CCAI 2020 大会特邀报告(II)

微软亚洲研究院：

多语种和多模态任务的预训练模型

临菲信息技术港

CCAI 2020 第二天大会特邀报告。ppt 根据视频摘录。

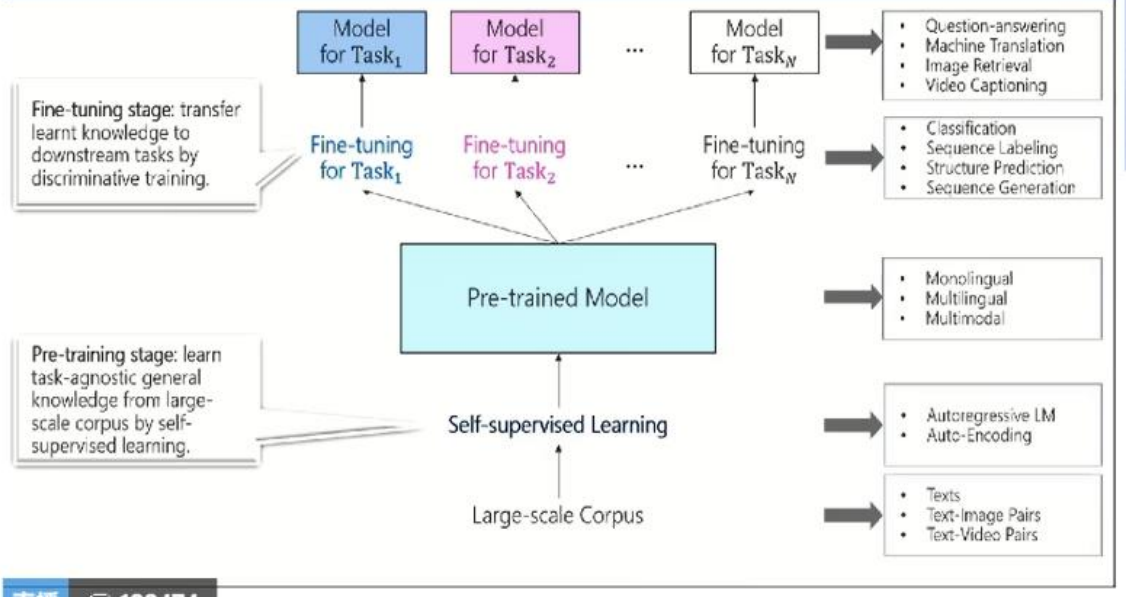
Pre-trained Models in Multi-Lingual and Multi-Moduality Tasks

报告人：周明，微软亚洲研究院副院长。

Agenda

1. Pre-trained models
2. Pre-trained models in multi-lingual tasks
3. Pre-trained models in multi-modality tasks
4. Summary

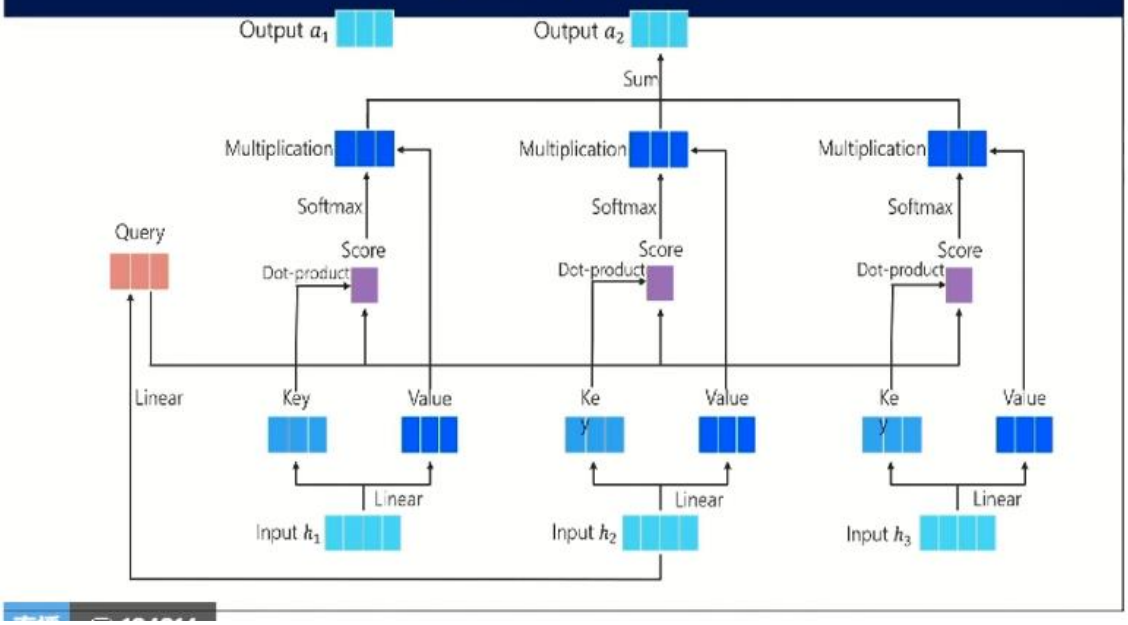
Pre-trained Model: A New Paradigm of NLP



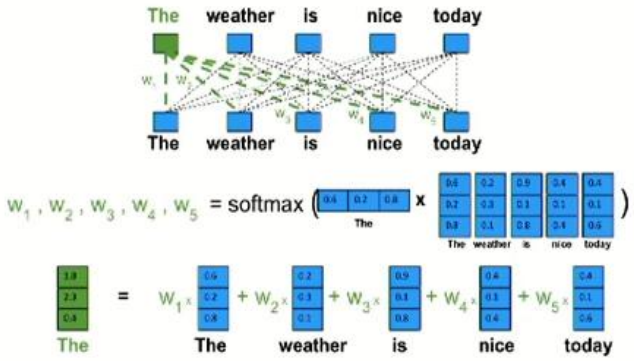
Why Pre-trained Models?

1. Pre-trained models embed task-agnostic general knowledge.
   - Syntactic knowledge and semantic knowledge are implicitly encoded
2. Pre-trained models transfer learnt knowledge to downstream tasks.
   - Tasks with low-resource annotations tasks and languages
3. Pre-trained models supports almost all NLP tasks with SOTA results
   - Natural language understanding tasks and generation tasks
4. Provide a scalable solution to various applications
   - Supporting a new task only needs finetuning via task-specific labelled data

Self-Attention

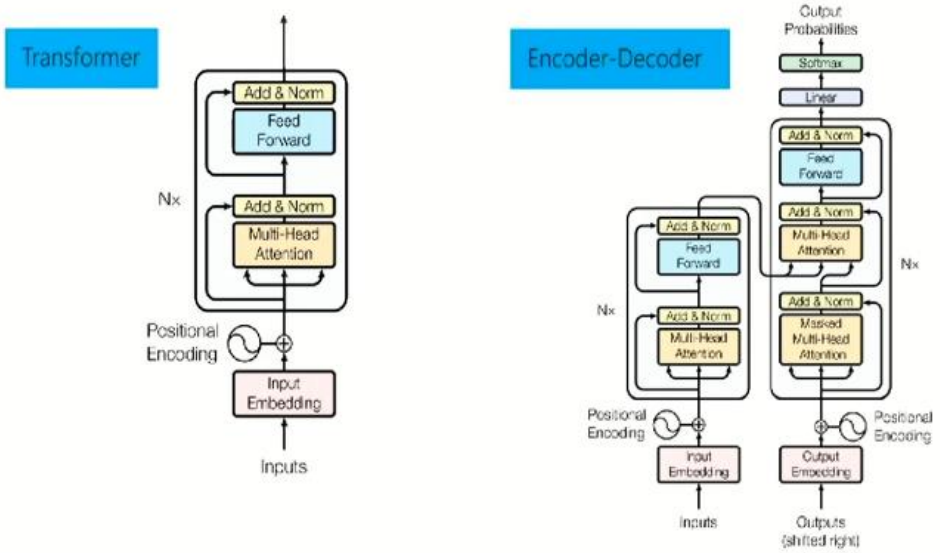


A Simplified Example for Self-Attention



A simplified example of self-attention in Transformer

Key Technologies (1): Transformer as Backbone



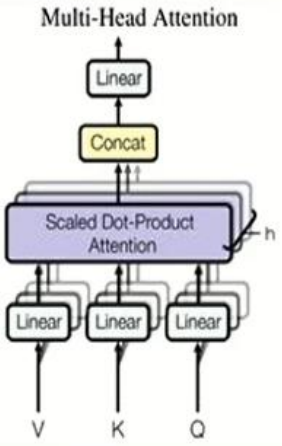
Multi-Head Attention

Multi-Head Attention

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

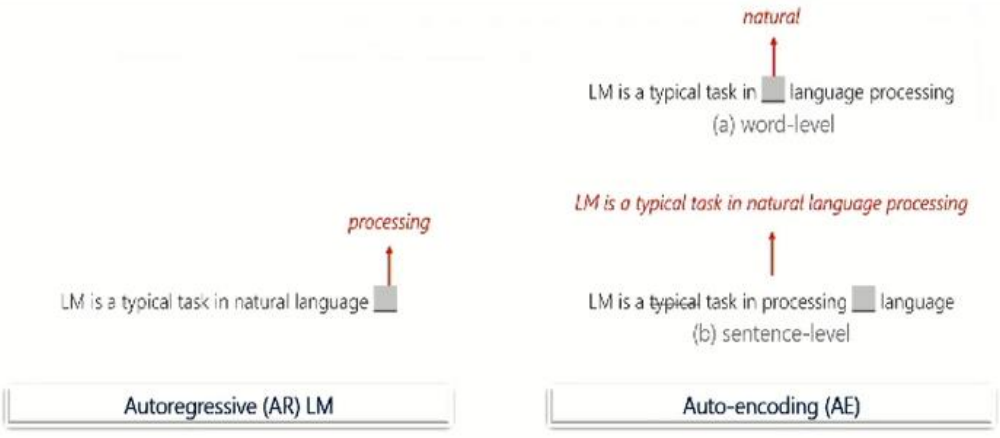
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

- Multi-head attention allows the model to jointly attend to **information from different representation subspaces** at different positions.



https://arxiv.org/pdf/1706.03762.pdf

Key Technologies (2): Pre-training by Self-supervised Learning



Self-supervised learning is a form of unsupervised learning where the data itself provides the supervision.

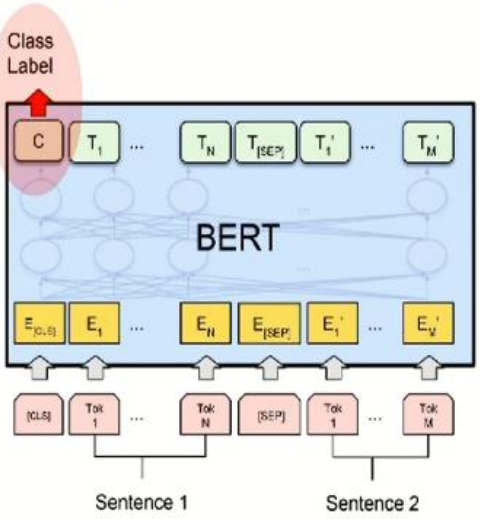
Key Technologies (3): Fine-tuning by Discriminative Training

(take BERT-based Sentence Pair Matching as an example)

Given the final hidden vector $C \in \mathbb{R}^H$ of the first input token ([CLS]), fine-tune BERT by a standard classification loss with C and W:

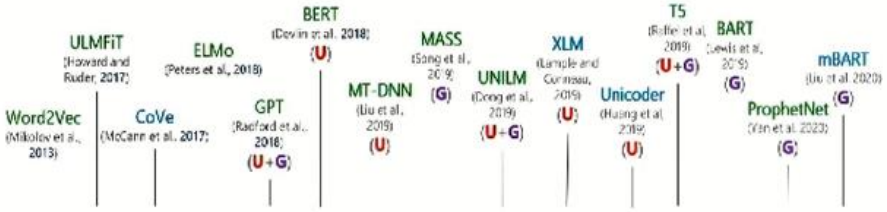
$$\log(\text{softmax}(CW^{\mathrm{T}}))$$

where $W \in \mathbb{R}^{K \times H}$ is a classification layer, K is the number of labels.



Roadmap

GREEN: monolingual pre-trained models
BLUE: multilingual pre-trained models
U: for understanding tasks
G: for generation tasks



Timeline of Pre-trained Models for Natural Language

Pre-trained Models for Natural Language Processing: A Survey

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, Xuanjing Huang

Recently, the emergence of pre-trained models (PTMs) has brought natural language processing (NLP) to a new era. In this survey, we provide a comprehensive review of PTMs for NLP. We first briefly introduce language representation learning and its research progress. Then we systematically categorize existing PTMs based on a taxonomy with four perspectives. Next, we describe how to adapt the knowledge of PTMs to the downstream tasks. Finally, we outline some potential directions of PTMs for future research. This survey is purposed to be a hands-on guide for understanding, using, and developing PTMs for various NLP tasks.

https://arxiv.org/abs/2003.08271

NLP Tasks

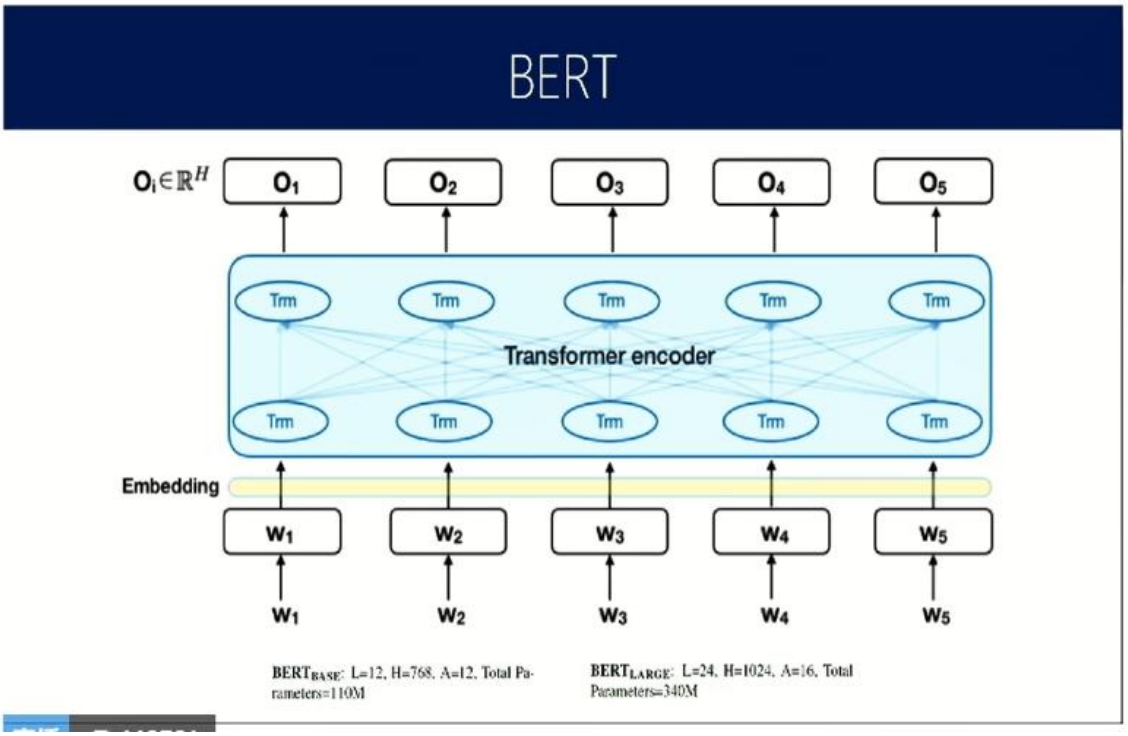
- Machine Translation
- Search Engine
- Semantic Parsing
- Question Answering
- Chatbot & Dialogue
- Paraphrase Classification
- Text Entailment
- Sentiment Analysis
- ...

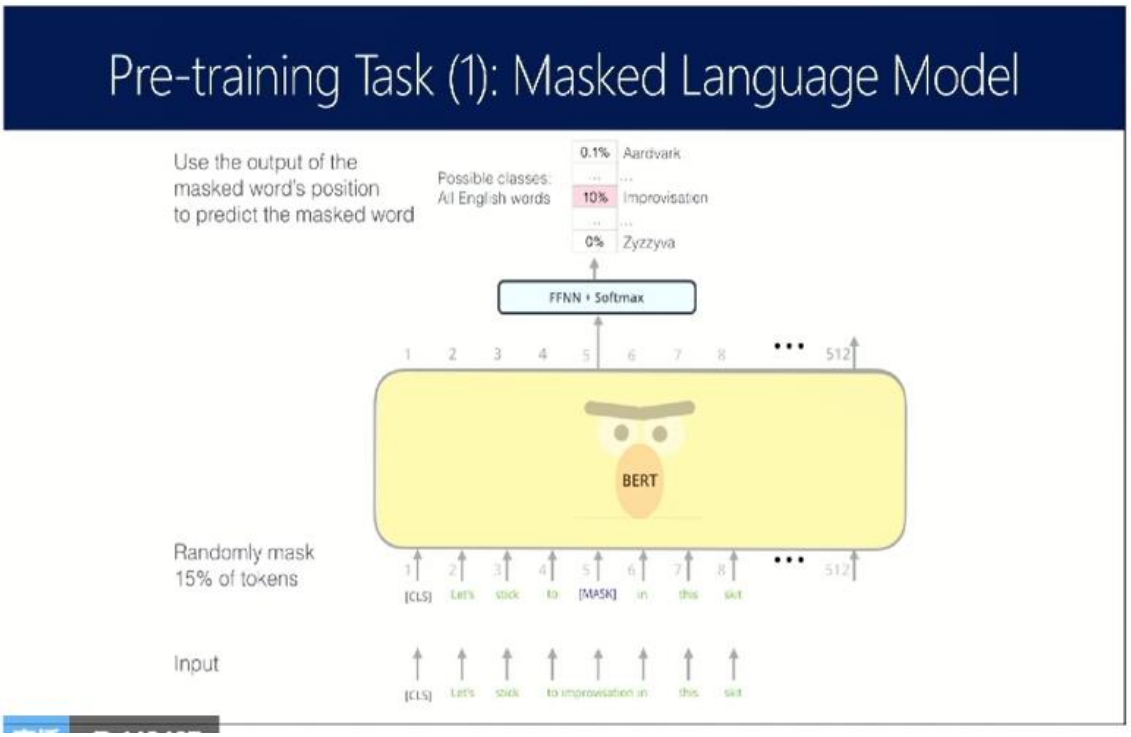
Connections and Differences Between (Monolingual) Pre-trained Models

Model Name	Model Usage	Model Backbone	Model Contribution
GPT (OpenAI)	Understanding & Generation	Transformer Encoder	1st unidirectional pre-trained LM based on Transformer
BERT (Google)	Understanding	Transformer Encoder	1st bidirectional pre-trained LM based on Transformer
MT-DNN (MS)	Understanding	Transformer Encoder	use multiple understanding tasks in pre-training
MASS (MS)	Generation	Separate Transformer Encoder-Decoder	use masked span prediction for generation tasks
UniLM (MS)	Understanding & Generation	Unified Transformer Encoder-Decoder	unify understanding and generation tasks in pre-training with different attention masks
RoBERTa (FB)	Understanding	Transformer Encoder	use better pre-training tricks, such as dynamic masking, large batches, removing NSP, data sampling
ERNIE (Baidu)	Understanding	Transformer Encoder	prove noun phrase masking and entity masking are better than word masking
SpanBERT (FB)	Understanding	Transformer Encoder	prove random span masking is better than others
XLNet (Google)	Understanding	Transformer Encoder	unify autoregressive LM and autoencoding tasks in pre-training with the two-stream self-attention
T5 (Google)	Generation	Separate Transformer Encoder-Decoder	use a separate encoder-decoder for understanding and generation tasks and prove it is the best choice; compare different hyper-parameters and show the best settings
BART (FB)	Generation	Separate Transformer Encoder-Decoder	try different text noising methods for generation tasks
ELECTRA (Google)	Understanding	Transformer Generator-Discriminator	use a simple but effective GAN-style pre-training task
ProphetNet (MS)	Generation	Separate Transformer Encoder-Decoder	use future n-gram prediction for generation tasks with the n-stream self-attention

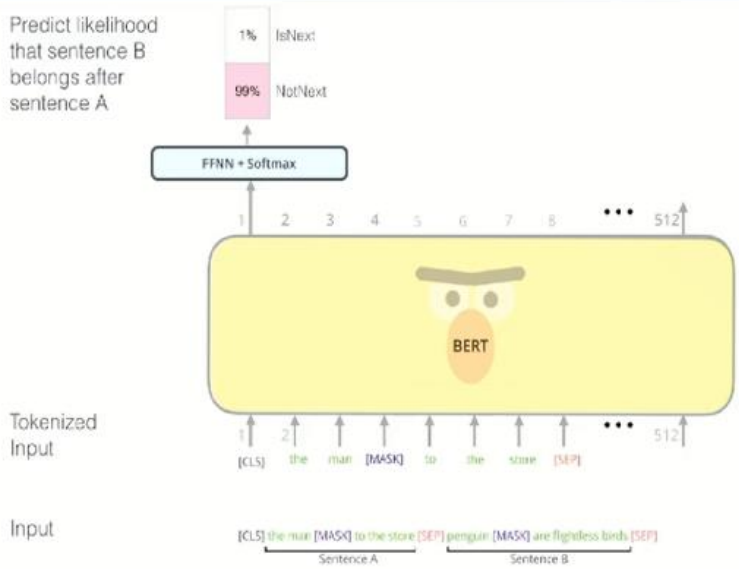
Push the Boundary of Pre-trained Models

- Model size (extremely big models, e.g. Turing 17B, GPT-3 175B parameters...)
- Pre-training methods and models (pre-training tasks, masking policy, network structures ...)
- From single language, to multi-lingual, to multi-modal
- Model compression, knowledge distilling, economical models for practical needs

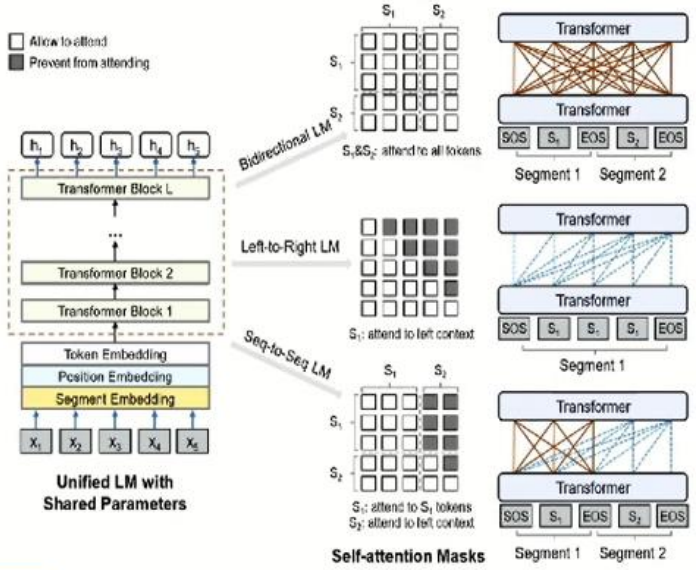

BERT
$O_i \in \mathbb{R}^H$
O1
O2
O3
O4
O5
Trm
Transformer encoder
Embedding
W1
W2
W3
W4
W5
BERTBASE: L=12, H=768, A=12, Total Parameters=110M
BERTLARGE: L=24, H=1024, A=16, Total Parameters=340M


Pre-training Task (1): Masked Language Model
Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
10% Improvisation
0% Zyzzyva
FFNN + Softmax
BERT
Randomly mask 15% of tokens
Input
[CLS]
[MASK]

Pre-training Task (2): Next Sentence Prediction



UniLM (Dong et al., 2019)



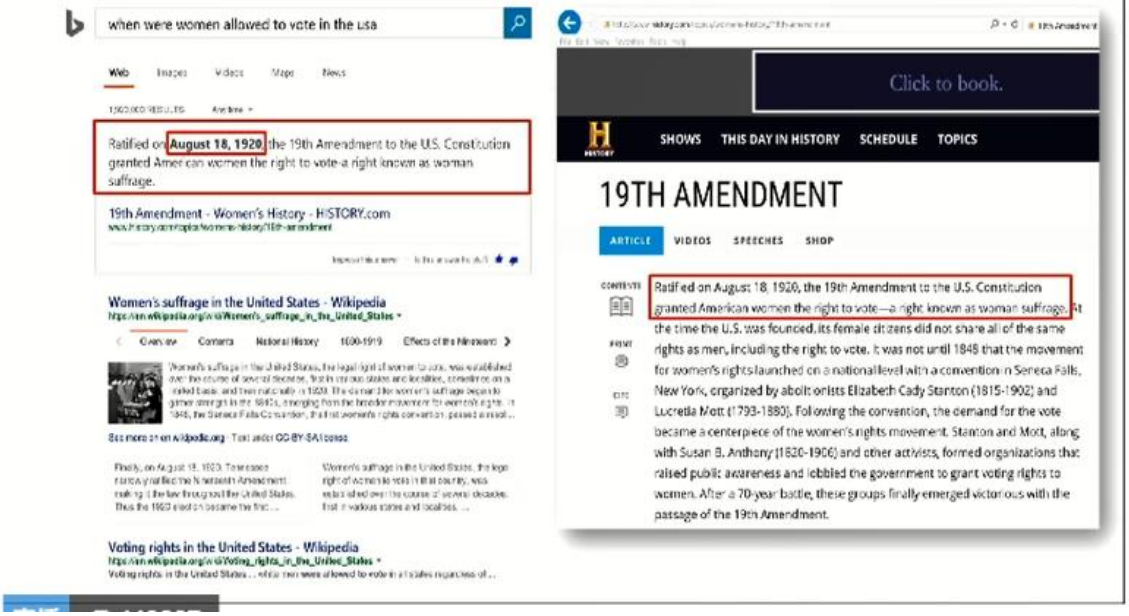
	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Summarization			
LEAD-3	40.42	17.62	36.67
Best Extractive	43.25	20.24	39.63
Abstractive Summarization			
PGNet	39.53	17.28	37.98
Bottom-Up	41.22	18.68	38.34
S2S-ELMo	41.56	18.94	38.47
UniLM	**43.47**	**20.30**	**40.63**

Table 3: Evaluation results on CNN/DailyMail. Models in the first block are extractive systems listed here for reference, while the others are abstractive models. The results of the best reported extractive model are taken from (Liu, 2019).

	EM	F1
RMR+ELMo (Hu et al., 2018)	71.4	73.7
BERT$_{\text{LARGE}}$	78.9	81.8
UniLM	**80.5**	**83.4**

Table 4: Extractive question answering results on the SQuAD development set.

Application: Question Answering



Application: Question Generation

Text Passage

in 1066[1,2] **, duke william ii**[3] of normandy conquered england killing king harold ii at the battle of hastings. **the invading normans and their descendants**[4] replaced the anglo-saxons as the ruling class of england.

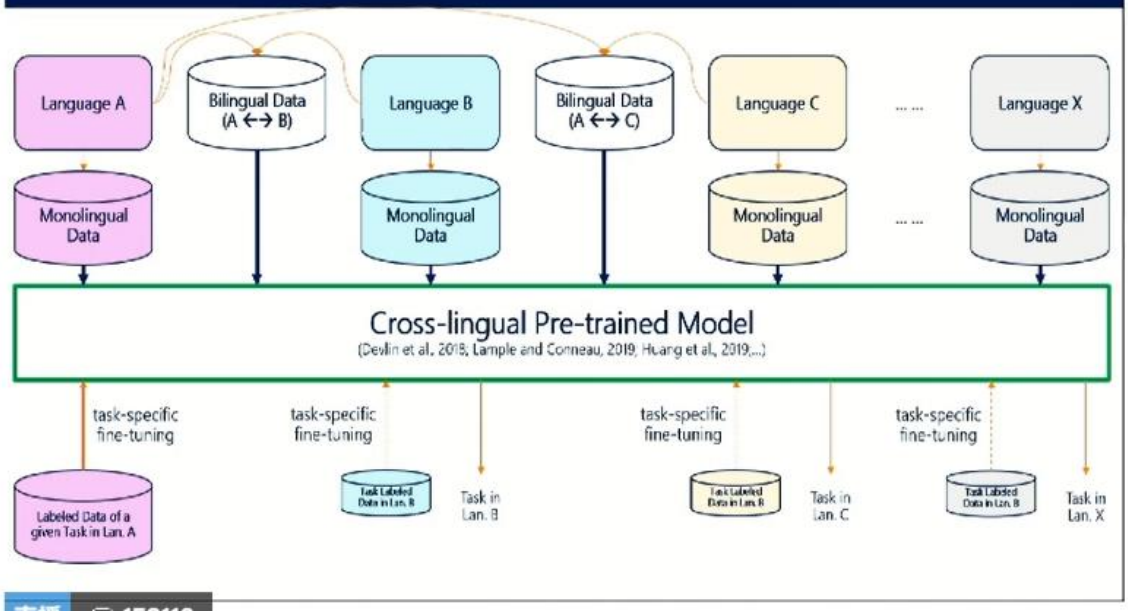
Questions Generated by our System

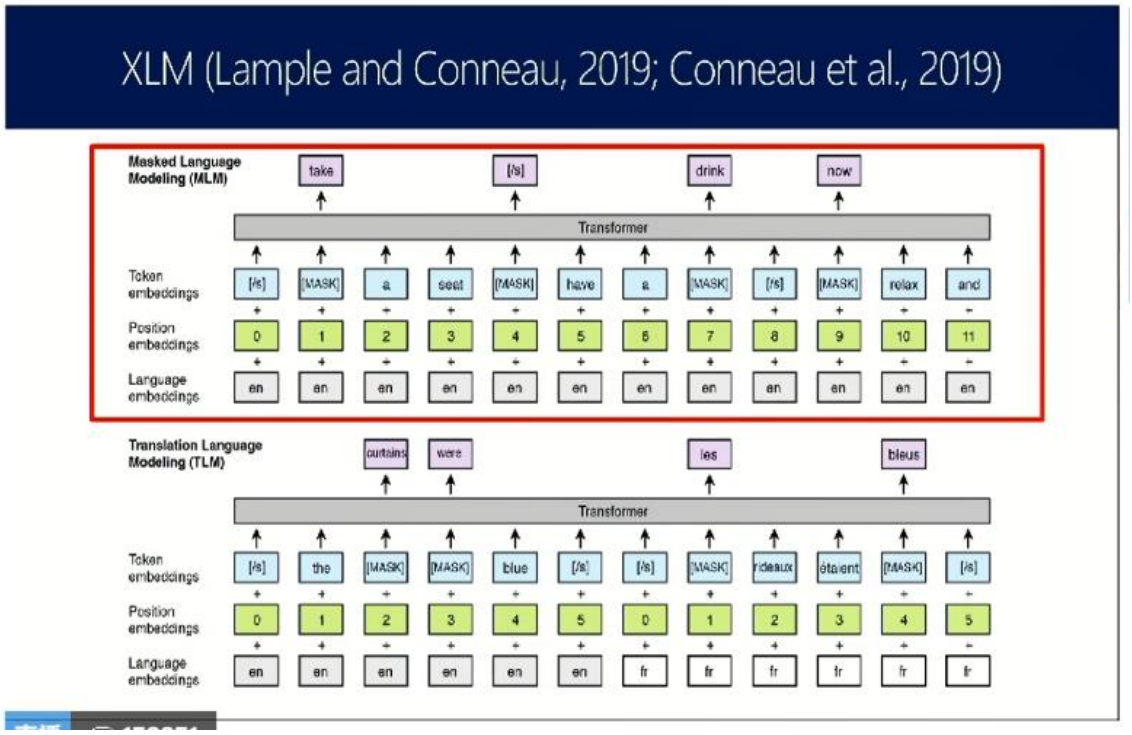
1) when did the battle of hastings take place?
2) in what year was the battle of hastings fought?
3) who conquered king harold ii at the battle of hastings?
4) who became the ruling class of england?

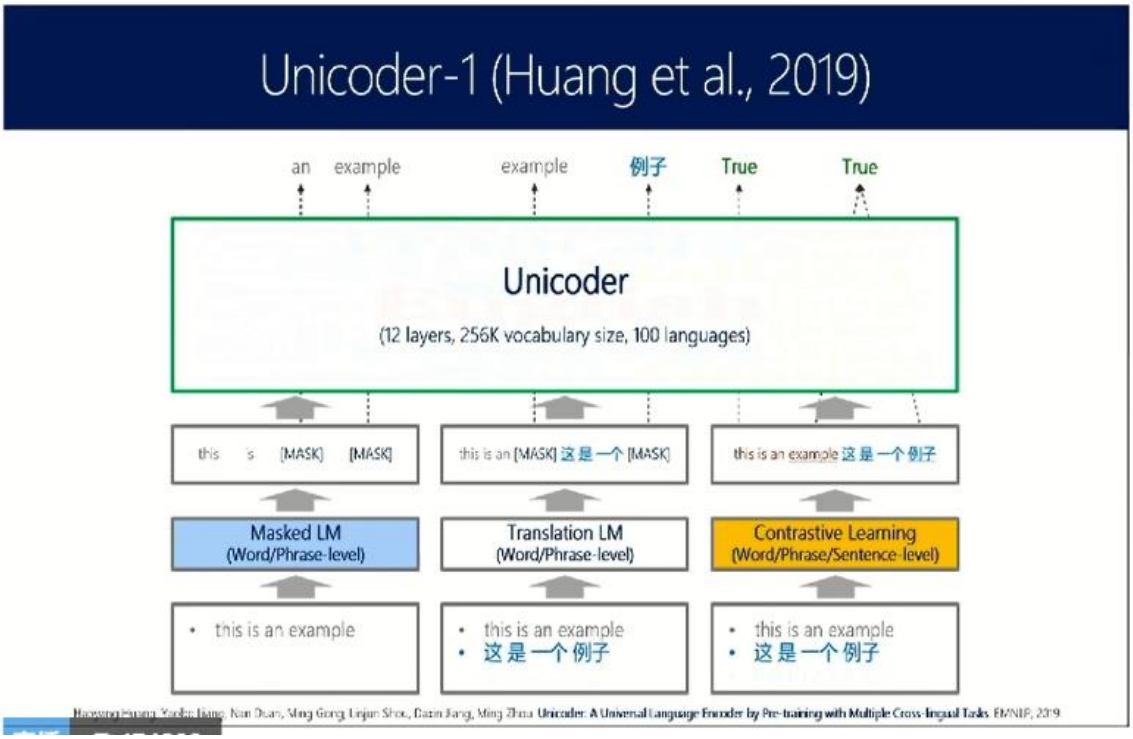
Agenda

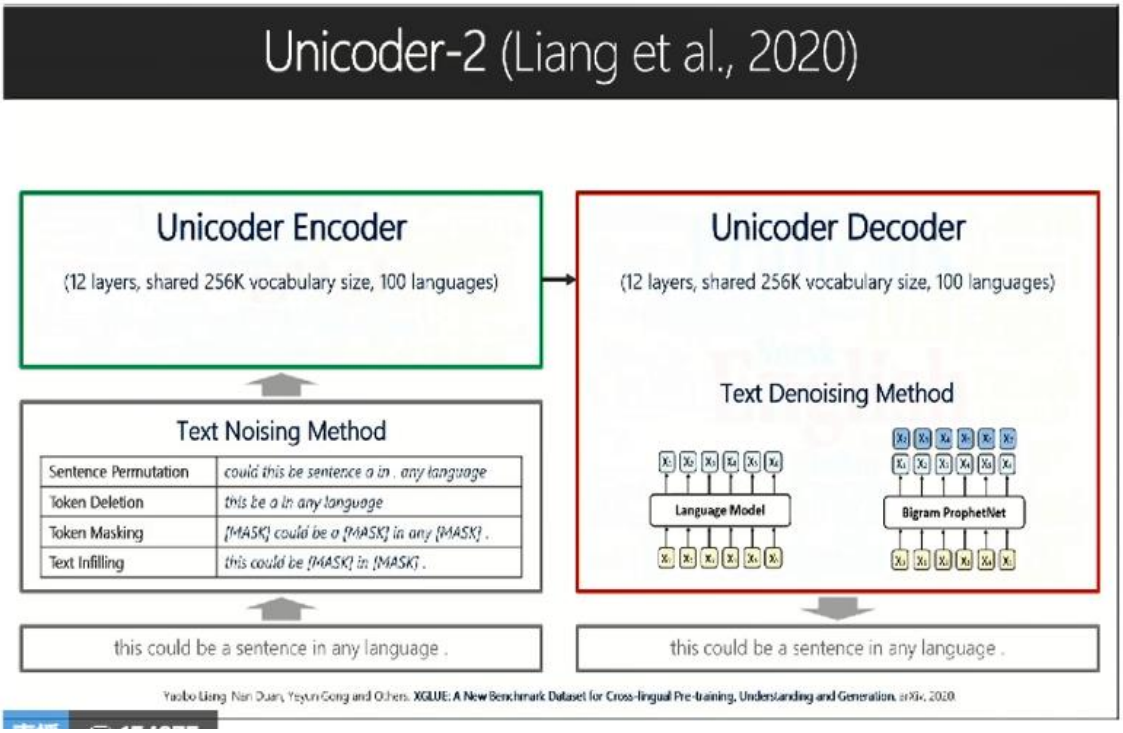
1. Pre-trained models
2. Pre-trained models in multi-lingual tasks
3. Pre-trained models in multi-modality tasks
4. Summary

Cross-Language Pre-training




XLM (Lample and Conneau, 2019; Conneau et al., 2019)
Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
Position embeddings
Language embeddings
Translation Language Modeling (TLM)
Transformer


Unicoder-1 (Huang et al., 2019)
an example
example
例子
True
True
Unicoder
(12 layers, 256K vocabulary size, 100 languages)
this is [MASK] [MASK]
this is an [MASK] 这是一个 [MASK]
this is an example 这是一个例子
Masked LM (Word/Phrase-level)
Translation LM (Word/Phrase-level)
Contrastive Learning (Word/Phrase/Sentence-level)
this is an example
this is an example
这是一个例子
this is an example
这是一个例子



Tasks in XGLUE

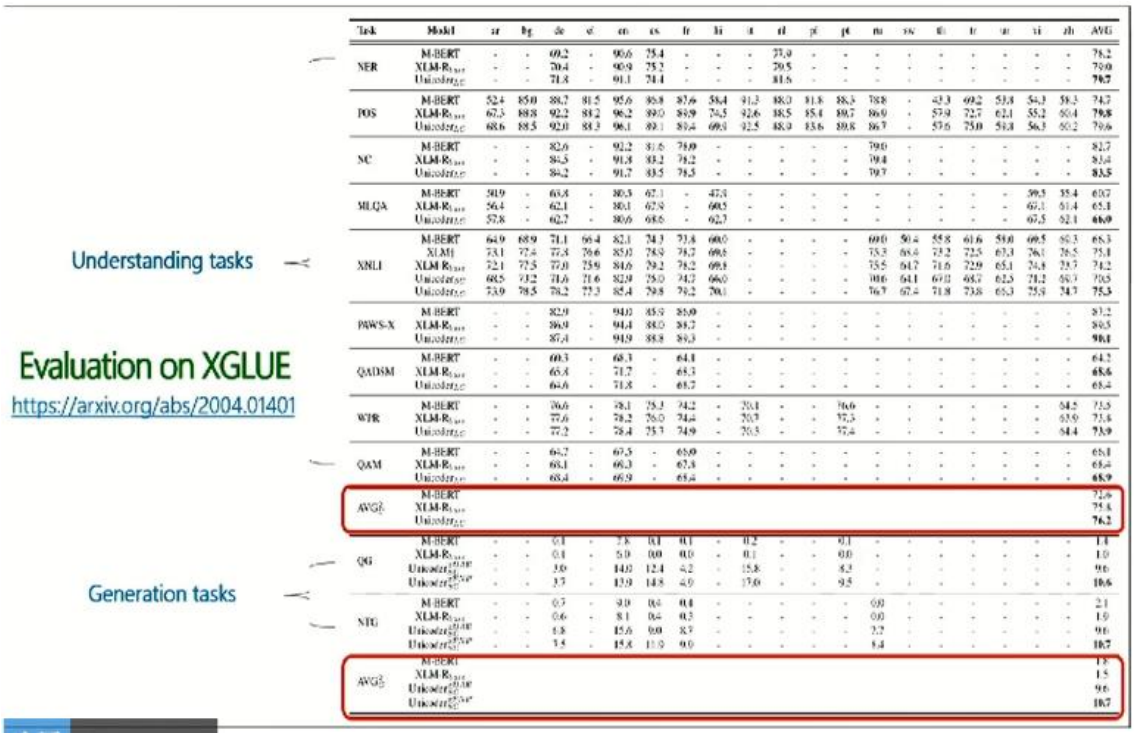
Task	# of Languages	|Train|[*]	|Dev|[*,†]	|Test|[*,†]	Metric	Data Source
NER	4	15.0K	2.8K	3.4K	F1	ECI Multilingual Text Corpus
POS	18	25.4K	1.0K	0.9K	ACC	UD Tree-banks (v2.5)
NC*	5	100K	10K	10K	ACC	Commercial News Website
MLQA	7	87.6K	0.6K	5.7K	F1	Wikipedia
XNLI	15	433K	2.5K	5K	ACC	MultiNLI Corpus
PAWS-X	4	49.4K	2K	2K	ACC	Wikipedia
QADSM*	3	100K	10K	10K	ACC	Commercial Search Engine
WPR*	7	100K	10K	10K	nDCG	Commercial Search Engine
QAM*	3	100K	10K	10K	ACC	Commercial Search Engine
QG*	6	100K	10K	10K	BLEU-4	Commercial Search Engine
NTG*	5	300K	10K	10K	BLEU-4	Commercial News Website

Table 2: 11 downstream tasks in XGLUE. For each task, training set is only available in English. |Train|[*] denotes the number of labeled instances in the training set. |Dev|[*,†] and |Test|[*,†] denote the average numbers of labeled instances in the dev sets and test sets, respectively. * denotes the corresponding dataset is constructed by this paper.

Task	ar	bg	de	el	en	es	fr	hi	it	nl	pl	pt	ru	sw	th	tr	ur	vi	zh
NER			✓		✓	✓				✓									
POS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
NC*			✓		✓	✓	✓						✓						
MLQA	✓		✓		✓	✓		✓										✓	✓
XNLI	✓	✓	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	✓	✓	✓
PAWS-X			✓		✓	✓	✓												
QADSM*			✓		✓		✓												
WPR*			✓		✓	✓	✓		✓			✓							✓
QAM*			✓		✓		✓												
QG*			✓		✓	✓	✓		✓			✓							
NTG*			✓		✓	✓	✓						✓						

Table 3: The 19 languages covered by the 11 downstream tasks: *Arabic* (ar), *Bulgarian* (bg), *German* (de), *Greek* (el), *English* (en), *Spanish* (es), *French* (fr), *Hindi* (hi), *Italian* (it), *Dutch* (nl), *Polish* (pl), *Portuguese* (pt), *Russian* (ru), *Swahili* (sw), *Thai* (th), *Turkish* (tr), *Urdu* (ur), *Vietnamese* (vi), and *Chinese* (zh). All these 6 new tasks with * are labeled by human, except es, it and pt datasets in QG (80+% accuracy) are obtained by an in-house QA ranker.

NER: Named Entity Recognition
POS: Part-of-Speech Tagging
NC: News Classification
MLQA: Multilingual MRC
XNLI: Natural Language Inference
PAWS-X: Paraphrase Classification
QADSM: Query-Ads Matching
WPR: Web Page Ranking
QAM: Question-Answer Matching
QG: Question Generation
NTG: News Title Generation



Application: Multilingual Question Answering

English translation: ***where is the largest sugar factory in the world***

English translation: ***The sugar refinery of the Algerian group Cevital produces 2.7 million tonnes of sugar a year, making it the largest refinery in the world. This refinery doubled its exports of white sugar from 377,000 tonnes to 600,000 tonnes in 2012.***

English translation: ***delete browser history windows 10***

English translation: ***In the tab "General" you will find the sub-item "Browser History". Click on the "Delete ..." button there. A window with the name "delete browser history" will open. Check the data you want to delete. Click on "Delete". The history has now been removed.***

Application: Multilingual News Headline Generation

en	Input News	if you 're planning a trip to europe , you probably want to check some famous landmarks off your list . but there are certain tourist traps you 're better off missing . susana victoria perez has more .
	Golden Title	do yourself a favor and avoid these tourist traps in europe
	Unicoder$^{xDAE}_{SC}$	tourist traps you should avoid in europe
fr	Input News	alain juppe , candidat a la primaire de la droite , " ne se sent pas engage " par les investitures decidees par le parti les republicains preside par nicolas sarkozy , a affirme jeudi a l' afp son directeur de campagne , gilles boyer . " c' est un processus mene a la hussarde . il n' y a pas de volonte d' equilibre et de rassemblement " , a-t-il denonce , en affirmant que " l' accord politique " entre les differents candidats a la primaire " n' a pas ete respecte " .
	Golden Title	legislatives : juppe " ne se sent pas engage " par les investitures
	Unicoder$^{xDAE}_{SC}$	alain juppe : " ne se sent pas engage " par les investitures
de	Input News	vermutlich zur verteidigung seines reviers hat ein aggressiver bussard in baden-wurttemberg einen radfahrer zu fall gebracht . der sich dabei schwer verletzte . wie die polizei in ludwigsburg am freitag mitteilte , attackierte der greifvogel den 51-jahrigen am vortag auf einem radweg entlang einer landesstraße . der bussard flog demnach so tief auf den radler zu , dass dieser ausweichen musste und sturzte . den angaben zufolge erlitt der mann schwere verletzungen und wurde von rettungskraften in ein krankenhaus gebracht . " aus luftiger hohe , von einem laternenmast aus , beobachtete der raubvogel anschließend die unfallaufnahme " , hieß es im polizeibericht .
	Golden Title	aggressiver bussard bringt radfahrer zu fall
	Unicoder$^{xDAE}_{SC}$	aggressiver bussard in ludwigsburg stürzes radler
es	Input News	despues de la marcha de bruce willis por problemas de agenda , steve carrell le sustituira asi en la nueva pelicula que prepara woody allen . segun informa variety , el actor se une al reparto ya formado por blake lively , parker posey , kristen stewart , jesse eisenberg , jeannie berlin ,corey stoll , anna camp , y ken stott , entre otros . como siempre , los detalles de la trama son aun un secreto aunque el rodaje se encuentre actualmente en marcha . por otro lado , aun no hay fecha de estreno ni distribuidora para la pelicula sin titulo de woody allen . sin embargo , el director tiene aun pendiente de estreno su ultimo filme con emma stone y joaquin phoenix titula da irrational man que se estrenara el proximo 25 de septiembre .
	Golden Title	steve carrell sustituye a bruce willis en la nueva pelicula de woody allen
	Unicoder$^{xDAE}_{SC}$	steve carrell sustituira a steve carrell en woody allen

Summary of Multilingual Pre-trained Models

Multilingual pre-trained models alleviate the low-resource issue in multiple languages

- Fine-tune on rich-resource languages (such as EN) and then apply to low-resource languages
- Show strong cross-lingual transfer capabilities in zero-shot or few-shot settings
- Achieve state-of-the-art results on various cross-lingual benchmarks.

Multilingual pretrained models can benefit many real-world applications

- Multilingual Search/QA/Ads/News/Text Summarization/...
- Low-resource Neural Machine Translation

Multilingual pretrained models still have many challenges

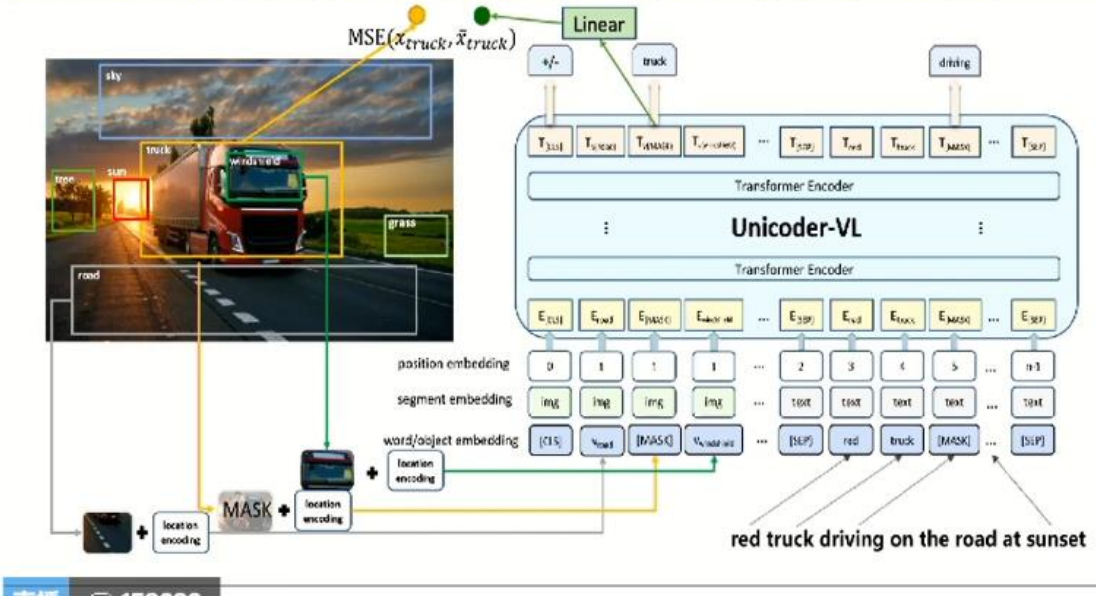
- Lacks training tasks fully leveraging the characteristics of multi-lingual data
- High cost to train huge parameters due to large vacab (250K)
- Transfer learning is weaker for languages in different families of language

Agenda

1. Pre-trained models
2. Pre-trained models in multi-lingual tasks
3. Pre-trained models in multi-modality tasks
4. Summary

Image-Language Pre-training

ViLBERT (Lu et al., 2019); Unicoder-VL (Li et al., 2019); VL-BERT (Su et al., 2019); UNITER (Chen et al., 2019)



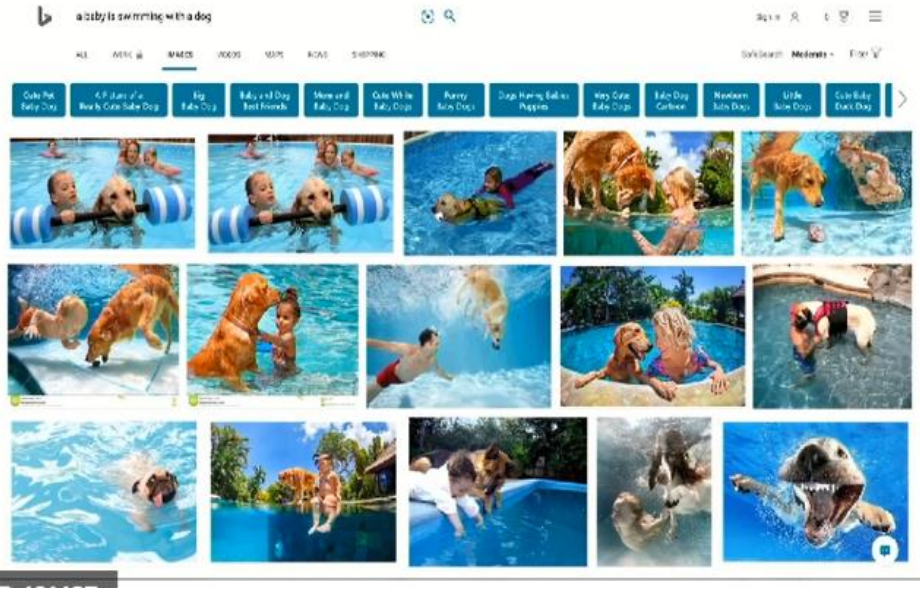
Evaluation Results

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

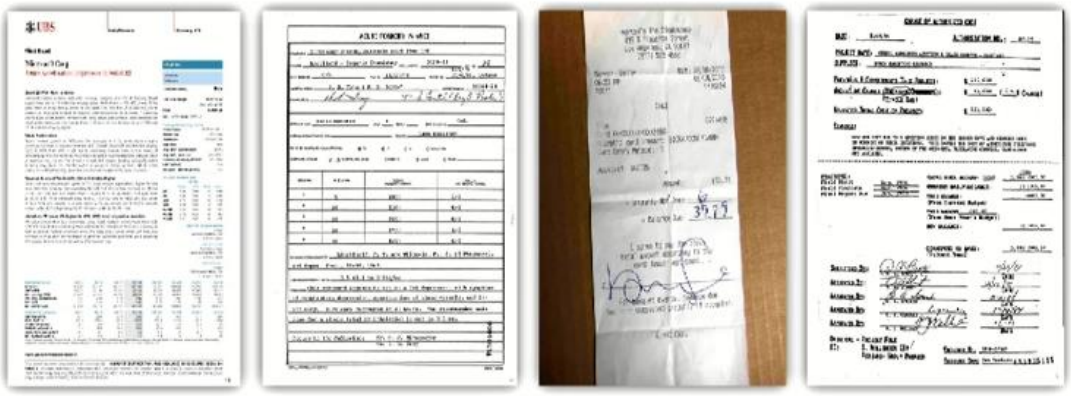
Model	Text-to-Image Retrieval (MSCOCO)			Image-to-Text Retrieval (MSCOCO)		
	R@1	R@5	R@10	R@1	R@5	R@10
UNITER (Chen et al., 2019)	48.4	76.7	85.9	63.3	87.0	93.1
Unicoder-VL (Li et al., 2020)	50.5	78.7	87.1	66.4	89.8	94.4

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

Application: Image Search



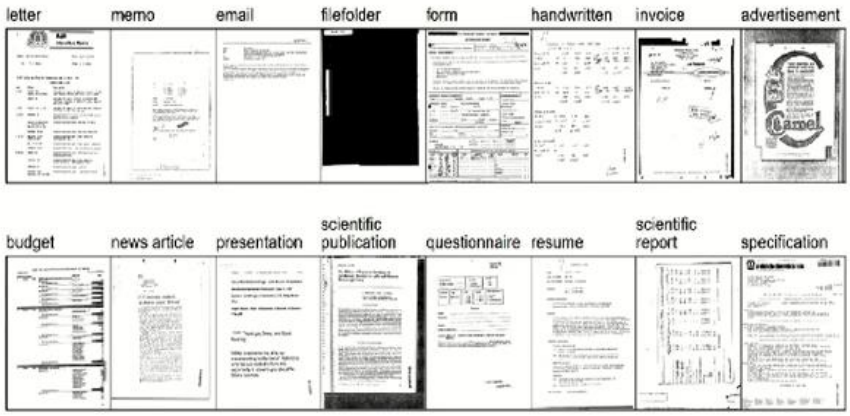
Document Understanding



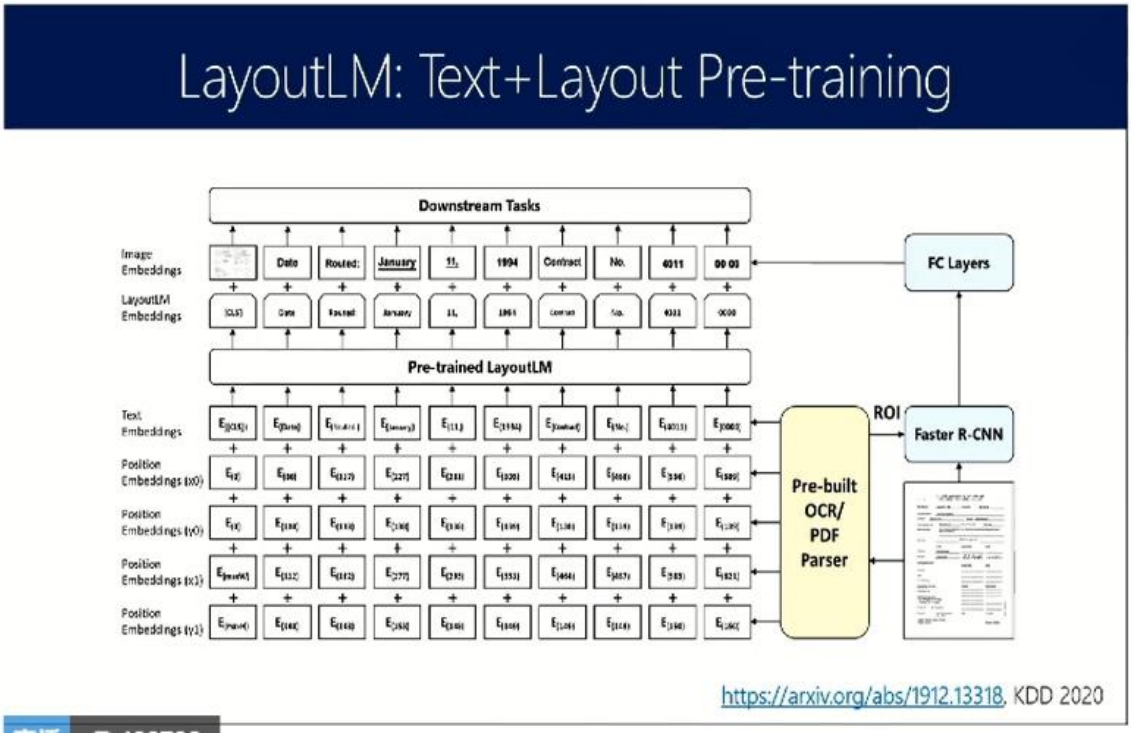
Information extraction from digital-born/scanned documents

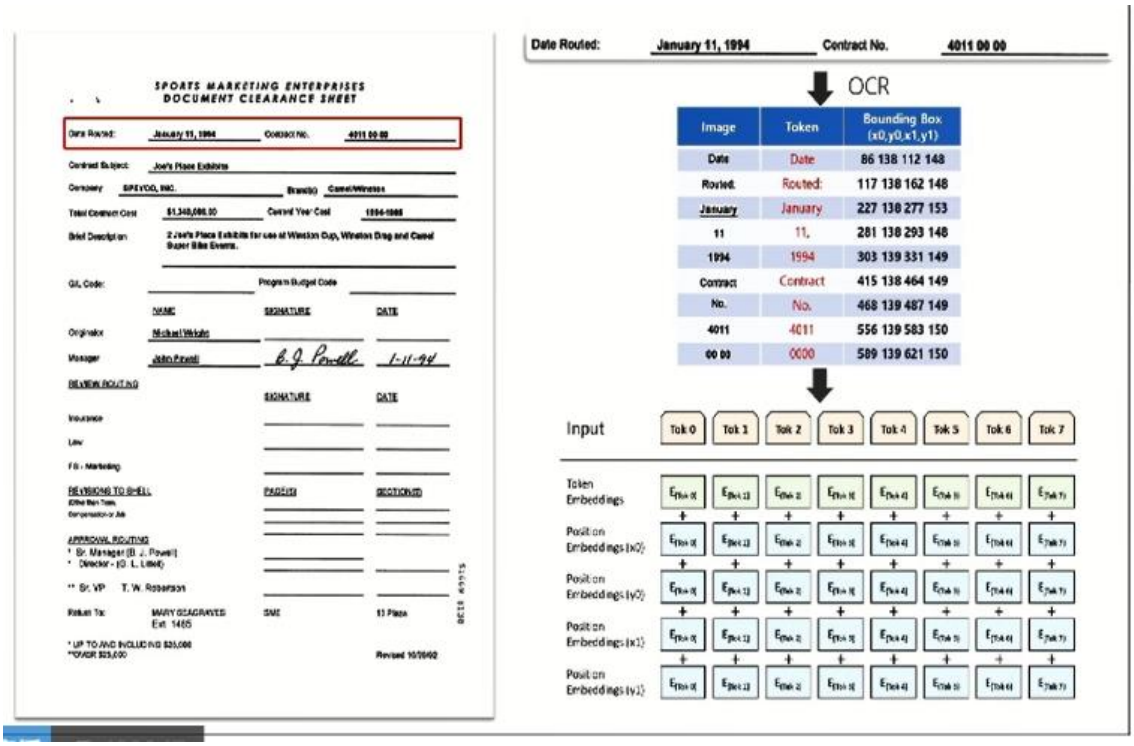
LayoutLM: Pre-training for text with rich **Layout** and **Style** information
https://arxiv.org/pdf/1912.13318.pdf, KDD 2020

Training Dataset for LayoutLM

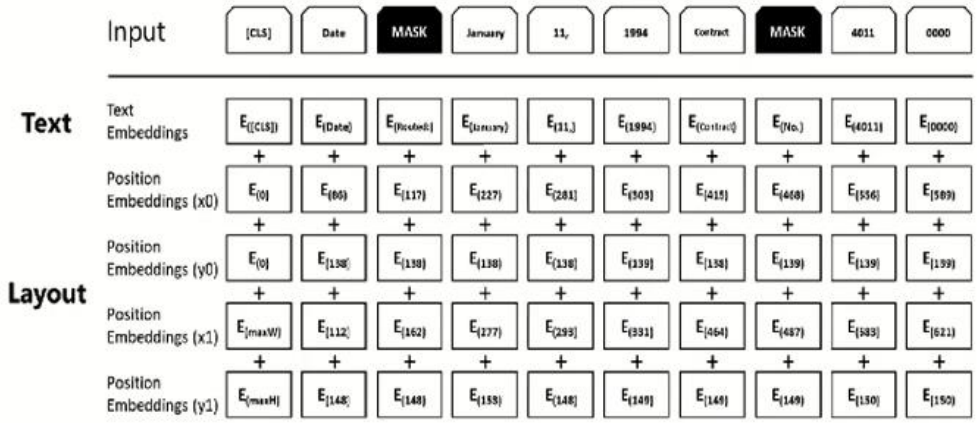


11 million scanned document images from IIT-CDIP Test Collection 1.0
https://ir.nist.gov/cdip/

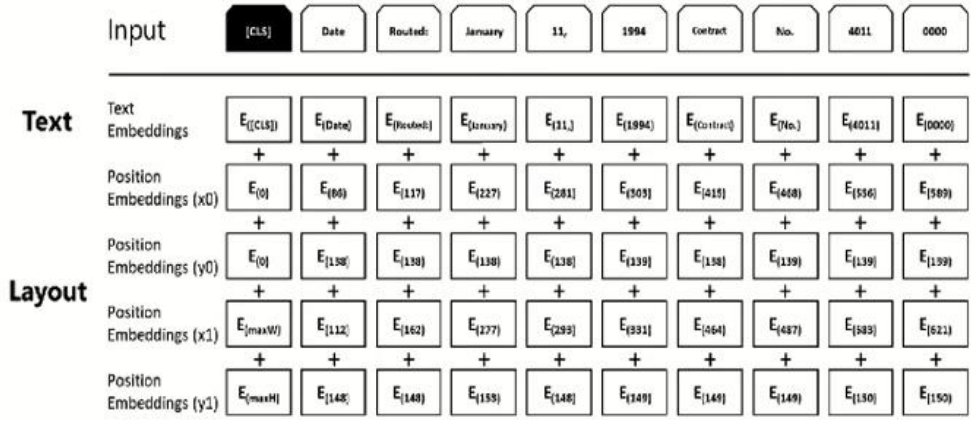

LayoutLM: Text+Layout Pre-training
Downstream Tasks
FC Layers
Image Embeddings
LayoutLM Embeddings
Pre-trained LayoutLM
Text Embeddings
Position Embeddings (x0)
Position Embeddings (y0)
Position Embeddings (x1)
Position Embeddings (y1)
ROI
Faster R-CNN
Pre-built OCR/ PDF Parser
https://arxiv.org/abs/1912.13318, KDD 2020


SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET
Date Routed: January 11, 1994
Contract No. 4011 00 00
OCR
Image
Token
Bounding Box (x0,y0,x1,y1)
Date Date 86 138 112 148
Routed Routed 117 138 162 148
January January 227 138 277 153
11 11, 281 138 293 148
1994 1994 303 139 331 149
Contract Contract 415 138 464 149
No. No. 468 139 487 149
4011 4011 556 139 583 150
00 00 0000 589 139 621 150
Input
Tok 0
Tok 1
Tok 2
Tok 3
Tok 4
Tok 5
Tok 6
Tok 7
Token Embeddings
Position Embeddings (x0)
Position Embeddings (y0)
Position Embeddings (x1)
Position Embeddings (y1)

Training Task#1: Mask Visual-Language Model



Training Task#2: Document Image Classification



Form Understanding with LayoutLM

[Task] Sequence labeling (B-I-O class labels) for key-value from forms
[Data] 149 training, 50 testing
[Metric] Precision, Recall, F1
[Baseline] Pre-trained BERT and RoBERTa

FUNSD: Form Understanding in Noisy Scanned Documents
https://guillaumejaume.github.io/FUNSD/

Results on Form Understanding

Modality	Model	Precision	Recall	F1	Parameters
Text only	$BERT_{BASE}$	0.547	0.671	0.603	110M
	$RoBERTa_{BASE}$	0.632	0.69	0.66	125M
	$BERT_{LARGE}$	0.611	0.709	0.656	340M
	$RoBERTa_{LARGE}$	0.674	0.738	0.704	355M
[Base]Text+Layout	$LayoutLM_{BASE}$ (1M)	0.691	0.774	0.73	113M
	$LayoutLM_{BASE}$ (11M)	0.76	0.816	0.787	113M
[Base]Text+Layout+Image	$LayoutLM_{BASE}$ (1M)	0.71	0.782	0.744	160M
	$LayoutLM_{BASE}$ (11M)	0.768	0.82	0.793	160M
[Large]Text+Layout	$LayoutLM_{LARGE}$ (1M)	0.717	0.805	0.759	343M
	$LayoutLM_{LARGE}$ (11M)	**0.793**	**0.854**	**0.822**	343M
[Large]Text+Layout+Image	On-going	-	-	-	-

Pre-training data: IIT-CDIP Test Collection 1.0 https://ir.nist.gov/cdip/
Fine-tuning data: FUNSD dataset with 149 forms, test on 50 forms

Results on Receipt Understanding

Modality	Model	Precision	Recall	F1	Parameters
Text only	$BERT_{BASE}$	0.9099	0.9099	0.9099	110M
	$RoBERTa_{BASE}$	0.9107	0.9107	0.9107	125M
	$BERT_{LARGE}$	0.9200	0.9200	0.9200	340M
	$RoBERTa_{LARGE}$	0.9280	0.9280	0.9280	355M
[Base]Text+Layout	$LayoutLM_{BASE}$ (1M)	0.9380	0.9380	0.9380	113M
	$LayoutLM_{BASE}$ (11M)	0.9438	0.9438	0.9438	113M
[Base]Text+Layout+Image	$LayoutLM_{BASE}$ (1M)	0.9416	0.9416	0.9416	160M
	$LayoutLM_{BASE}$ (11M)	0.9467	0.9467	0.9467	160M
[Large]Text+Layout	$LayoutLM_{LARGE}$ (1M)	0.9416	0.9416	0.9416	343M
	$LayoutLM_{LARGE}$ (11M)	**0.9604**	**0.9604**	**0.9604**	343M
Baseline	Ranking 1st in ICDAR 2019	0.9402	0.9402	0.9402	-

ICDAR 2019 Robust Reading Challenge on Key Information Extraction from Scanned Receipts
626 training, 347 testing, https://rrc.cvc.uab.es/?ch=13&com=tasks

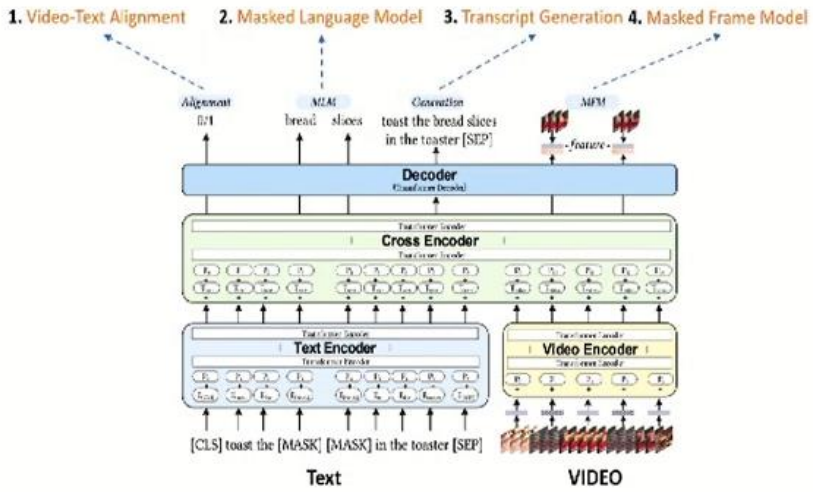
Results on Document Image Classification

Modality	Model	ACC	Parameters
Text only	$BERT_{BASE}$	89.81	110M
	$RoBERTa_{BASE}$	90.06	125M
	$BERT_{LARGE}$	89.92	340M
	$RoBERTa_{LARGE}$	90.11	355M
[Base]Text+Layout	$LayoutLM_{BASE}$ (1M)	91.48	113M
	$LayoutLM_{BASE}$ (11M)	91.78	113M
[Base]Text+Layout+Image	$LayoutLM_{BASE}$ (1M)	94.31	160M
	$LayoutLM_{BASE}$ (11M)	**94.42**	160M
[Large]Text+Layout	$LayoutLM_{LARGE}$ (1M)	91.88	343M
	$LayoutLM_{LARGE}$ (11M)	91.90	343M
Previous SOTA	LadderNet (Sarkhel & Nandi, 2019)	92.77	-
	EfficientNetB4 (DeGange, 2019)	92.81	-
	MultiModal (Dauphinee et al., 2019)	93.07	-

RVL-CDIP dataset (320K training, 40K validation, 40K testing), https://www.cs.cmu.edu/~aharley/rvl-cdip/

Video-Language Pre-training

VideoBERT (Sun al., 2019); CBT (Sun et al., 2019); Unicoder-VL (Luo et al., 2020)



Video-Language Datasets

Pre-training Data:

1. HowTo100M: 136M video clips with captions from 1.2M Youtube videos.

(Research) Evaluation Data:

1. YouCook2: 2000 long untrimmed videos from 89 cooking recipes.
2. MSR-VTT: 10K web video clips with 41.2 hours and 200K clip-sentence pairs

https://www.di.ens.fr/willow/research/howto100m/

http://youcook2.eecs.umich.edu/

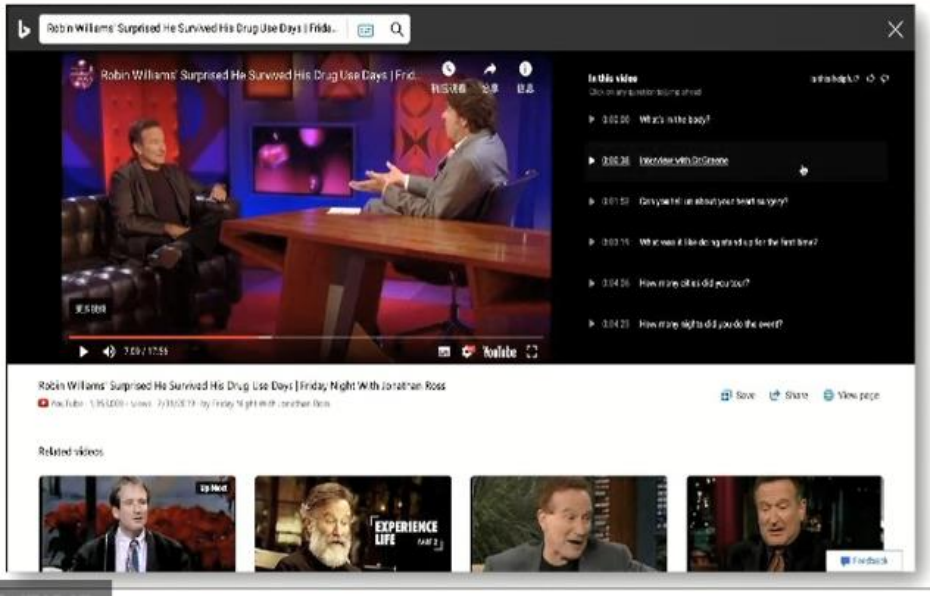
Evaluation Results

Unicoder-VL beats results from Google (CBT) and Baidu (ActBERT) on video retrieval and video captioning respectively.

Task	Video Retrieval						Video Captioning					
	YouCook2			MSR-VTT			YouCook2 (w/o transcript)			YouCook2 (w/ transcript)		
	R@1	R@5	R@10	R@1	R@5	R@10	BLEU-4	R-L	CIDEr	BLEU-4	R-L	CIDEr
SoTA	9.6	26.7	38.0	8.6	23.4	33.1	5.4	30.6	0.6	9.0	36.7	1.1
Unicoder-VL	10.0	27.5	38.8	15.4	39.5	52.3	6.1	31.5	0.6	10.4	38.0	1.2
Δ	0.4 ↑	0.8 ↑	0.8 ↑	6.8 ↑	16.1 ↑	19.2 ↑	0.7 ↑	0.9 ↑	-	1.4 ↑	1.3 ↑	0.1 ↑

Blue numbers indicates the best result for a task.

Application: Video Chaptering



Summary of Multimodal Pre-trained Models

Multimodal pre-trained models are in the early stage

- Use Transformer as backbone following NLP pre-trained models
- Learn joint representations from limited existing vision-language datasets
- Achieve state-of-the-art results on downstream tasks, but still far from practical

Multimodal pretrained models still have many challenges

- The size of image/video-language dataset is smaller compared with NLP dataset
- CV models are not jointly trained with NLP models
- Object categories detected by CV is limited and mistakes propagate to pre-trained model training
- Hard to train deep transformers due to huge cost for training

Summary and Future Work

Pretrained models in NN-NLP

- Full set of NN-NLP: End-to-End training with large scale data and big models including expression learning, encoder-decoder, attention model, transformer
- Transfer-learning with pre-trained models from language, to multi-lingual and multi-modality tasks

In the future, we will need to further explore:

- New pre-training tasks, new model architectures, for more modalities, with smaller model sizes and faster training
- How to empower with interpretation mechanisms to NN-NLP?

相关阅读：

CCAI 2020 大会特邀报告

CCAI 2020 大会特邀报告(II): Thoughtful Artificial Intelligence

临菲信息技术港

临菲信息技术港公众号

临菲学堂