.6974	0.7426	0.7518	0.7565
ProphetNet-Ads	78.05	84.28	86.24	87.54	0.7133	0.7472	0.7542	0.7580
Merged Tri+Tri-Ads	81.34	86.83	88.45	89.39	/	/	/	/
Merged Above	86.56	90.11	91.34	92.15	/	/	/	/

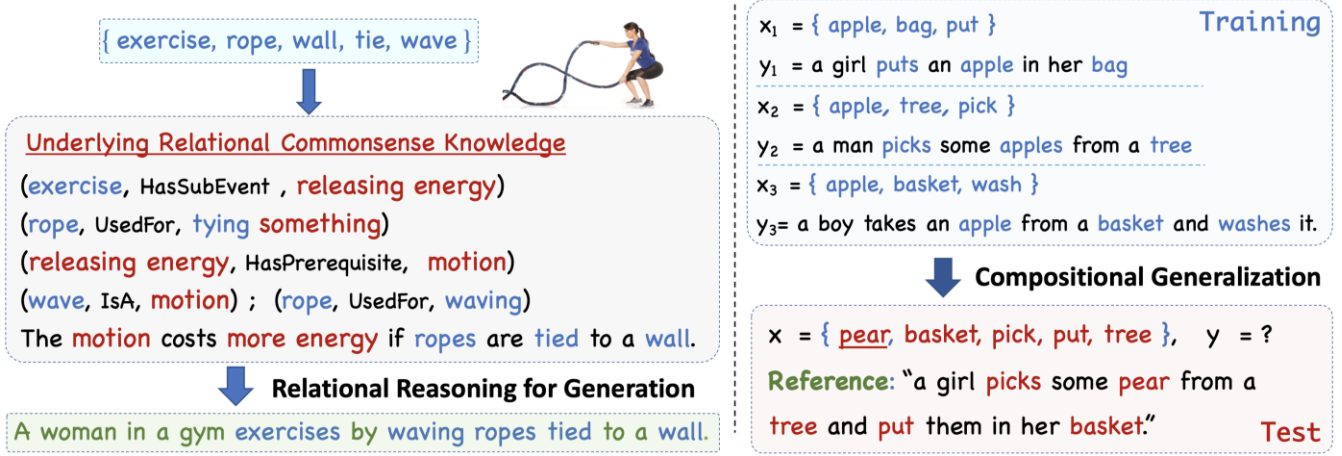
Experimental results

An Enhanced Knowledge Injection Model for Commonsense Generation. *COLING 2020*.

CommonGen (Lin et al., 2019) Task

- Task

- Key challenges



- Data
  - Concepts from captions (Flickr, etc)
  - Crowd sourcing for dev and test sets
- Metrics
  - BLEU, CIDEr, SPICE

Motivation

- External knowledge related to the scene of given concepts
  - The prototype would introduce scenario knowledge to find out reasonable combination of concepts.
  - Prototypes would provide the missing key concepts to complete a coherent scenario.

Concepts	front, guitar, microphone, sit	ear, feel, pain, pierce
Prototype	A singer performed the song standing in front of the **audiences** while **playing guitar**.	He expresses severe pain as he tries to **pierce his hand**.
NLG + Prototype	A singer sitting in front of the audiences while playing guitar	He expresses severe pain as he pierce his ear

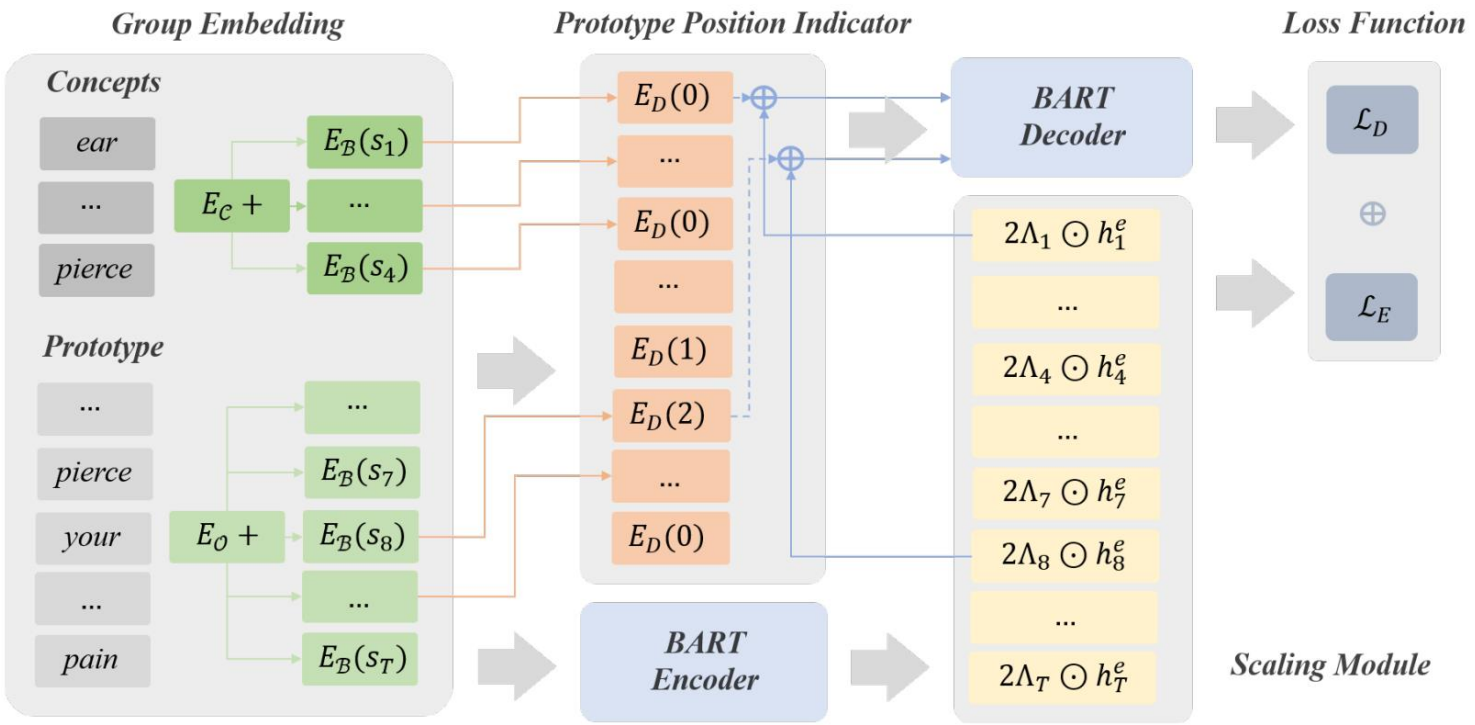
Enhanced Knowledge Injection

- How to better utilize the prototype with the framework of pretrained model?
- Enhanced Knowledge Injection
  - Tokens in prototype make various contribution in the generation.
  - Scaling Module
    - Discriminate important tokens.
    - Assign weights to tokens in the prototype.
  - Prototype Position Indicator
    - Tokens closer to the concept words in prototype are more important for scene description.
    - Mark the relative position of different tokens in the prototype

Problem Setup

- Concepts: $\mathcal{C} = (c_1, \ldots, c_{n_c})$
- Prototype: $\mathcal{Q} = (o_1, \ldots, o_{n_o})$
- Target: $\mathcal{T} = (t_1, \ldots, t_{n_t})$
- Build a generation model $G_\theta : \mathcal{C} \times \mathcal{Q} \rightarrow \mathcal{T}$

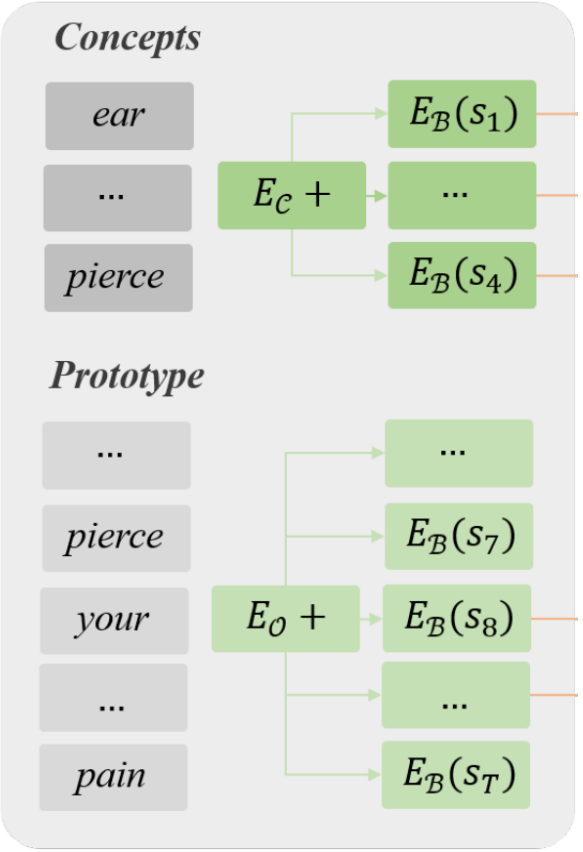
Overall Framework



- Group Embedding(GE)
- Encoder with Scaling Module(SM)
- Decoder with Prototype Position Indicator(PPI)

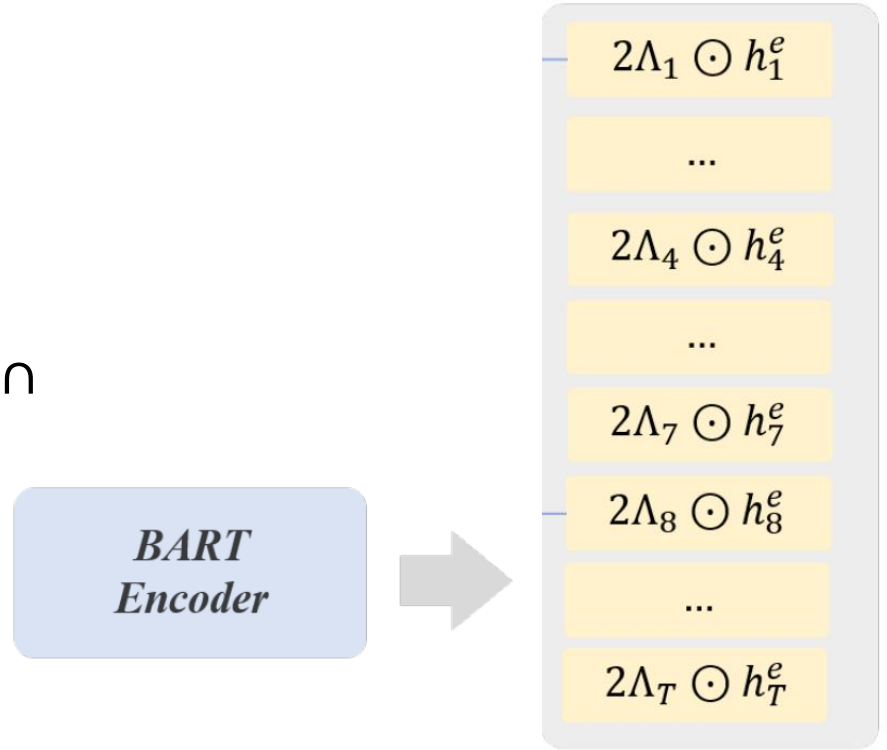
Group Embedding

- Model input: Concept + Prototype
  - $\mathcal{S} = [\mathcal{C}, \mathcal{O}] = [c_1, \ldots, c_{n_c}, o_1, \ldots, o_{n_o}]$
- Add group embedding on top of the original BART embedding function.
  - $E(c_j) = E_B(c_j) + E_{\mathcal{C}}$
  - $E(o_k) = E_B(o_k) + E_{\mathcal{O}}$
  - E_B is the original embedding function of BART, $E_{\mathcal{C}}$ and $E_{\mathcal{O}}$ are two group embeddings for $\mathcal{C}$ and $\mathcal{O}$.



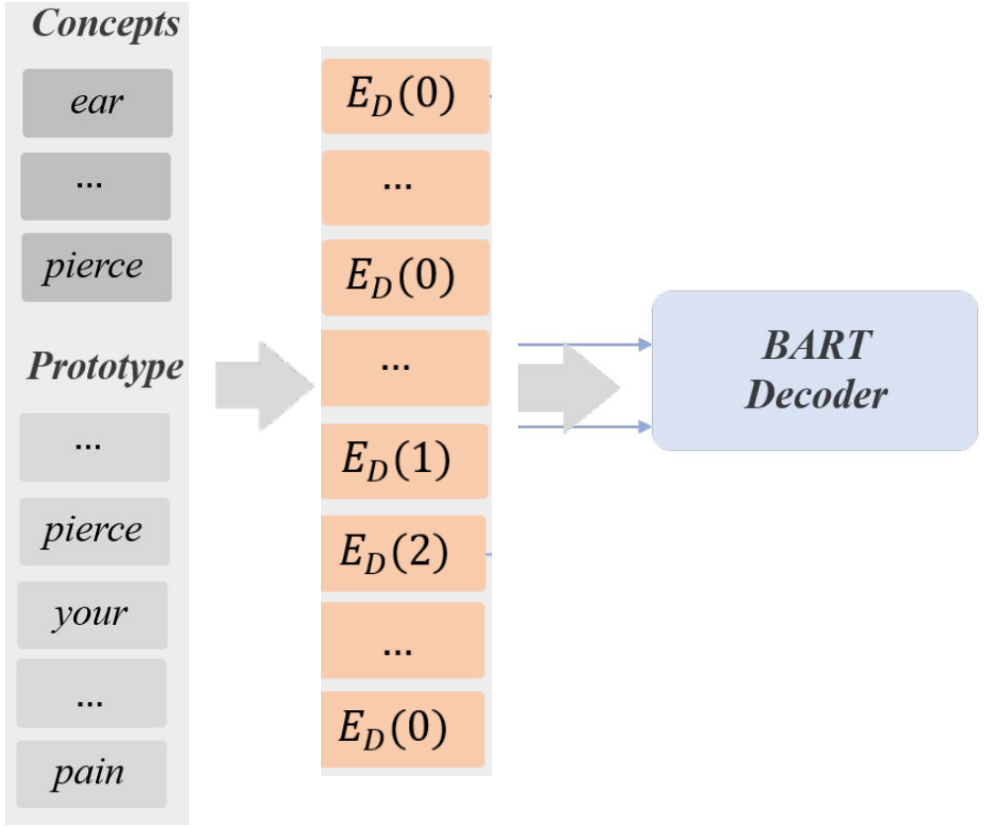
Scaling Module

- Injecting the co-occurrence between the input $\mathcal{S}$ and the target $\mathcal{T}$ into the scaling module.
- Cross entropy for scaling module.
- $\mathcal{L}_E = -\sum_{s_v \in \mathcal{S}} \left(\mathcal{I}_{\{s_v \in \mathcal{T}\}} log M(\Lambda_v) + \mathcal{I}_{\{s_v \in \mathcal{T}\}} log(1 - M(\Lambda_v)) \right)$
- where M is the mean function, assign positive label for $\mathcal{S} \cap \mathcal{T}$ and negative label for others.



Prototype Position Indicator

- Surrounding tokens of concepts in prototype tend to describe how these concept words interact with the scenario.
  - Pick up the positions of concept tokens in prototype as multiple position centers.
  - Compute the smallest distance from token itself to those concept tokens in prototype.



Prototype Collection

- In-Domain Corpus
  - CommonGen is to describe a common scenario in our daily life.
  - Datasets of image captioning or video captioning are related.
  - We utilize VaTeX, SNLI, Activity and the training set of CommonGen as the external plain text knowledge datasets.
- Out-of-Domain Corpus
  - Employ Wikipedia as our external knowledge dataset.
- The number of retrieved prototypes whose concepts co-occur in GT sentences across $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$.

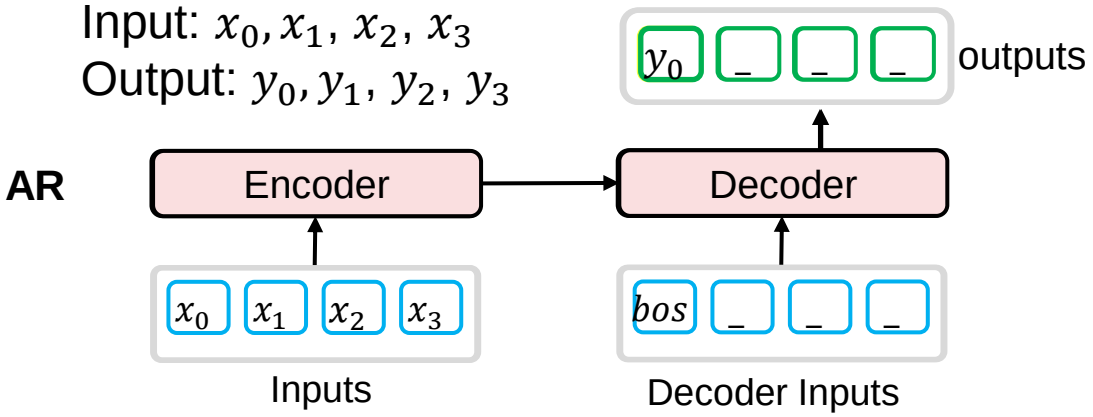
	1	2	3	4	5
$\mathcal{D}_{in}$	2,179	17,664	16,356	2,538	332
$\mathcal{D}_{out}$	3,009	21,441	12,278	2,069	272

Experimental Results

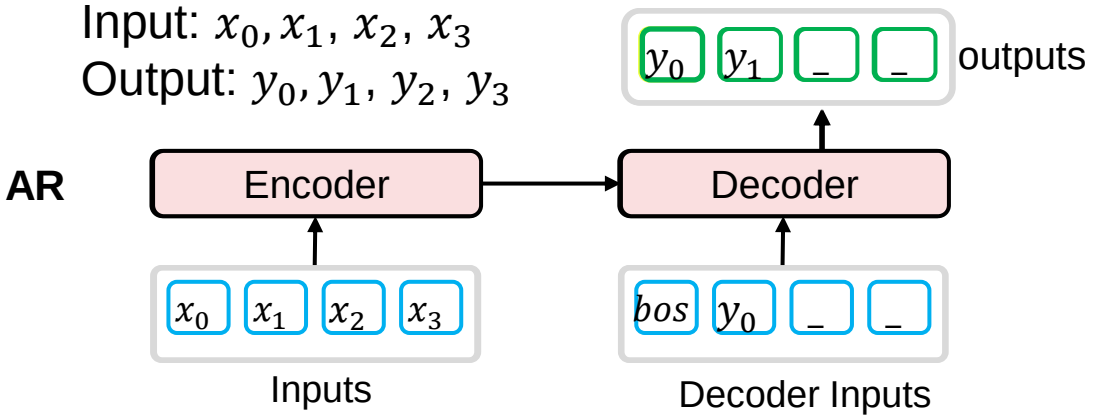
	Model	BLEU-4	METEOR	CIDEr	SPICE
Models without pre-training	$bRNN\text{-}CopyNet$	2.00	12.70	3.99	10.60
	$LevenTrans$	4.00	13.30	3.72	14.00
Pretrained Language Generation model	$GPT\text{-}2$	19.40	24.40	11.06	24.50
	$UniLM$	27.50	29.40	14.92	29.90
	$T5$	27.20	30.00	14.58	30.60
	$BART$	24.90	30.50	13.32	30.10
Models with $\mathcal{D}_{out}$	$Retrieve_{\mathcal{D}_{out}}$	7.50	18.40	4.95	15.00
	$BART_{\mathcal{D}_{out}}$	30.30	31.50	15.82	31.80
	$EKI\text{-}BART_{\mathcal{D}_{out}}$	32.10	32.00	16.80	32.50
Models with $\mathcal{D}_{in}$	$Retrieve_{\mathcal{D}_{in}}$	26.40	29.90	12.91	27.90
	$BART_{\mathcal{D}_{in}}$	32.40	32.30	16.43	32.70
	$EKI\text{-}BART_{\mathcal{D}_{in}}$	**36.10**	**33.80**	**17.80**	**33.40**

BANG: Bridging Autoregressive and Non-autoregressive Generation with Large Scale Pre-training. *ICML 2021*

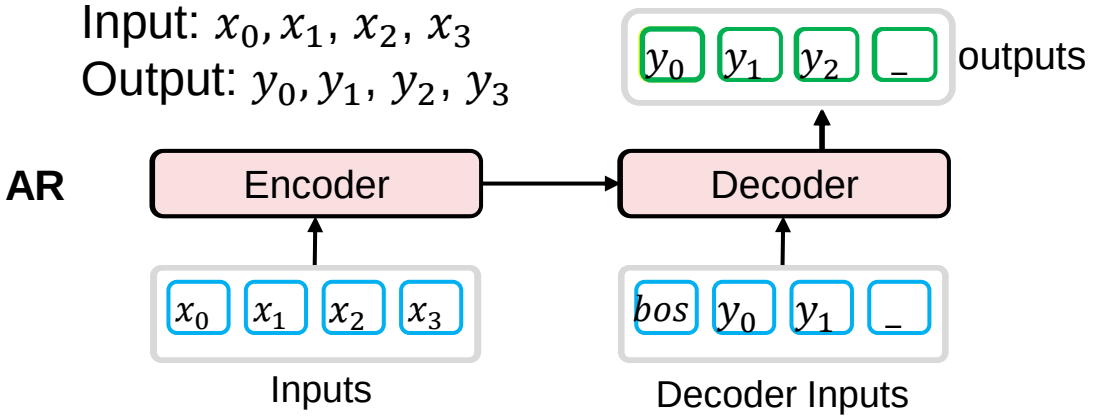
Architecture of BANG



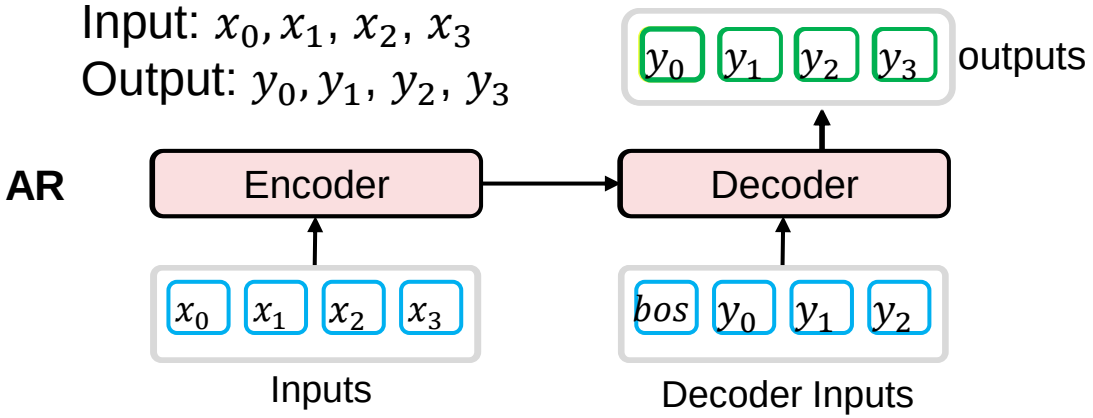
Architecture of BANG



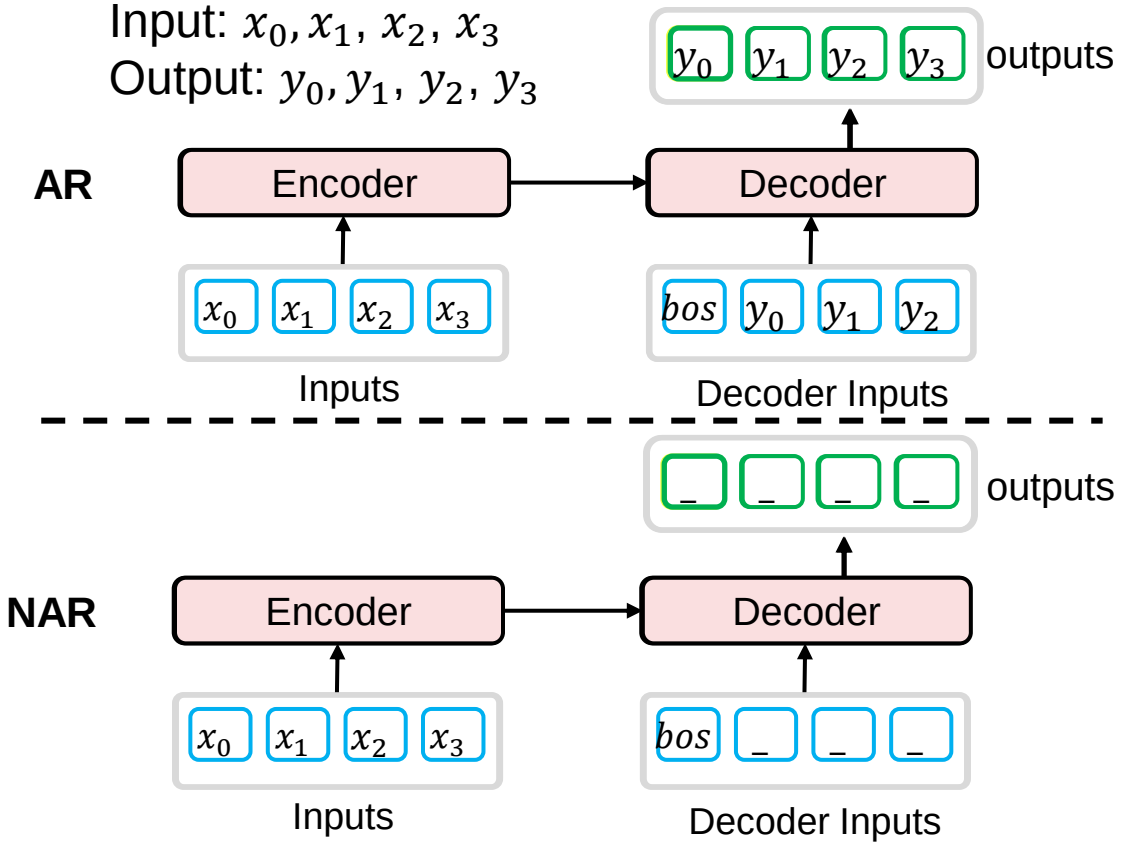
Architecture of BANG



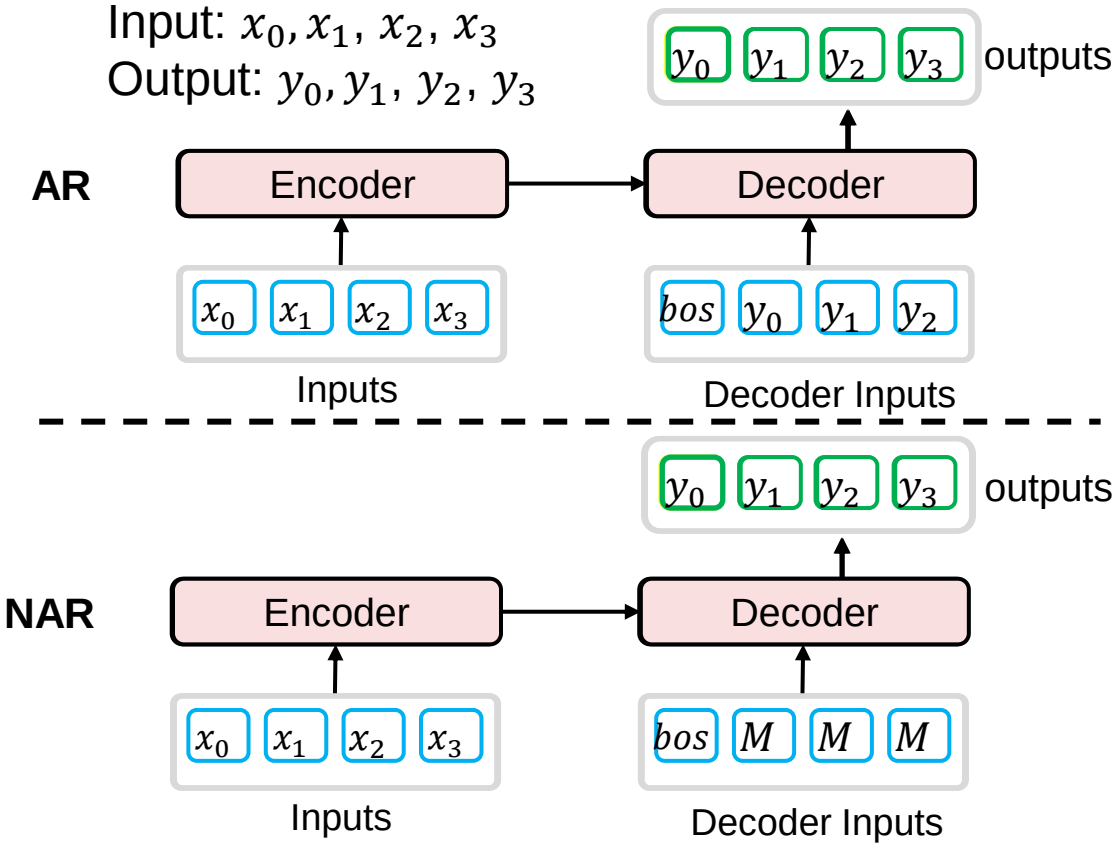
Architecture of BANG



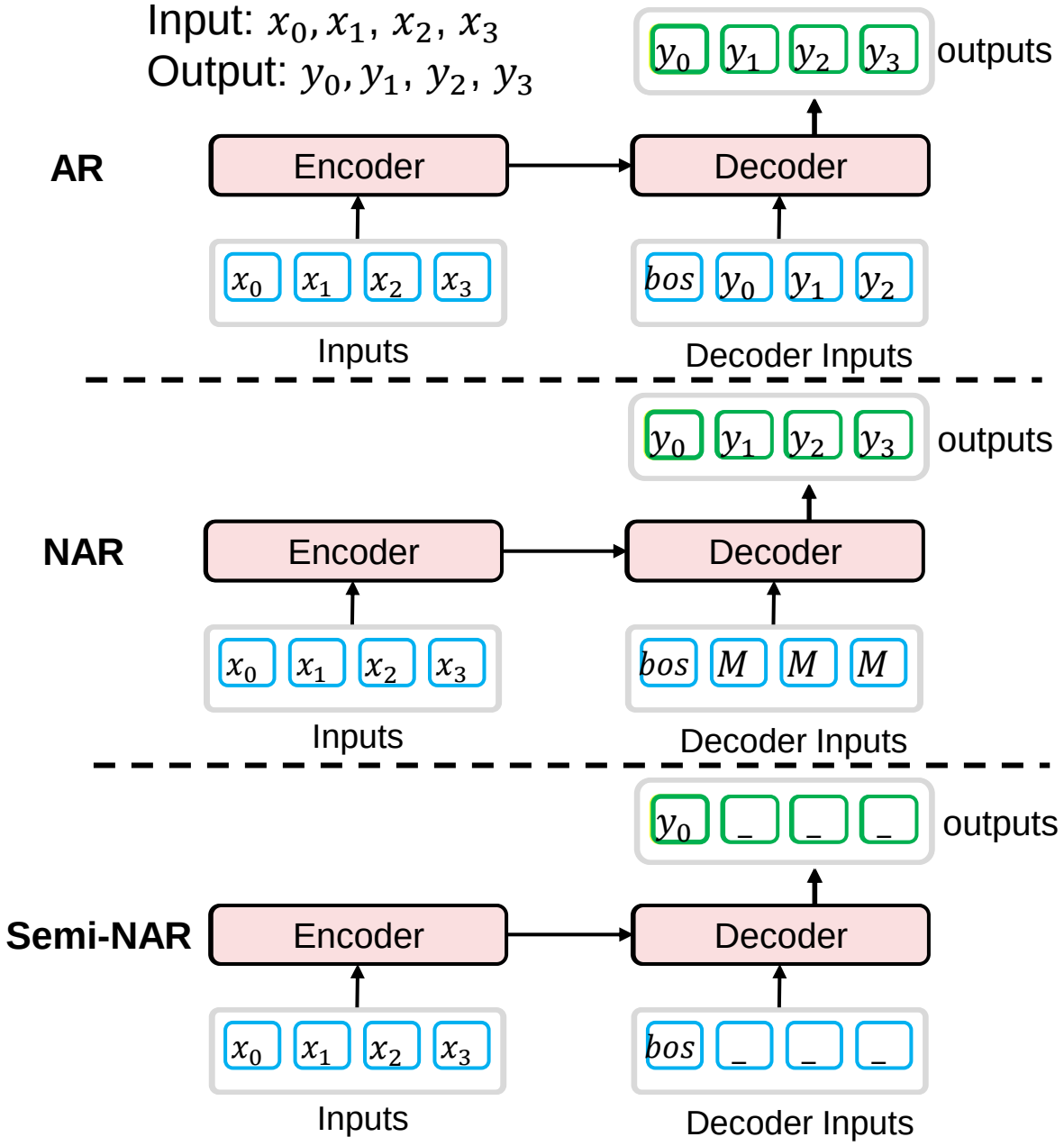
Architecture of BANG



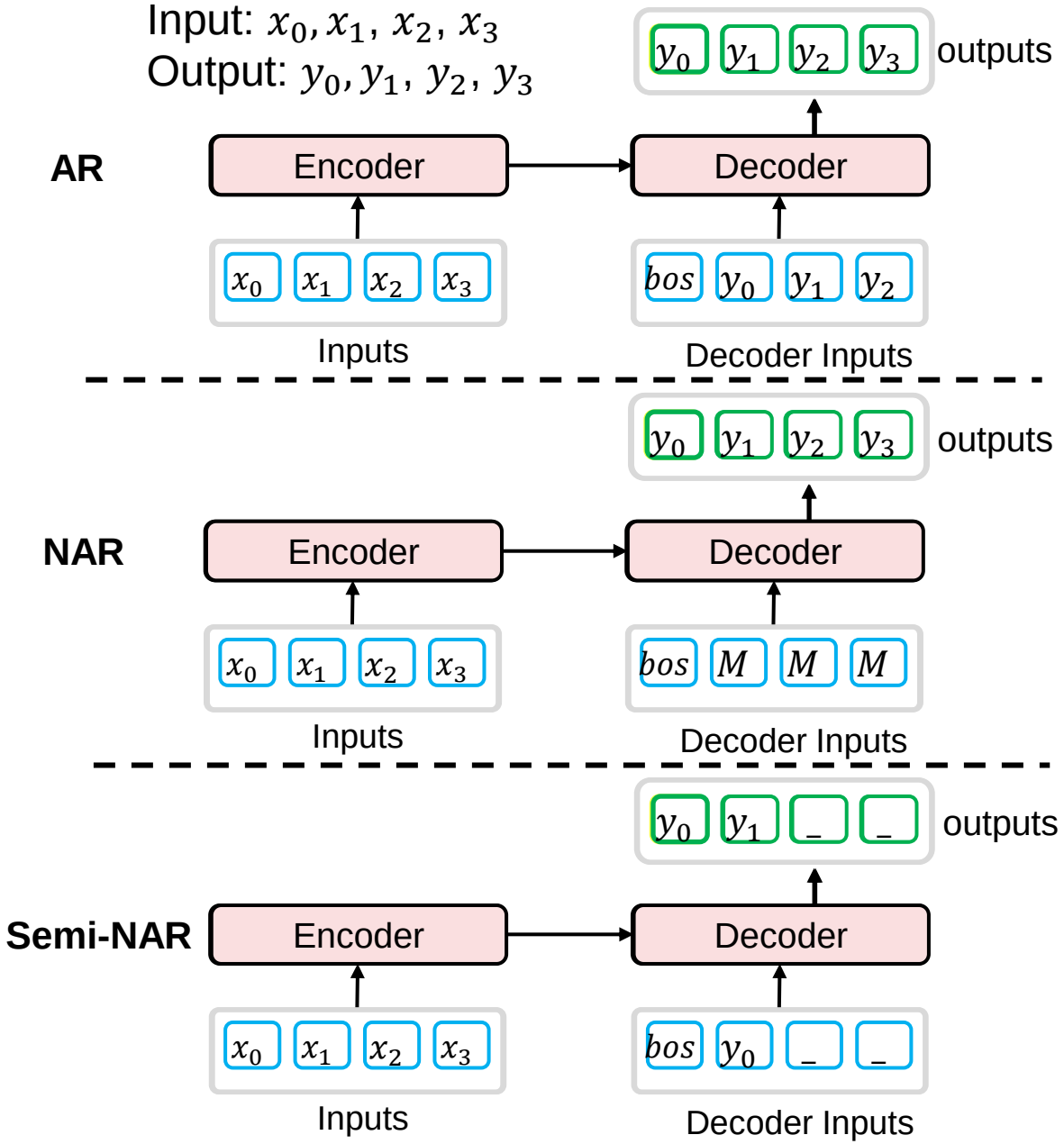
Architecture of BANG



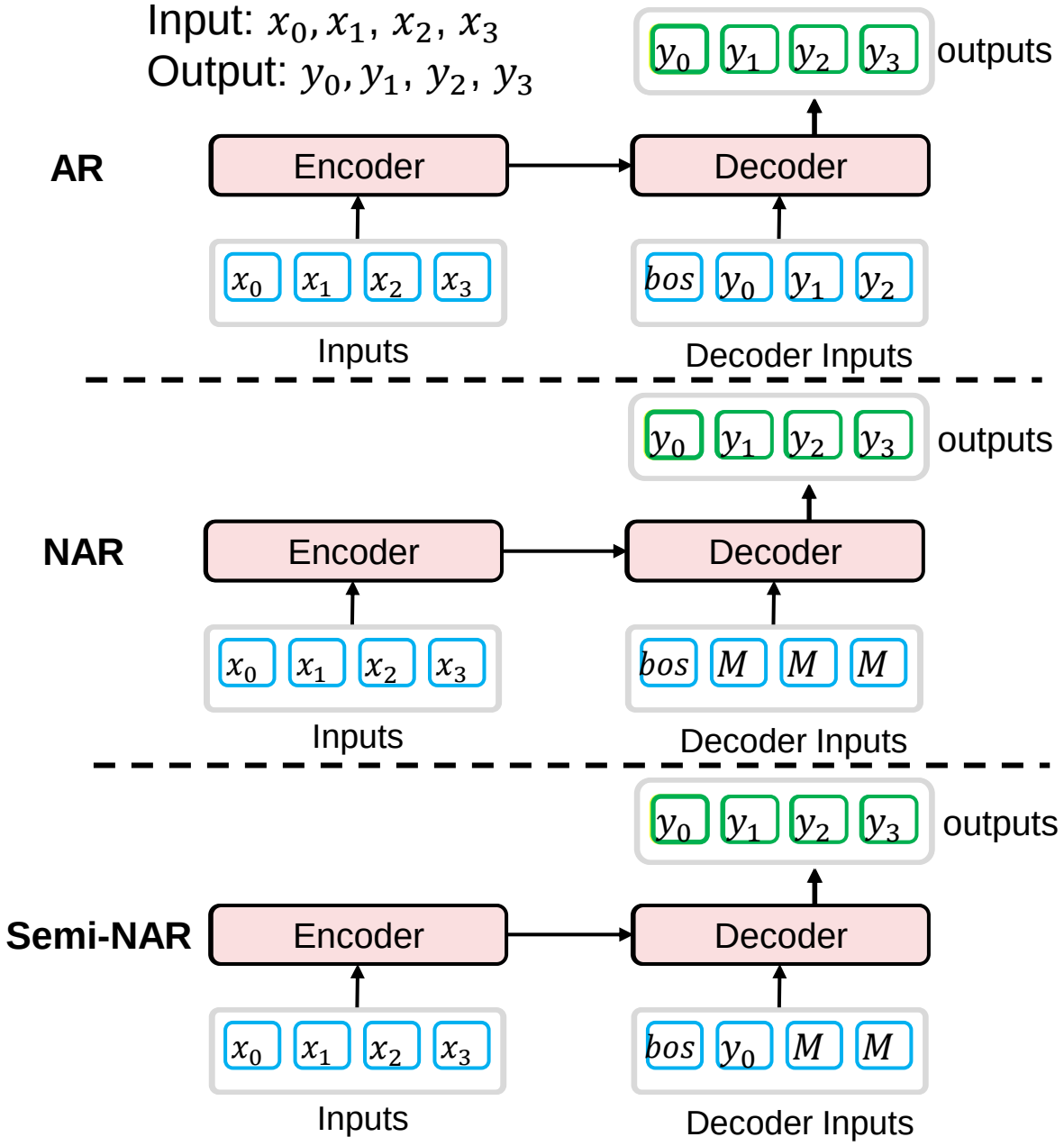
Architecture of BANG



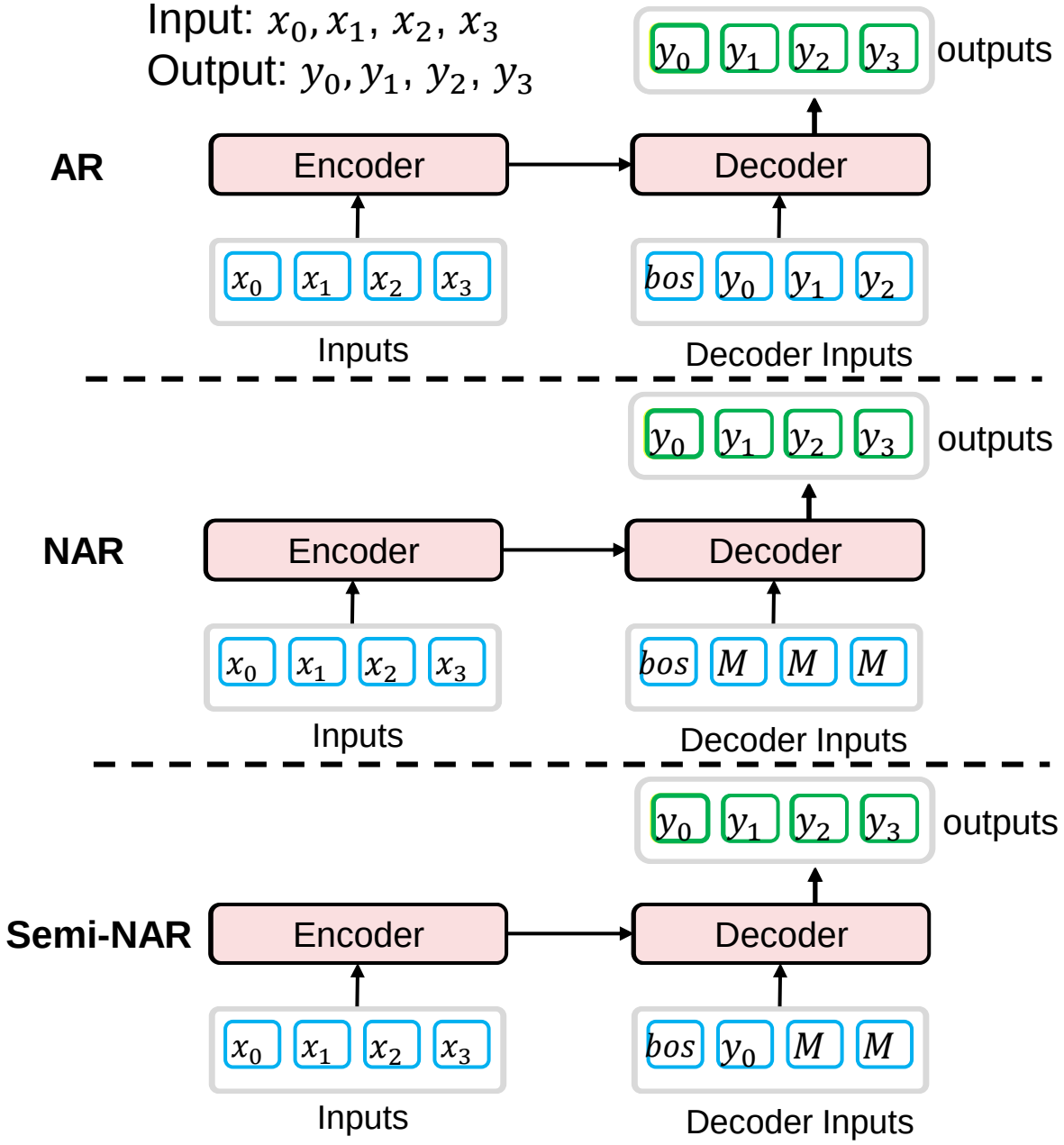
Architecture of BANG

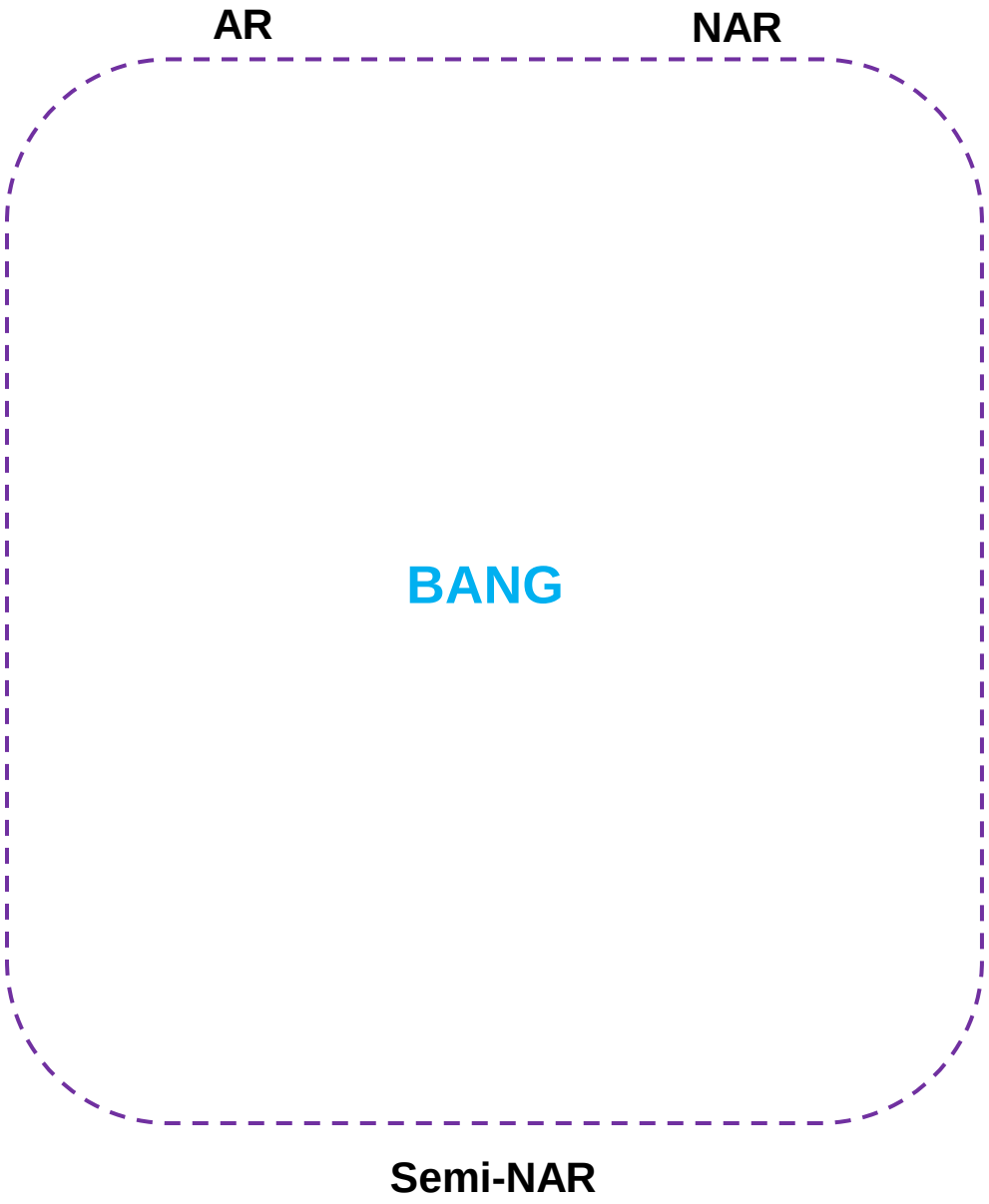


Architecture of BANG

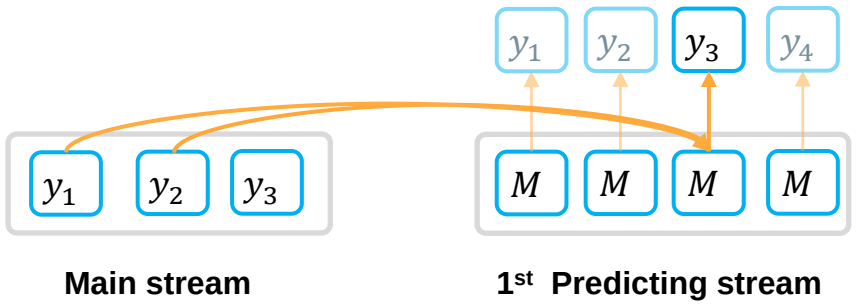


Architecture of BANG

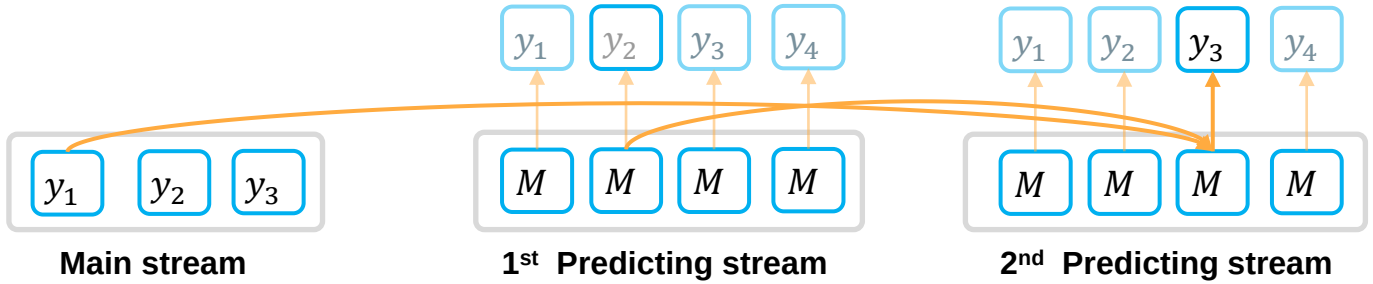




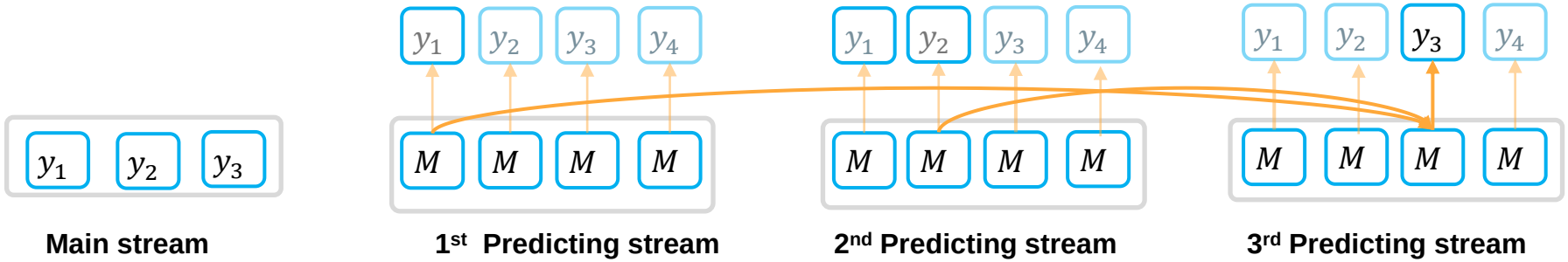
Cross-Visible N-Stream Self-Attention



AR



Semi-NAR



NAR

Experiments

- **Pretraining data**

  16GB English language corpora.

- **Downstream tasks**
  - SQuAD for Question Generation
  - XSum for Abstractive Summarization
  - PersonaChat for Response Generation

Results of Question Generation

Pattern	Methods	SQuAD 1.1				Latency
		ROUGE-L	BLEU-4	METEOR	OVERALL	ms/Sample
Semi-NAR	InsT (Stern et al., 2019)	29.98	2.34	8.15	13.49 (+0.00)	67.61 (4.3x)
	iNAT (Lee et al., 2018)	32.34	3.16	9.18	14.89 (+1.40)	31.59 (2.0x)
	CMLM (Ghazvininejad et al., 2019)	29.60	3.89	9.70	14.40 (+0.91)	106.84 (6.8x)
	LevT (Gu et al., 2019)	30.81	2.68	9.40	14.30 (+0.81)	116.41 (7.4x)
	BANG	**47.39**	**17.62**	**21.69**	**28.90 (+15.41)**	111.11 (7.1x)
NAR	NAT (Gu et al., 2017)	31.51	2.46	8.86	14.28 (+0.02)	17.11 (1.1x)
	iNAT (Lee et al., 2018)	32.44	2.33	8.84	14.54 (+0.28)	16.52 (1.1x)
	CMLM (Ghazvininejad et al., 2019)	31.58	2.51	8.85	14.31 (+0.05)	16.41 (1.0x)
	LevT (Gu et al., 2019)	31.38	2.27	9.14	14.26 (+0.00)	27.52 (1.8x)
	BANG	**44.07**	**12.75**	**18.99**	**25.27 (+11.01)**	**15.69 (1.0x)**
AR	Transformer (Vaswani et al., 2017)	29.43	4.61	9.86	14.63(+0.00)	159.49(10.2x)
	MASS (Song et al., 2019)	**49.48**	20.16	**24.41**	31.35 (+16.72)	N/A
	BART (Lewis et al., 2019)	42.55	17.08	23.19	27.61 (+12.98)	N/A
	ProphetNet (Qi et al., 2020)	48.00	19.58	23.94	30.51 (+15.88)	N/A
	BANG	49.32	**21.40**	24.25	**31.66 (+17.03)**	N/A

Results of Abstractive Summarization

Pattern	Methods	XSum				Latency
		ROUGE-1	ROUGE-2	ROUGE-L	OVERALL	ms/Sample
Semi-NAR	InsT (Stern et al., 2019)	17.65	5.18	16.05	12.96 (+0.00)	63.37 (4.0x)
	iNAT (Lee et al., 2018)	26.95	6.88	22.43	18.75 (+5.79)	31.27 (2.0x)
	CMLM (Ghazvininejad et al., 2019)	29.12	7.70	23.04	19.95 (+6.99)	113.64 (7.1x)
	LevT (Gu et al., 2019)	25.33	7.40	21.48	18.07 (+5.11)	101.01 (6.3x)
	BANG	**34.71**	**11.71**	**29.16**	**25.19 (+12.23)**	109.77 (6.9x)
NAR	NAT (Gu et al., 2017)	24.04	3.88	20.32	16.08 (+0.22)	17.47 (1.1x)
	iNAT (Lee et al., 2018)	24.02	3.99	20.36	16.12 (+0.26)	16.94 (1.1x)
	CMLM (Ghazvininejad et al., 2019)	23.82	3.60	20.15	15.86 (+0.00)	16.88 (1.1x)
	LevT (Gu et al., 2019)	24.75	4.18	20.87	16.60 (+0.74)	27.72 (1.7x)
	BANG	**32.59**	**8.98**	**27.41**	**22.99 (+7.13)**	**15.97 (1.0x)**
AR	Transformer (Vaswani et al., 2017)	30.66	10.80	24.48	21.98(+0.00)	262.47(16.4x)
	MASS (Song et al., 2019)	39.70	17.24	31.91	29.62 (+7.64)	N/A
	BART (Lewis et al., 2019)	38.79	16.16	30.61	28.52 (+6.54)	N/A
	ProphetNet (Qi et al., 2020)	39.89	17.12	32.07	29.69 (+7.71)	N/A
	BANG	**41.09**	**18.37**	**33.22**	**30.89 (+8.91)**	N/A

Results of Response Generation

Pattern	Methods	PersonaChat					Latency
		BLEU-1	BLEU-2	D-1	D-2	OVERALL	ms/Sample
Semi-NAR	InsT (Stern et al., 2019)	12.63	9.43	0.1	0.3	5.62 (+0.00)	65.27 (4.4x)
	iNAT (Lee et al., 2018)	41.17	32.13	0.1	1.1	18.63 (+13.01)	43.25 (2.9x)
	CMLM (Ghazvininejad et al., 2019)	**44.38**	**35.18**	0.1	0.8	20.12 (+14.50)	105.82 (7.1x)
	LevT (Gu et al., 2019)	24.89	18.94	0.1	0.6	11.13 (+5.51)	80.26 (5.4x)
	BANG	39.82	30.72	**1.9**	**14.2**	**21.66 (+16.04)**	109.17 (7.3x)
NAR	NAT (Gu et al., 2017)	31.53	24.17	0.1	0.8	14.15 (+2.20)	17.86 (1.2x)
	iNAT (Lee et al., 2018)	30.56	23.38	0.1	0.7	13.69 (+1.74)	16.40 (1.1x)
	CMLM (Ghazvininejad et al., 2019)	**31.44**	**24.06**	0.1	0.6	14.05 (+2.10)	16.26 (1.1x)
	LevT (Gu et al., 2019)	26.92	20.47	0.0	0.4	11.95 (+0.00)	27.56 (1.9x)
	BANG	31.11	23.90	**2.5**	**22.7**	**20.05 (+8.10)**	**14.89 (1.0x)**
AR	Transformer (Vaswani et al., 2017)	41.56	32.95	0.3	0.8	18.90(+0.00)	138.31(9.3x)
	MASS (Song et al., 2019)	41.06	35.75	1.4	6.9	21.28 (+2.38)	N/A
	BART (Lewis et al., 2019)	**47.60**	**39.36**	1.1	6.1	**23.54(+4.64)**	N/A
	ProphetNet (Qi et al., 2020)	46.00	38.40	1.3	7.3	23.25 (+4.35)	N/A
	BANG	45.77	35.54	**1.4**	**8.4**	22.78 (+3.88)	N/A

Poolingformer: Long Document Modeling with Pooling Attention. *ICML 2021*

NLG Scenario with Long Input

Long input

News Article

As in years past, a lot of the food trends of the year were based on creating perfectly photogenic dishes. An aesthetically pleasing dish, however, doesn't mean it will stand the test of time. In fact, it's not uncommon for food trends to be all the hype one year and die out the next. From broccoli coffee to "bowl food," here are 10 food trends that you likely won't see in 2019.

...[15 sentences with 307 words are abbreviated from here.]

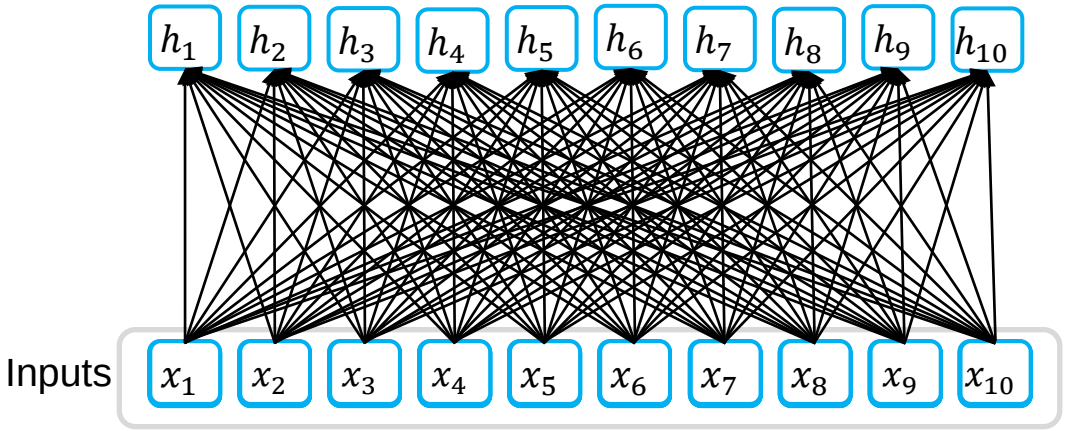
In 2018, restaurants all over the US decided it was a good idea to place gold foil on everything from ice cream to chicken wings to pizza resulting in an expensive food trend. For example, the Ainsworth in New York City sells $1,000 worth of gold covered chicken wings. It seems everyone can agree that this is a food trend that might soon disappear.

Short output

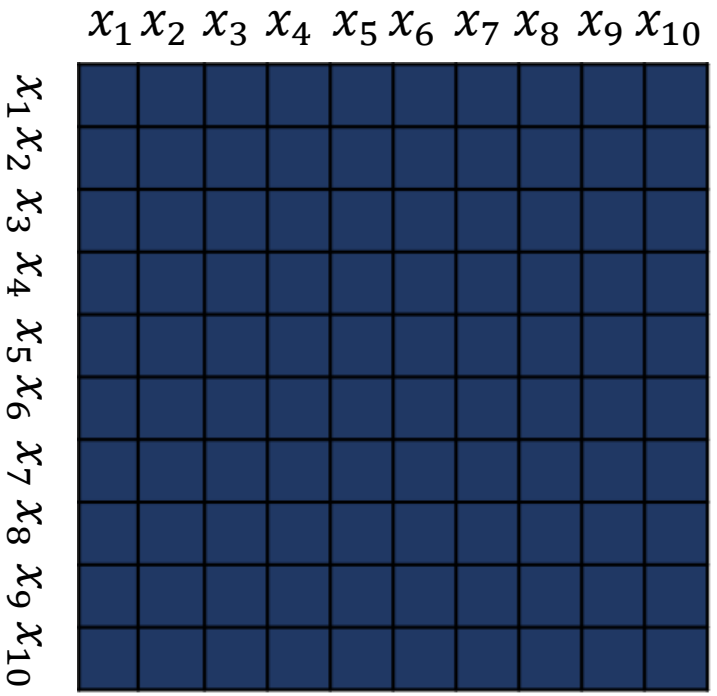
News Headline

10 food trends that you likely won't see in 2019

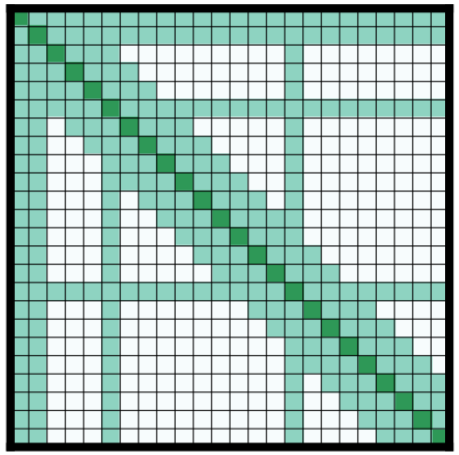
Transformer Encoder



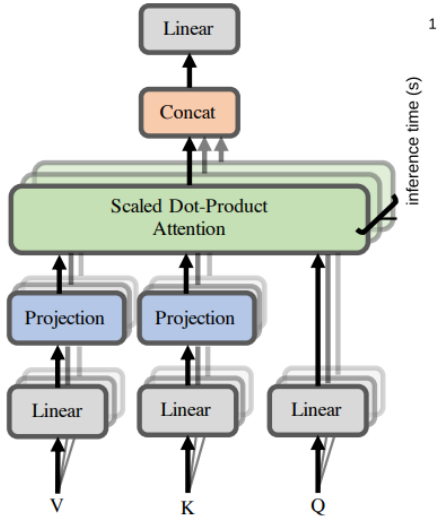
$$\mathcal{O}(n^2)$$

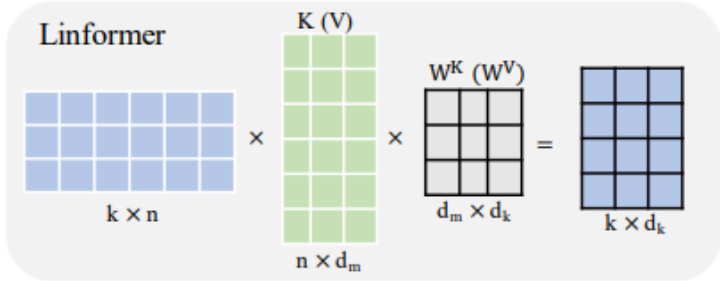


Longformer (Beltagy et al., 2020) and Linformer (Wang et al., 2020)



Longformer Global Token





Linformer

Global vector in Longformer :

- For downstream tasks, set some tokens as global tokens,

while other tokens obtain further (global) information through global Token

Global vector in Linformer:

- Pooling a long sequence to a set of global vectors, which means global information of whole sentence.

All tokens attention to the global vectors.

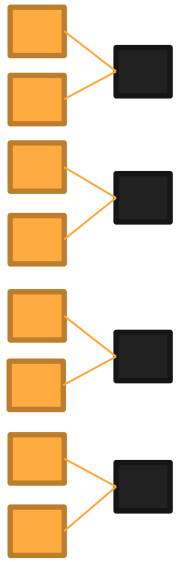
Motivation of Poolingformer

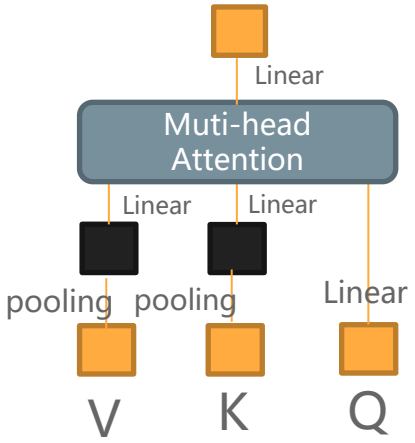
Motivation:

This global information confuses positional information.
All tokens should pay different attention to global information in different locations

Method:

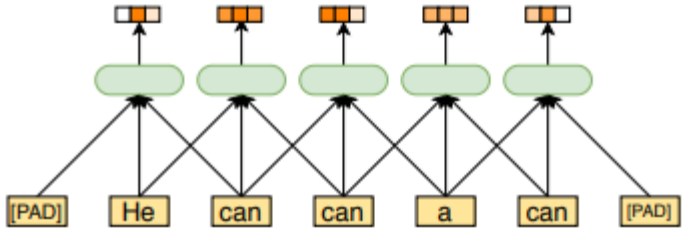
Pooling different blocks of a sequence into global vectors.
Each vector represents the features of different blocks, and other tokens can freely assign attention weights to different blocks.





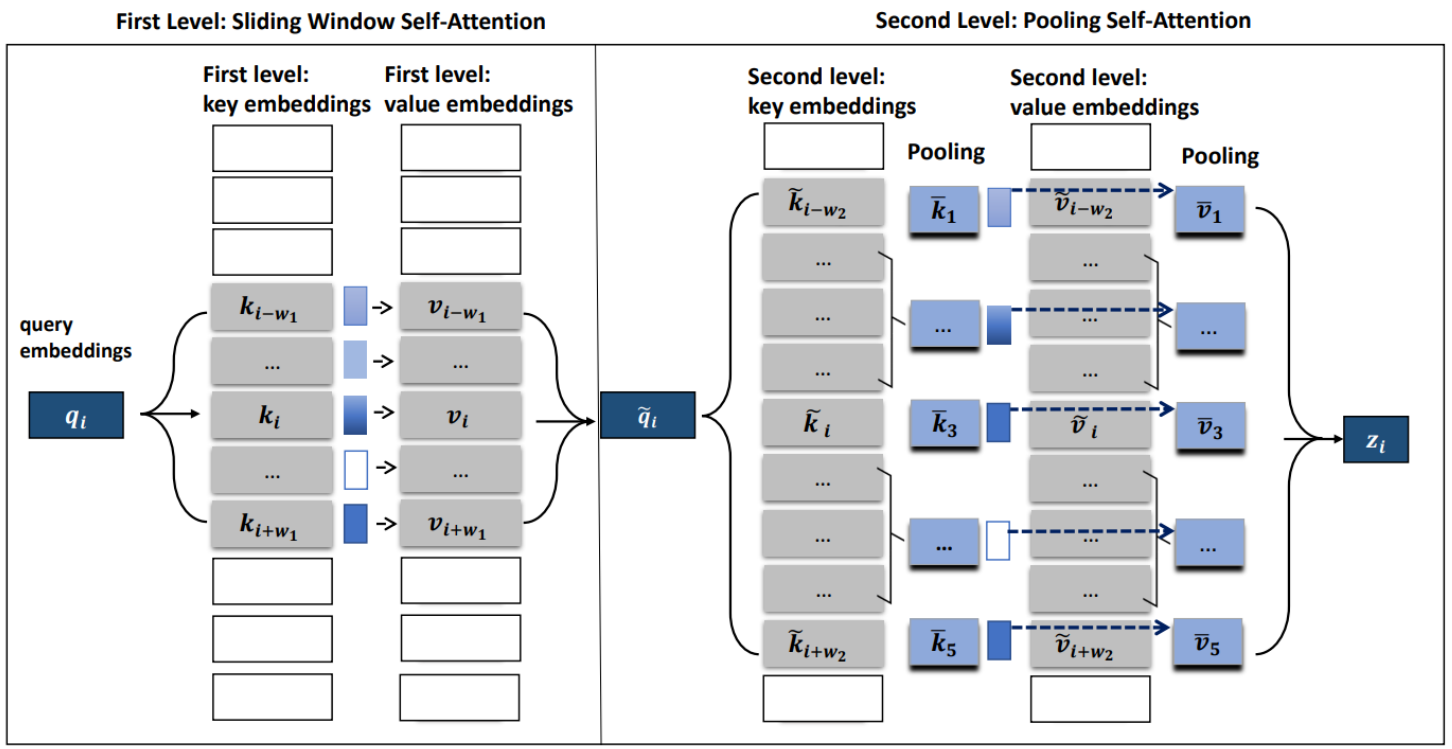
Get Global Vectors

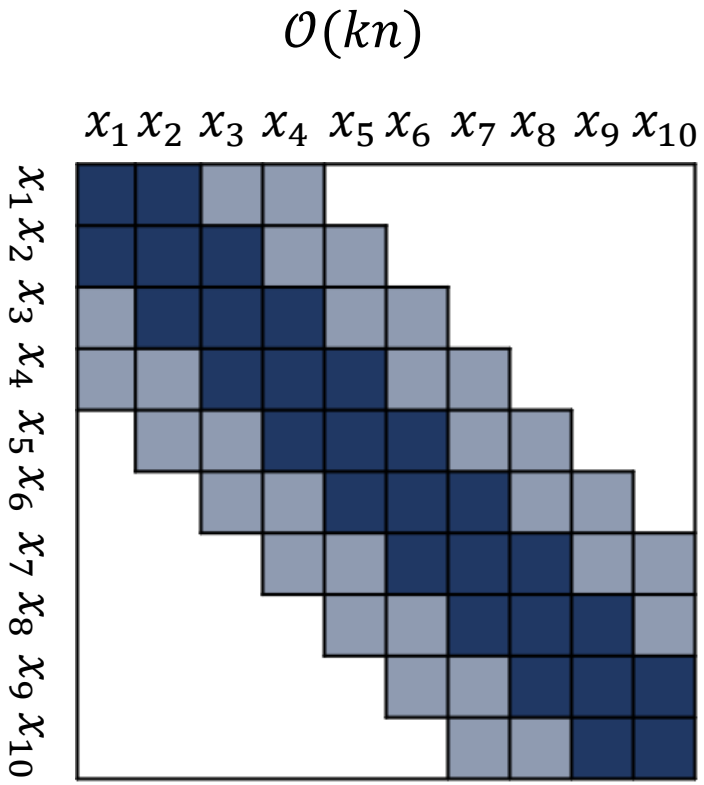
Locally dynamic lightweight convolution way. (Jiang et al., 2020)



(Dynamically determine the weight of token in the window)

Poolingformer Encoder



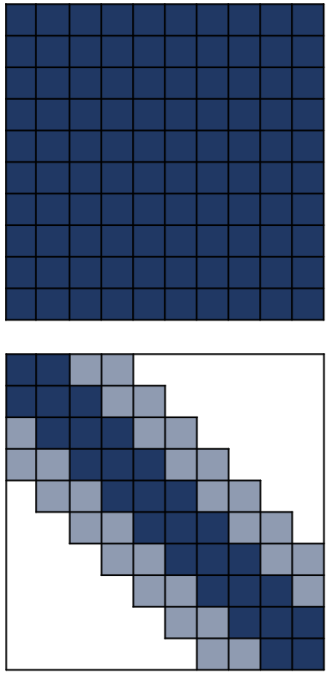


Computational Complexity

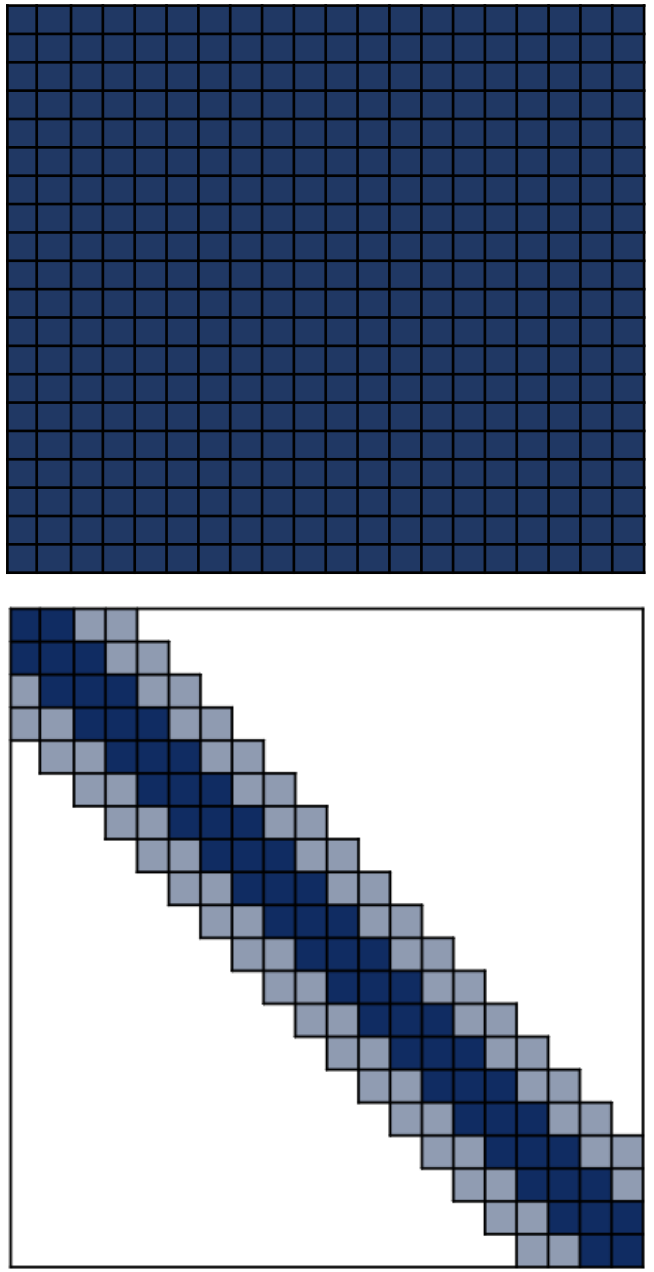
Transformer

Poolingformer

n = 10



n = 20



Computational Complexity

Model	Complexity
Transformer (Vaswani et al., 2017)	$\mathcal{O}(n^2)$
Reformer (Kitaev et al., 2020)	$\mathcal{O}(n \log n)$
Cluster-Former (Wang et al., 2020b)	$\mathcal{O}(n \log n)$
Longformer (Beltagy et al., 2020)	$\mathcal{O}(n)$
BigBird (Zaheer et al., 2020)	$\mathcal{O}(n)$
Poolingformer	$\mathcal{O}(n)$

Computational complexity of several related models.

Poolingformer for Document-level Summarization

Model	ROUGE-1	ROUGE-2	ROUGE-L
Sent-PTR-512	42.32	15.63	38.06
Extr-Abst-TLM-512	41.62	14.69	38.03
PEGASUS-512	44.21	16.95	38.83
Dancer-512	45.01	17.60	40.56
BigBird-16k	46.63	19.02	41.77
LED-4k	44.40	17.94	39.76
LED-16k	46.63	19.62	41.83
Poolingformer-4k	47.86	19.54	42.35
Poolingformer-16k	**48.47**	**20.23**	**42.69**

Poolingformer for Document-level QA

Google's NQ

Long Answer Leaderboard

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall	R@P = 90	R@P = 75	R@P = 50
1	PoolingFormer	PoolingFormer_Team	To-Be-Released	1/14/21	0.79823	0.7847	0.81224	0.62448	0.8379	0.88748
2	ClusterFormer	Dynamics_365_AI_Research	Microsoft	9/28/20	0.77969	0.78471	0.77473	0.56986	0.80654	0.86686
3	ETC-large	philly_pham	Google Research	5/25/20	0.7778	0.77476	0.78087	0.52204	0.79864	0.86598
4	ReflectionNet-ensemble	Wide_Field	Microsoft STCA NLP Group	2/9/20	0.77185	0.76791	0.77583	0.53345	0.78526	0.85238
5	ClusterFormer	Dynamics_365_AI_Research	Microsoft	9/19/20	0.76351	0.76151	0.76552	0.54354	0.77495	0.85852

Short Answer Leaderboard

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall	R@P = 90	R@P = 75	R@P = 50
1	ReflectionNet-ensemble	Wide_Field	Microsoft STCA NLP Group	2/9/20	0.64114	0.70445	0.58827	0.35046	0.54355	0.66144
2	PoolingFormer	PoolingFormer_Team	To-Be-Released	1/14/21	0.61629	0.70369	0.5482	0.17567	0.509	0.63995
3	RoBERTa-mnlp-ensemble	GAAMA	IBM Research AI	1/6/20	0.61409	0.6961	0.54936	0.28223	0.50436	0.62747
4	RikiNet_v2	DREAM_Losin	anonymous	11/29/19	0.61302	0.67612	0.56069	0.18089	0.48432	0.6417
5	ClusterFormer	Dynamics_365_AI_Research	Microsoft	9/28/20	0.60944	0.62149	0.59785	0.29326	0.49506	0.64053

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

TyDi QA

Passage Answer Task

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall
1	PoolingFormer	PoolingFormer_Team	To-Be-Released	1/25/2021	79.53	80.42	78.79
2	BERT with language-clustered vocab	Google-Research	Google Research	6/4/2020	77.65	77.43	78.00
3	GAAMA (XLM-R) with ARES system	GAAMA	IBM Research AI	11/13/2020	72.56	73.55	72.12
4	tydiqa-baseline	tydiqa-team	Google Research	2/15/2020	64.40	62.32	67.13
			IBM				

Minimal Answer Task

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall
1	PoolingFormer	PoolingFormer_Team	To-Be-Released	1/25/2021	67.65	73.48	63.27
2	GAAMA (XLM-R) with ARES system	GAAMA	IBM Research AI	11/13/2020	66.08	70.78	62.20
3	BERT with language-clustered vocab	Google-Research	Google Research	6/4/2020	63.40	67.19	60.21
4	mBERT-mnlp-single	GAAMA	IBM Research AI	8/12/2020	53.19	61.47	47.28

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

Ablation Study

Setting	w_1	w_2	C	LA F1	SA F1
RoBERTa$_{base}$	-	-	-	63.8	43.2
Poolingformer$_{base}$	128	-	-	66.3	43.1
Poolingformer$_{base}$	256	-	-	67.4	43.4
Poolingformer$_{base}$	512	-	-	66.1	42.6
Poolingformer$_{base}$	128	256	4	67.9	45.0
Poolingformer$_{base}$	128	512	4	**68.7**	**45.2**
Poolingformer$_{base}$	128	2,048	8	66.9	42.6
Poolingformer$_{base}$	128	2,048	16	67.0	44.4

Performance of different window lengths on NQ dev set.

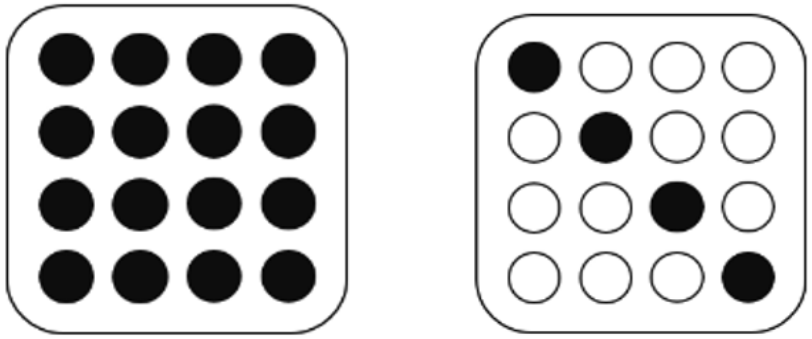
Setting	LA F1	SA F1
Poolingformer$_{base}$(Without $2nd$ level window)	66.3	43.1
Poolingformer$_{base}$(MEAN)	68.5	43.7
Poolingformer$_{base}$(MAX)	68.6	**45.3**
Poolingformer$_{base}$(LDConv)	**68.7**	45.2
Poolingformer$_{base}$(MeanLDConv)	67.7	44.1
Poolingformer$_{base}$(LDConv, *Mix*)	67.5	44.6
Poolingformer$_{base}$(LDConv, *Weight Sharing*)	67.2	44.2

Pooling and fusion approaches.

Mask Attention Networks: Rethinking and Strengthen Transformer. *NAACL 2021*

Mask Attention Networks: Rethinking and Strengthen Transformer

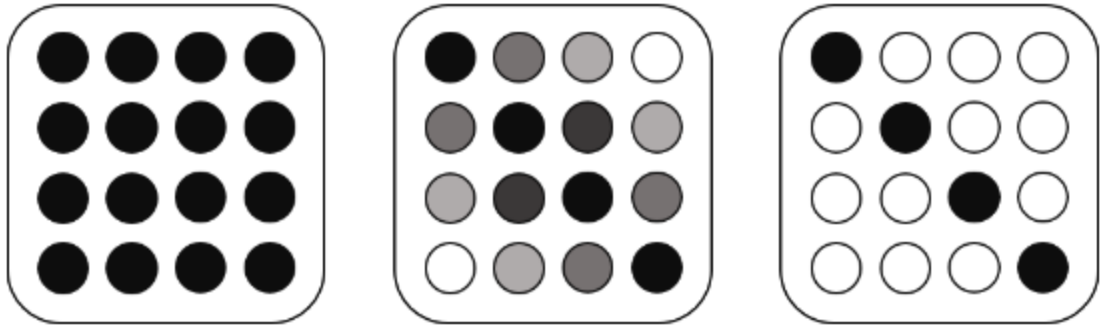
- Mask Attention Networks
  - Network has a mask matrix that element-wise multiplies a key-query attention matrix. SAN and FFN are two special cases of MANs
  - The mask matrix of SAN is an all-ones matrix for long range dependency modeling and capture the global semantics.
  - The mask matrix of FFN is an identity matrix for self-evolution.



(a) SAN (b) FFN

Mask Attention Networks: Rethinking and Strengthen Transformer

- How to strengthen Transformer in localness modeling and keep the advantage of SAN and FFN?
- Dynamic Mask Attention Network(DMAN)
  - The mask matrix depends on the query context and relative distance.
  - Tokens in specific neighborhood receive more attention.
  - A sequential layered structure: DMAN->SAN->FFN



(a) SAN (b) DMAN (c) FFN

Framework: Transformer

- Transformer Layer:

  Self-Attention Network(SAN) + Feed-Forward Network(FFN)

- Multihead Self-Attention Network

$$\mathcal{A}(Q,K,V) = \mathcal{S}(Q,K)V$$

$$\mathcal{S}(Q,K) = \left[\frac{\exp\left(Q_i K_j^T/\sqrt{d_k}\right)}{\sum_k \exp\left(Q_i K_k^T/\sqrt{d_k}\right)}\right]$$

$$H^l = \left[A^1, \cdots, A^I\right] W_H$$

$$A^i = \mathcal{A}\left(H^l W_Q^i, H^l W_K^i, H^l W_V^i\right)$$

- Feed Forward Network

$$H^{l+1} = \mathrm{ReLU}\left(H^l W_1\right) W_2$$

Framework: Mask Attention Networks

- Mask Attention Function: Adding a mask matrix on top of the attention function.

$$\mathcal{A}_M(Q,K,V) = \ S_M(Q,K)V$$
$$\mathcal{S}_M(Q,K) = \left[\frac{M_{i,j}\exp\left(Q_i K_j^T/\sqrt{d_k}\right)}{\sum_k M_{i,k}\exp\left(Q_i K_k^T/\sqrt{d_k}\right)}\right],$$

- Mask Attention Networks:

$$H^{l+1} = \mathcal{F}\left(\left[A^1_{M^1},\cdots,A^I_{M^I}\right]\right)W_H$$
$$A^i_{M^i} = \mathcal{A}_{M^i}\left(H^l W^i_Q, H^l W^i_K, H^l W^i_V\right)$$

Framework: Dynamic Mask Attention Network

- Mask other tokens that not in neighborhood.
- Static Mask:
  - Masking tokens not within b units

$$\mathbf{SM}[t,s] = \begin{cases} 0, & |\,t-s\,| > b \\ 1, & |\,t-s\,| \le b \end{cases}$$

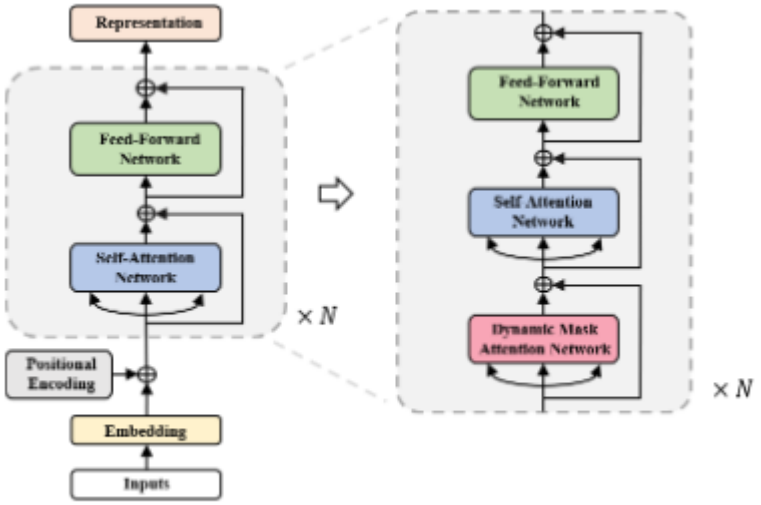
- Dynamic Mask Attention Network:
  - Mask value depends on the context h_t^l, head i and position t, s.

$$\mathbf{DM}_i^l[t,s] = \sigma\left(h_t^l W^l + P_{t-s}^l + U_i^l\right)$$

Framework: Collaboration of MANs

- Advantages of sub-networks of MANs.
  - SAN specializes in global semantic modeling.
  - FFN focuses on self-processing.
  - DMAN models local structure more effectively.

- Stacking: DMAN->SAN->FFN



a) Transformer: SAN→FFN

b) DMAN→SAN→FFN

Experimental Results

- Machine Translation

Model	IWSLT14 De-En		WMT14 En-De			
	small	params	base	params	big	params
Transformer (Vaswani et al., 2017)	34.4	36M	27.3	62M	28.4	213M
Convolutional Transformer (Yang et al., 2019b)	-	-	28.2	88M	28.7	-
Weighted Transformer (Ahmed et al., 2017)	-	-	28.4	65M	28.9	213M
Local Transformer (Yang et al., 2018)	-	-	28.5	89M	29.2	268M
Relative Transformer (Shaw et al., 2018)	-	-	26.8	-	29.2	-
Scaling NMT (Ott et al., 2018)	-	-	-	-	29.3	213M
Dynamic Conv (Wu et al., 2019)	35.2	-	-	-	29.7	213M
Ours	**36.3**	37M	**29.1**	63M	**30.4**	215M

Experimental Results

- Abstract Summarization

Model	CNN/Daily Mail				Gigaword			
	R-1	R-2	R-L	R-avg	R-1	R-2	R-L	R-avg
w/o Pre-train								
LEAD-3 (Nallapati et al. 2016)	40.42	17.62	36.67	31.57	-	-	-	-
PTGEN+Coverage (See, Liu, and Manning 2017)	39.53	17.28	36.38	31.06	-	-	-	-
Dynamic Conv (Wu et al. 2019)	39.84	16.25	36.73	30.94	-	-	-	-
Transformer (Vaswani et al. 2017)	39.50	16.06	36.63	30.73	37.57	18.90	34.69	30.38
Ours	**40.98**	**18.29**	**37.88**	**32.38**	**38.28**	**19.46**	**35.46**	**31.06**
w/ Pretrain								
S2S-ELMo (Edunov, Baevski, and Auli 2019)	41.56	18.94	38.47	32.99	-	-	-	-
BERTSUMABS (Liu and Lapata 2019)	41.72	19.39	38.76	33.29	-	-	-	-
MASS (Song et al. 2019)	42.12	19.50	39.01	33.54	38.73	19.71	35.96	31.46
Ours w/ Pretrain	**42.34**	**19.57**	**39.24**	**33.71**	**39.12**	**20.28**	**36.33**	**31.85**

Stacking Method Investigation

- We compare several different stacking methods
  - C#3, C#4, C#5 > C#1, C#2: DMAN contributes to improvement.
  - C#5, C#4 > C#3, C#2: SAN is necessary.
  - C#5 > C#4: Our stacking method is the best.

#	Method	BLEU
C#1	FFN→SAN→FFN	35.51
C#2	SAN→SAN→FFN	35.66
C#3	DMAN→DMAN→FFN	35.86
C#4	SAN→DMAN→FFN	35.91
C#5	DMAN→SAN→FFN	**36.35**

GLGE: A New General Language Generation Evaluation Benchmark. *ACL-Findings 2021*

GLGE: A New General Language Generation Evaluation Benchmark

Corpus	|Train|	|Dev|	|Test|	|Src.|	|Tgt.|	Input	Output	Metric
Abstractive Text Summarization								
CNN/DailyMail	287,113	13,368	11,490	822.3	57.9	article	summary	R-1/R-2/R-L
Gigaword	3,803,957	189,651	1,951	33.7	8.7	passage	headline	R-1/R-2/R-L
XSUM	204,017	11,327	11,333	358.5	21.1	article	summary	R-1/R-2/R-L
MSNews	136,082	7,496	7,562	310.7	9.7	article	headline	R-1/R-2/R-L
Answer-aware Question Generation								
SQuAD 1.1	75,722	10,570	11,877	149.4	11.5	answer/passage	question	R-L/B-4/MTR
MSQG	198,058	11,008	11,022	45.9	5.9	highlight/passage	question	R-L/B-4/MTR
Conversational Question Answering								
CoQA	108,647	3,935	4,048	354.4	2.6	history/passage	answer	F1-Score
Personalizing Dialogue								
PersonaChat	122,499	14,602	14,056	120.8	11.8	persona/history	response	B-1/B-2/D-1/D-2

GLGE Leaderboard

Leaderboard (11/24/2020-Present) ranked by **GLGE-Easy** Score (average score on 8 NLG tasks)

Model	Submission Date	GLGE Score	CNN/DailyMail	Gigaword	XSUM	MSNews	SQuAD 1.1	MSQG	CoQA	PersonaChat
CTRLgen (42Maru)	2021-03-08	37.6	46.2/22.6/43.1	39.2/20.0/36.5	45.7/22.4/37.3	45.4/26.1/41.5	51.3/23.0/26.8	39.8/10.6/24.3	75.1	51.0/40.2/1.1/6.8
P2DeNet (Anonymous)	2021-05-11	37.4	44.3/21.0/41.4	39.6/20.2/36.8	45.3/22.3/37.3	44.6/25.0/40.8	51.6/23.0/26.6	39.5/11.0/23.6	75.3	48.8/39.4/1.7/13.7
ProphetNet-large (GLGE Team)	2020-11-24	36.5	44.2/21.1/41.3	39.5/20.4/36.6	44.4/21.3/36.4	44.1/24.4/40.2	51.5/22.5/26.0	38.3/9.6/23.3	73.0	46.7/39.0/1.3/7.5
BART-large (GLGE Team)	2020-11-24	35.8	44.1/21.2/40.9	38.1/18.4/34.9	45.1/22.2/37.2	43.8/24.0/39.2	50.3/22.0/26.4	38.8/9.2/24.3	68.6	49.9/40.0/1.3/8.0
MASS-middle (GLGE Team)	2020-11-24	34.3	42.9/19.8/39.8	38.9/20.2/36.2	39.1/16.5/31.4	40.4/21.5/36.8	49.9/21.3/25.2	38.9/9.5/23.5	67.6	46.0/38.2/1.2/6.2
ProphetNet-base (GLGE Team)	2020-11-24	33.8	42.5/19.7/39.5	38.9/19.9/36.0	39.8/17.1/32.0	40.6/21.6/37.0	48.0/19.5/23.9	37.1/9.3/22.7	65.3	46.0/38.4/1.3/7.3
MASS-base (GLGE Team)	2020-11-24	33.6	42.1/19.5/39.0	38.7/19.7/35.9	39.7/17.2/31.9	39.4/21.0/36.1	49.4/20.1/24.4	38.9/10.2/23.3	65.4	41.0/35.7/1.4/6.9

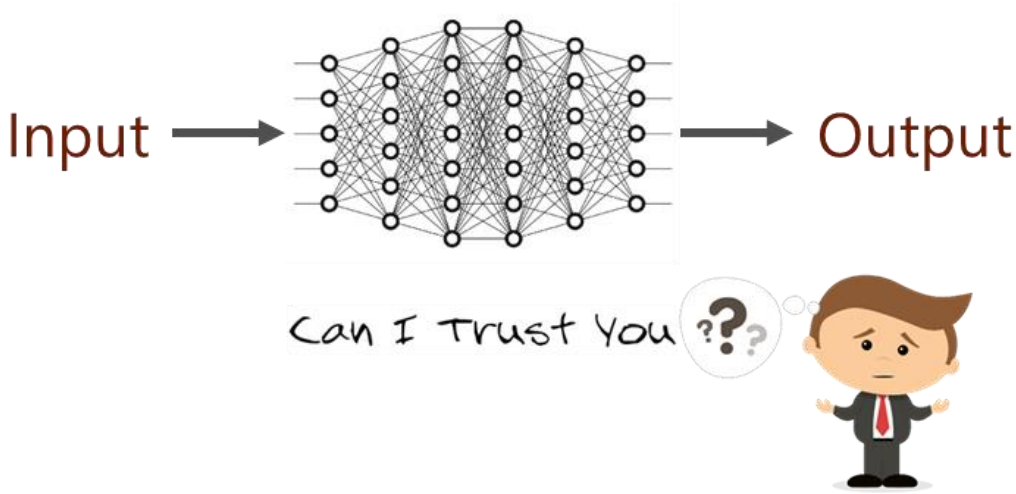
6). Conclusion & Future Work

Conclusion & Challenge

Pre-trained models can learn general representations and achieve SOTAs on many NLP/CV/SE tasks.

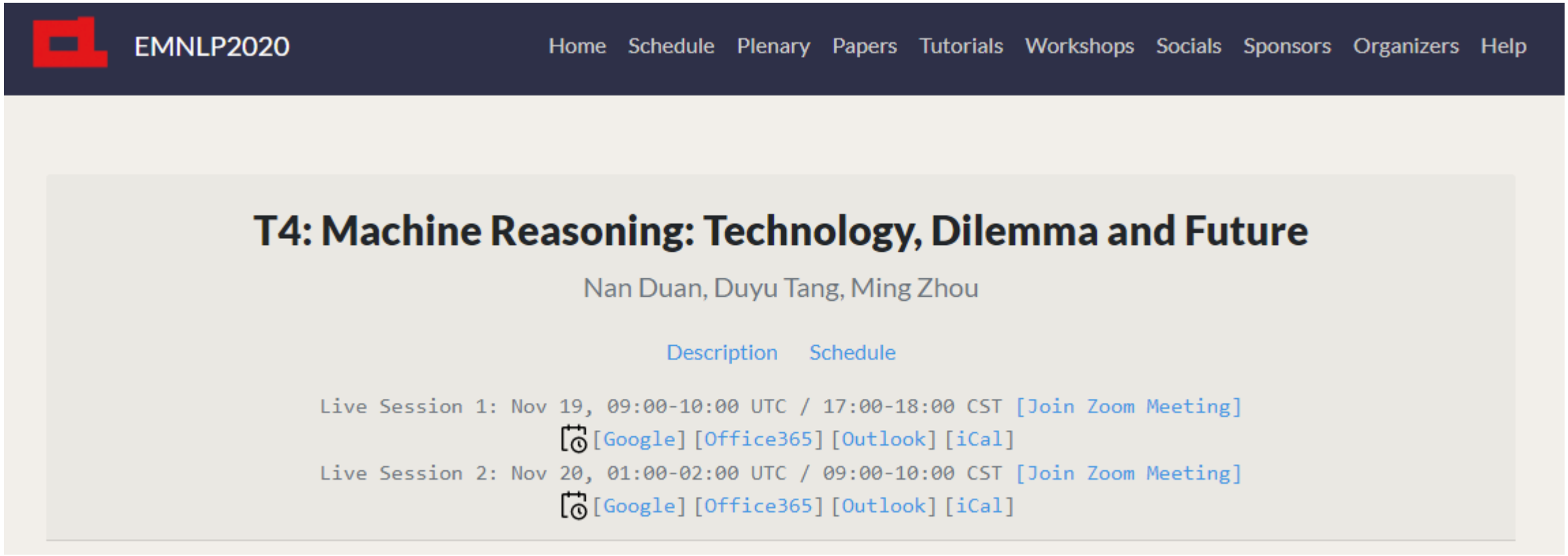
But they still have shortcomings,

- The predictions are not interpretable and trustable.
- The working mechanisms are hard to be analyzed or understood.
- Such models are usually hard to be integrated with existing knowledge.



Machine Reasoning (aka. Knowledge-aware NLU/NLG)

Machine Reasoning: an interpretable decision-making process that can solve problems or draw conclusions from what the system is told (i.e., facts and observations) and already knows (i.e., models, common sense and knowledge) under certain constraints.



EMNLP2020: Machine Reasoning: Technology, Dilemma and Future

4 Types of Machine Reasoning Frameworks

Symbolic Reasoning

Symbolic Knowledge
+ Inference based on Logic Rules

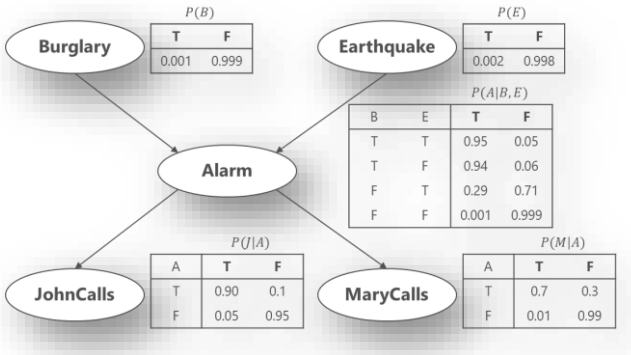
$\boldsymbol{KB_1}$: $\forall x\ cat(x) \Rightarrow like(x, fish)$
$\boldsymbol{KB_2}$: $\forall x \forall y\ (cat(x) \wedge like(x, y)) \Rightarrow eat(x, y)$
$\boldsymbol{KB_3}$: $cat(Tom)$

?

α: $eat(Tom, fish)$

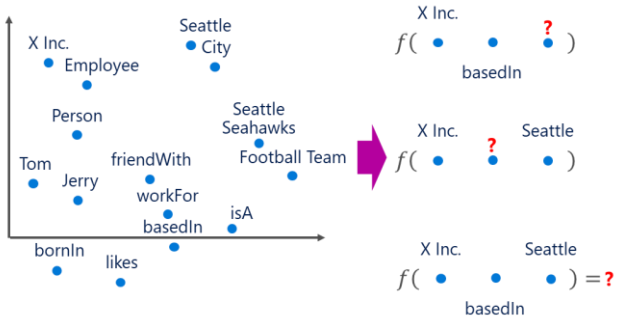
Probabilistic Reasoning

Probabilistic Symbolic Knowledge
+ Inference based on Probabilistic Graphical Models



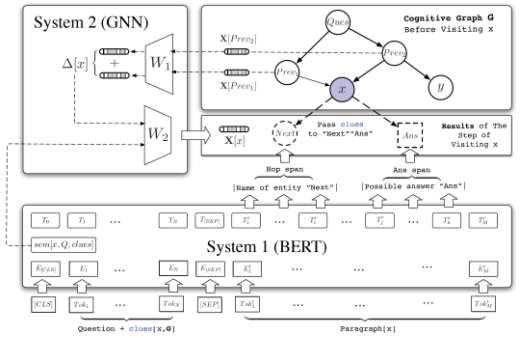
Neural-Symbolic Reasoning

Vector Representation of Symbolic Knowledge
+ Inference based on Neural Networks



Neural-Evidence Reasoning

Vector Representation of Non-Symbolic Evidence Knowledge
+ Inference based on Neural Networks



Pre-training (Perception) + Dense IR (Knowledge) + Reasoning (Cognition)

Machine Reasoning
(aka. Knowledge-aware NLU/NLG)

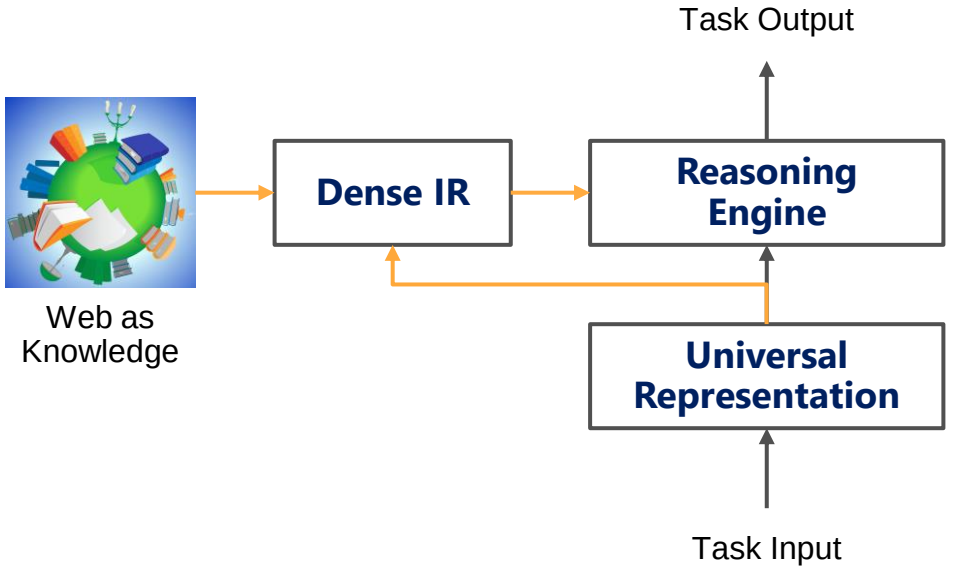
Solve NLU and NLG tasks based on (i) input representations from pre-trained models, (ii) external knowledge from dense IR and (iii) reasoning algorithms.

Dense Information Retrieval
from Heterogeneous Data Sources

Given a task input, retrieve related information from heterogeneous data sources (documents, databases, images, videos, etc.) based on their vector representations.

Universal Vision-Language Representation Learning

Learn universal representations from and for visual and multilingual text contents &
Learn common sense from visual contents for next generation NLP pre-trained models.



Thank YOU!

We are hiring Researchers and Interns, NOW!

Hiring Contact: nanduan@microsoft.com