

语音大模型的若干思考与猜测

欧智坚

清华大学电子工程系 (EE)

清华大学语音与智能实验室 (THU-SPMI)

它思科技 (TasiTech)

2023 SpeechHome语音技术研讨会, 2023/11/19



内容安排

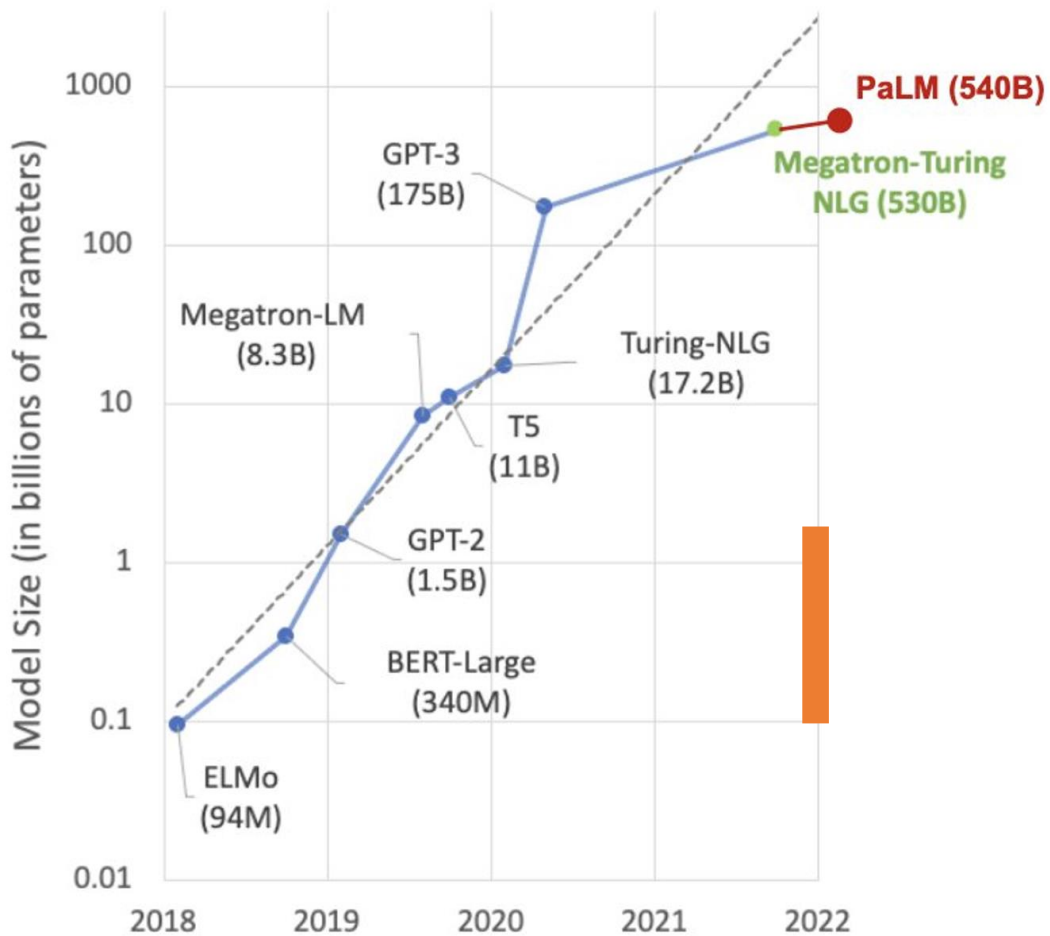
一、引言

二、语音大模型的若干思考

三、总结

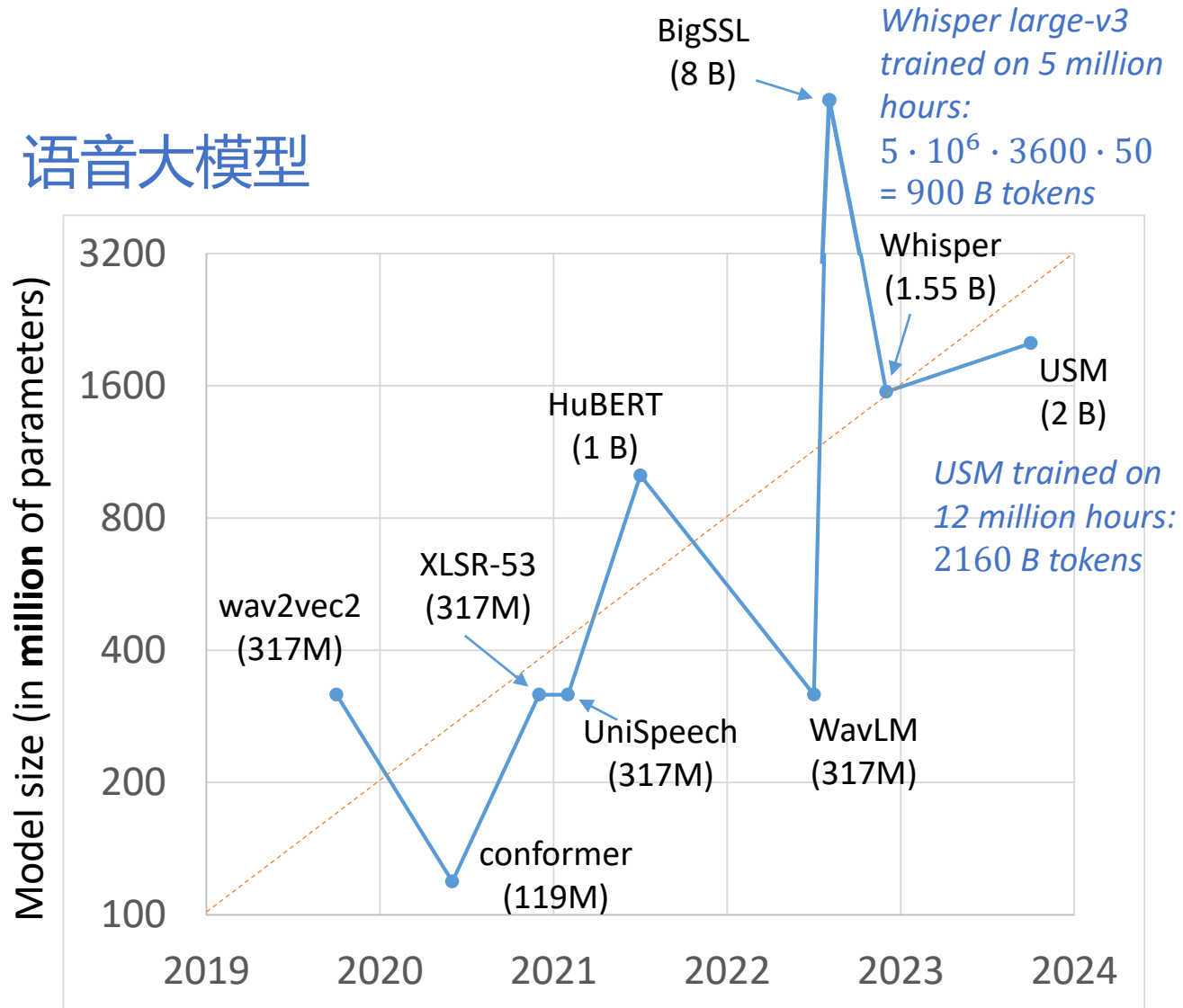
大模型在目前AI研究中备受关注!

语言大模型



© Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, 2022 Feb.

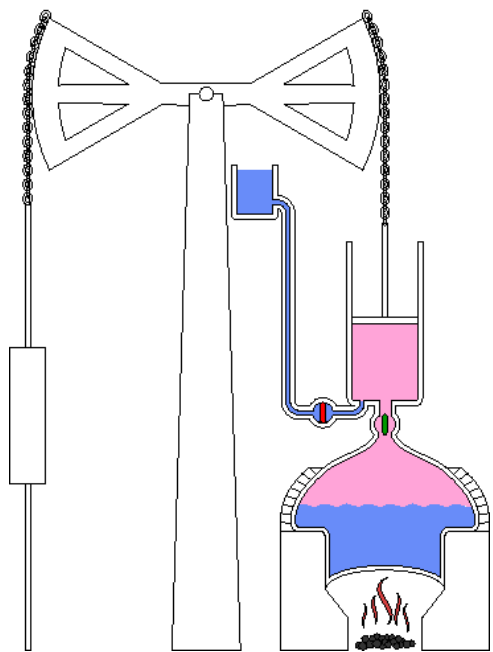
语音大模型



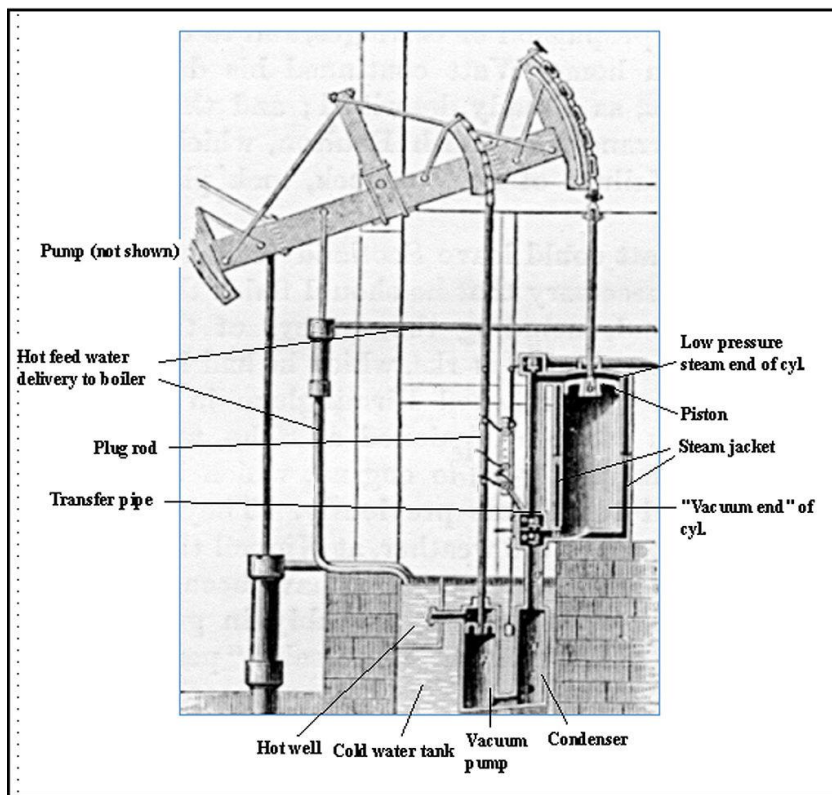
© 欧智坚, 2023 Nov.

大模型技术引发新工业革命？

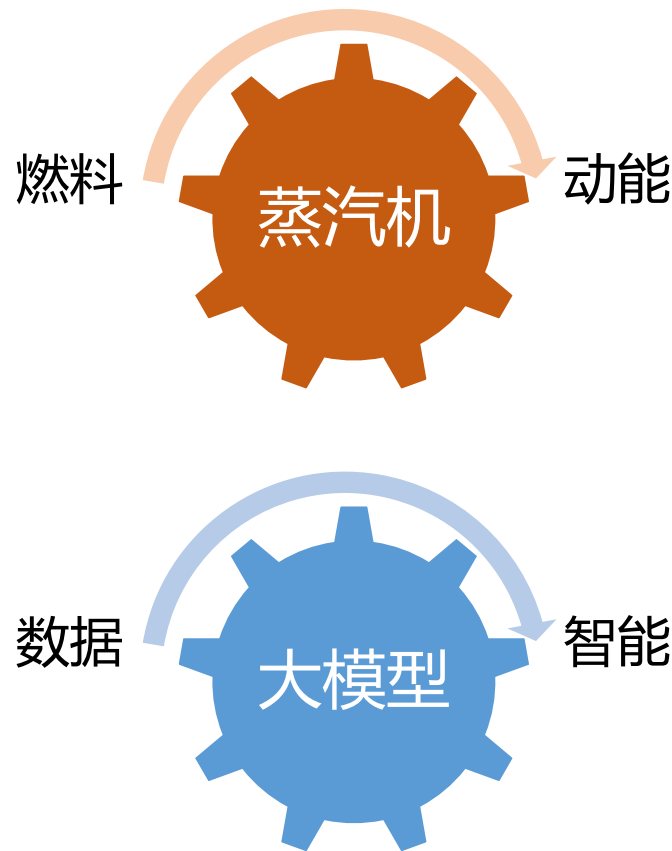
- 瓦特的分离式冷凝机，燃料->动能，引发了工业革命。
- 大模型从数据中学习获得智能，什么制约了AI技术规模化应用？



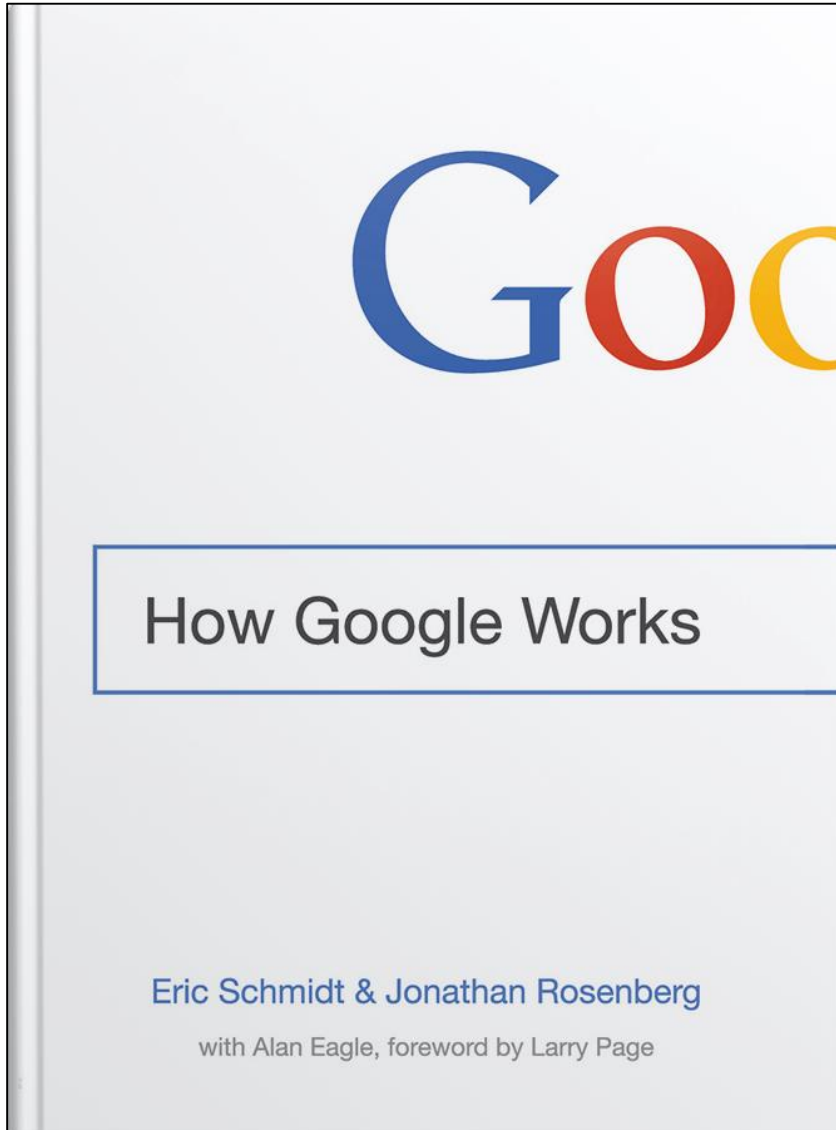
纽科门蒸汽机 (1712)



瓦特的分离式冷凝机 (1781)



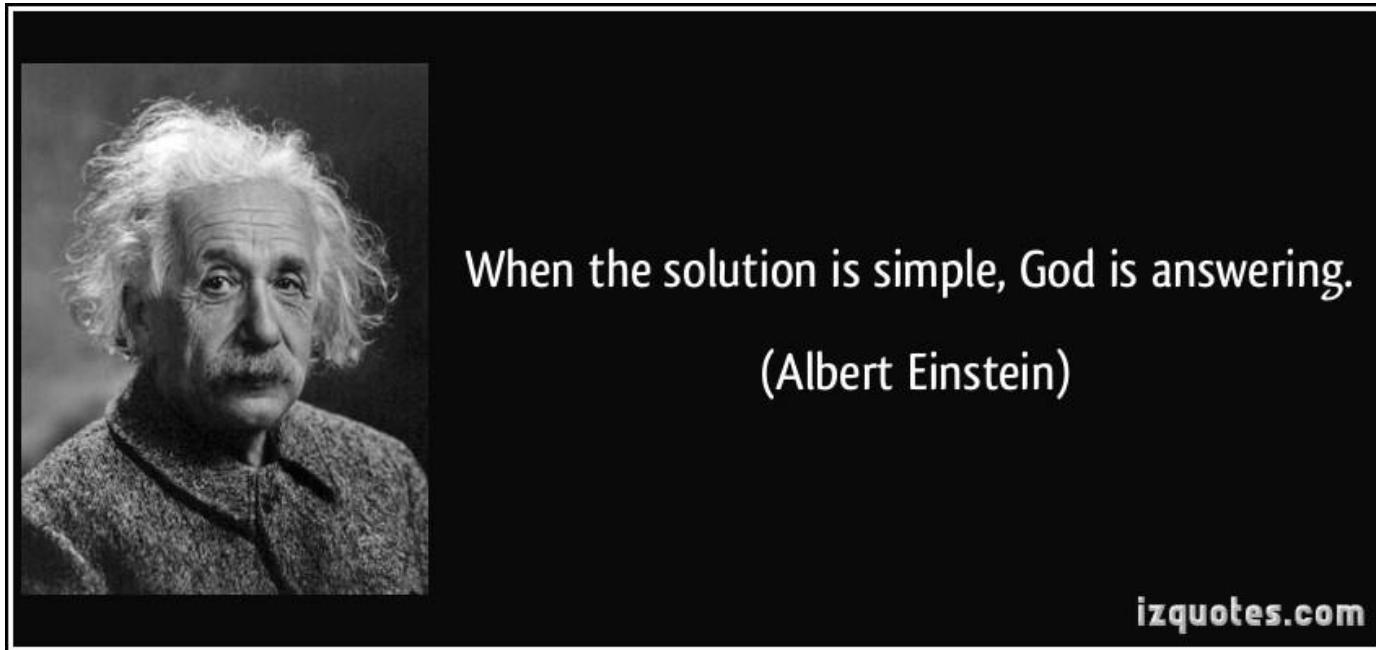
Motivation



... to think from **first principles** and real-world physics rather than having to accept the prevailing “wisdom”.

— Larry Page, 2014

未来之路 Path to the future



S³: Sound, Simple, Scalable
扎实, 简洁, 可伸缩

墨菲定律
Anything that can go wrong will go wrong.

奥卡姆剃须刀
若无必要, 勿增实体.

可伸可缩
在不同尺度上都能工作,
神经网络解决XOR问题

内容安排

一、引言

二、语音大模型的若干思考 from first principles

- Principled **unsupervised learning**, yes/no, how?
- End-to-end is all you need (for supervised learning)?
 - **AM and LM** fusion, yes/no, how?
 - **Multi-lingual** ASR needs phonetic knowledge or not, how?

三、总结

语音大模型的若干猜测

匿名填写

*01 未来语音大模型是否需要“有原则的无监督学习”?

- 是
- 否
- 不确定

*02 未来语音识别大模型是否需要“AM与LM融合”?

- 是
- 否
- 不确定

*03 未来多语言语音识别大模型是否需要“语音学知识指导建模”?

- 是
- 否
- 不确定

*04 未来3年能否实现语音大模型的“有原则的无监督学习”?

- 能
- 不好说
- 不需要



语音大模型的若干猜测

扫一扫二维码打开或分享...



内容安排

一、引言

二、语音大模型的若干思考 from first principles

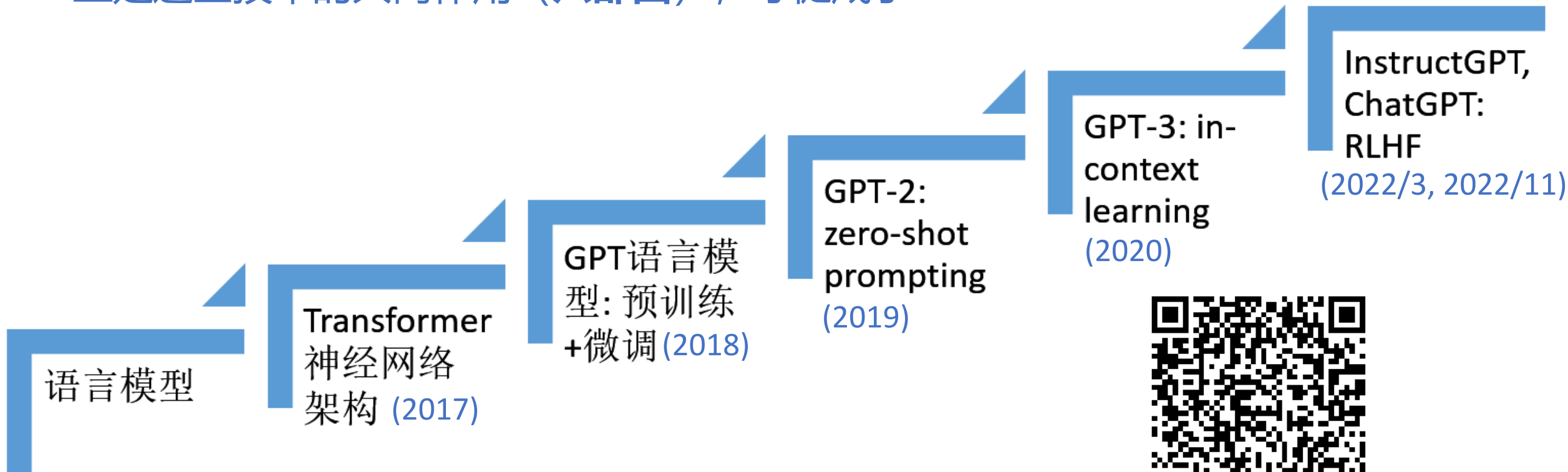
- **Principled unsupervised learning, yes/no, how?**
- End-to-end is all you need (for supervised learning)?
 - **AM and LM** fusion, yes/no, how?
 - **Multi-lingual** ASR needs phonetic knowledge or not, how?

三、总结

ChatGPT的进步

站在多年来人工智能研究的巨人肩膀上

正是这些技术的共同作用（**六部曲**），才促成了ChatGPT



“严谨谈谈ChatGPT的进步、不足及AGI挑战”
欧智坚，2023/3/16

语言模型 (LM, language model)

语言模型：人类自然语言的概率模型

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1})$$

当前符号 前面历史符号

自回归语言模型：

每个位置利用前面历史符号 x_1, x_2, \dots, x_{i-1} (即上文) ,

计算当前符号出现的(条件)概率 $P(x_i | x_1, \dots, x_{i-1})$

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

The best thing about AI is its ability to,

The best thing about AI is its ability to learn,

The best thing about AI is its ability to learn from,

The best thing about AI is its ability to learn from experience,

The best thing about AI is its ability to learn from experience.,

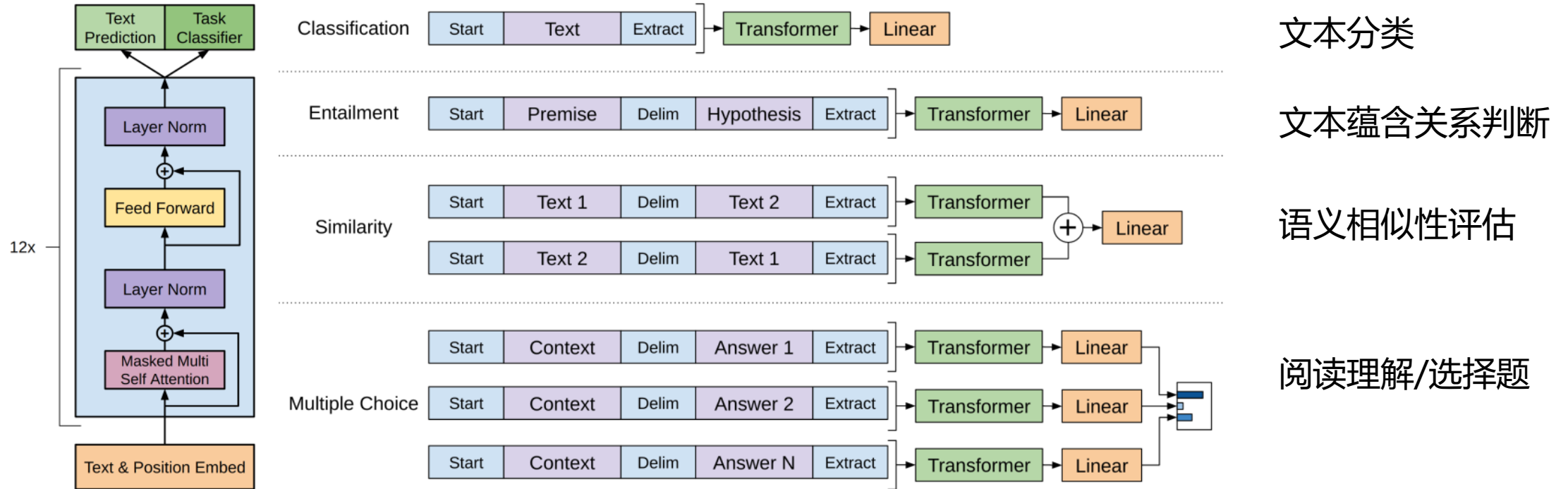
The best thing about AI is its ability to learn from experience. It,

The best thing about AI is its ability to learn from experience. It's,

The best thing about AI is its ability to learn from experience. It's not

GPT语言模型及预训练+微调技术 (2018)

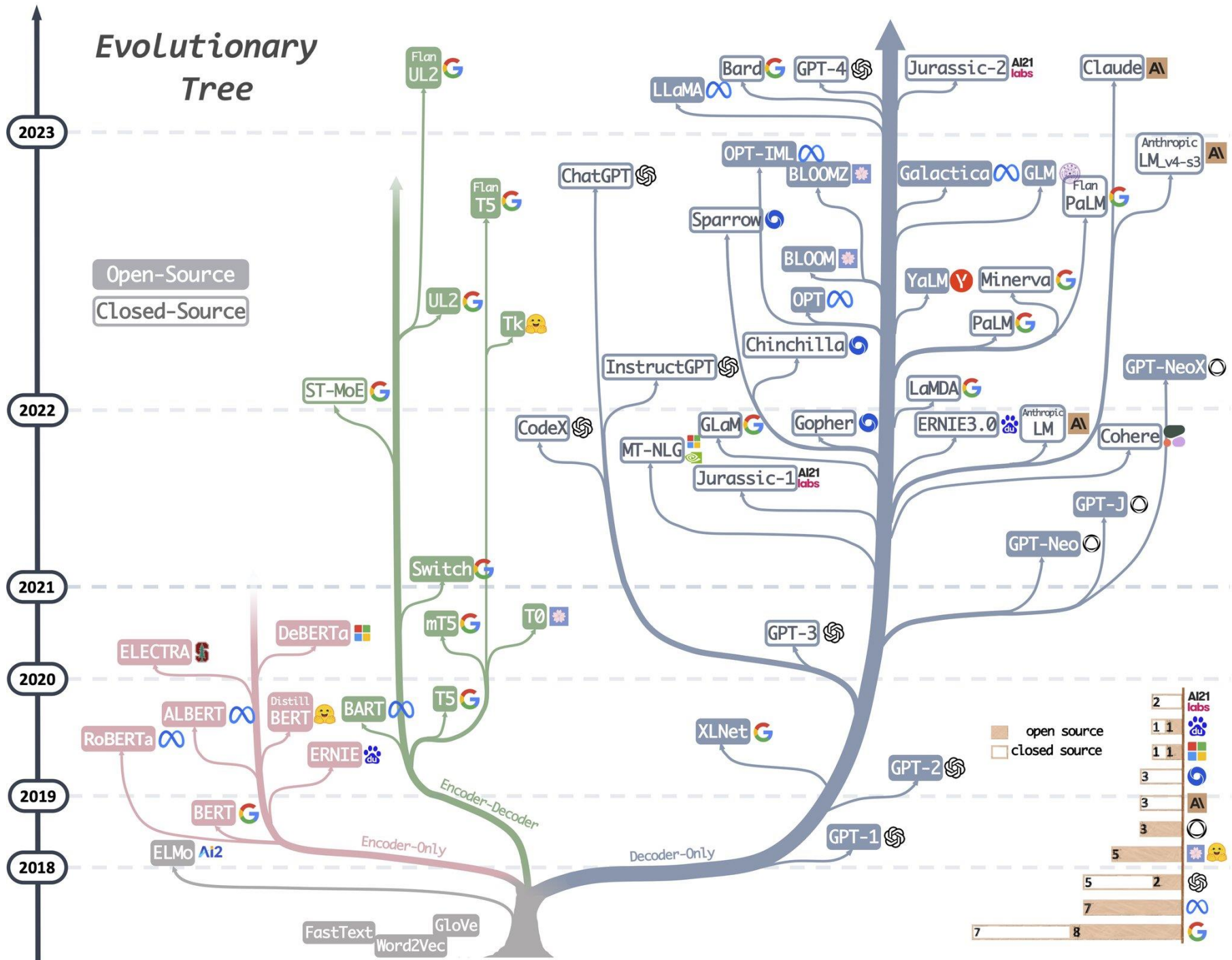
自然语言理解，包括范围广泛的不同任务



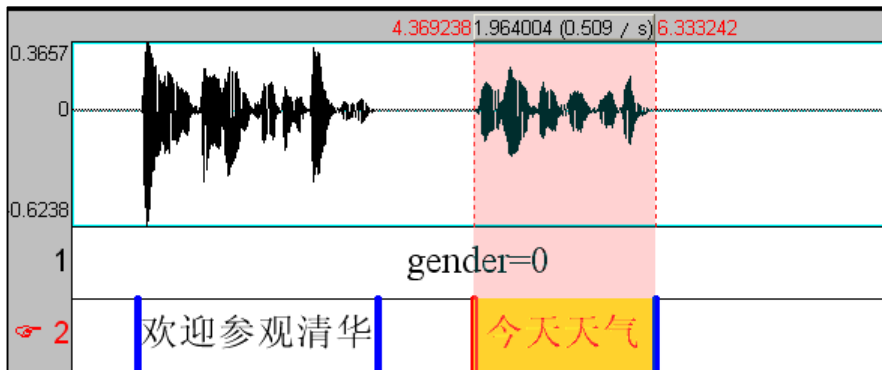
来自GPT原文[4] "Improving language understanding by **Generative Pre-Training**" (2018)

"Our work broadly falls under the category of semi-supervised learning for natural language."

无监督预训练 (unsupervised pre-training) 结合有监督微调 (supervised fine-tuning) ，
是一种**半监督学习**，其本质是协同进行**有监督学习**和**无监督学习**。



从有标数据 (x, y) 进行有监督学习, 巨大成功!

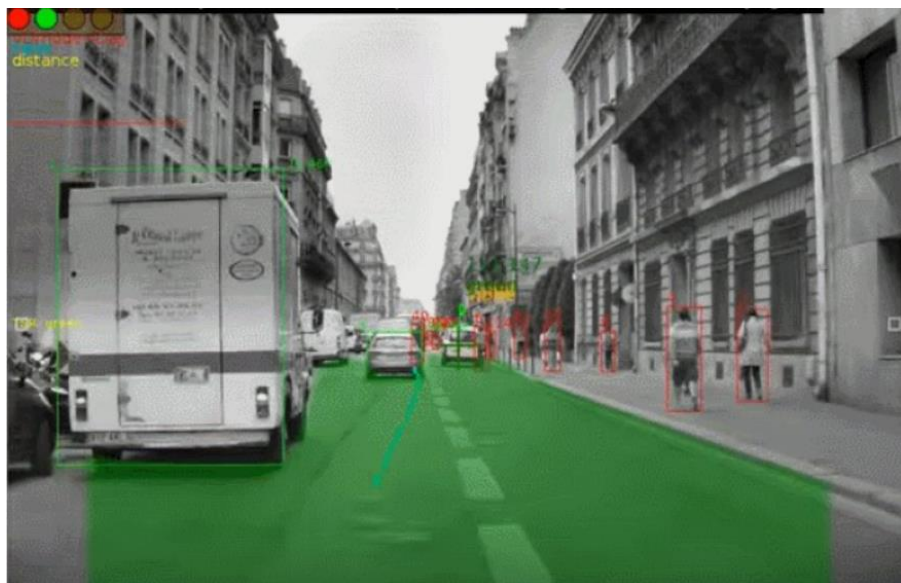


语音识别

ATIS UTTERANCE EXAMPLE IOB REPRESENTATION

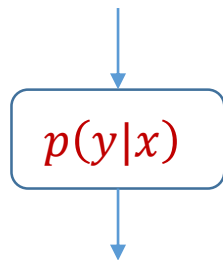
Sentence	<i>show</i>	<i>flights</i>	<i>from</i>	<i>Boston</i>	<i>To</i>	<i>New</i>	<i>York</i>	<i>today</i>
Slots/Concepts	O	O	O	B-dept	O	B-arr	I-arr	B-date
Named Entity	O	O	O	B-city	O	B-city	I-city	O
Intent	<i>Find Flight</i>							
Domain	<i>Airline Travel</i>							

意图识别, 语义填槽, 命名实体识别



目标检测与跟踪

输入: 上文 x

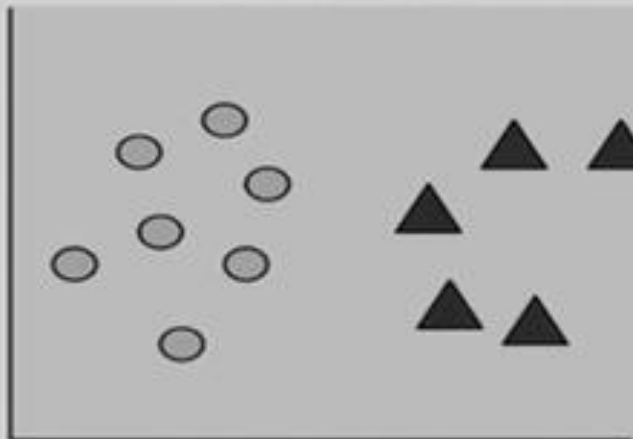


输出: 响应 y



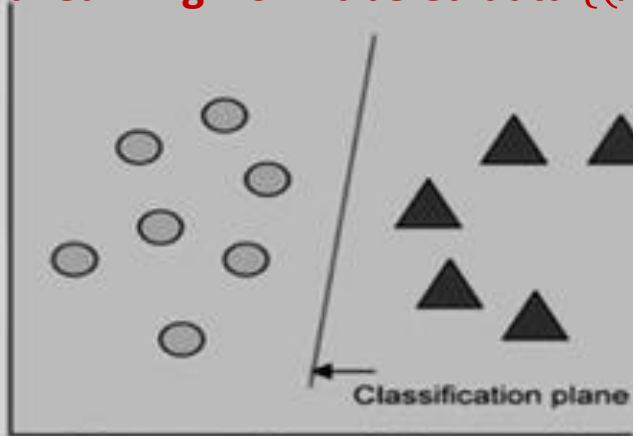
对话系统

Semi-supervised learning (SSL)

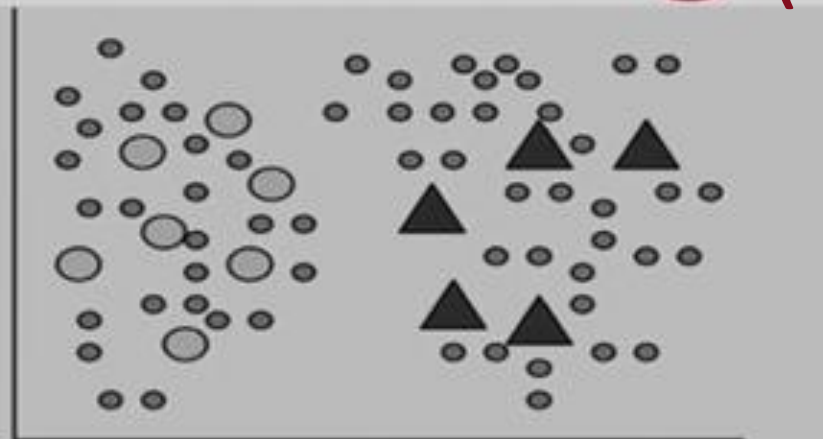


Labeled Data
(a)

Supervised learning from labeled data $\{(x, y)\}$

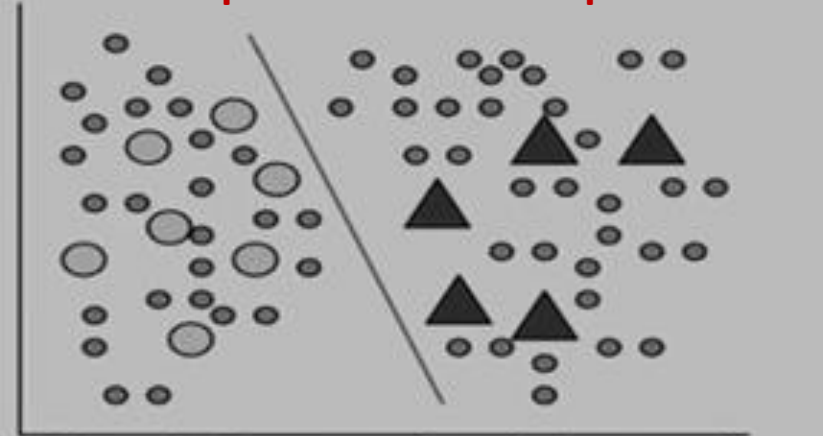


Supervised Learning
(c)



Labeled and Unlabeled Data
(b)

Collaborative supervised and unsupervised learning



Semi-Supervised Learning
(d)

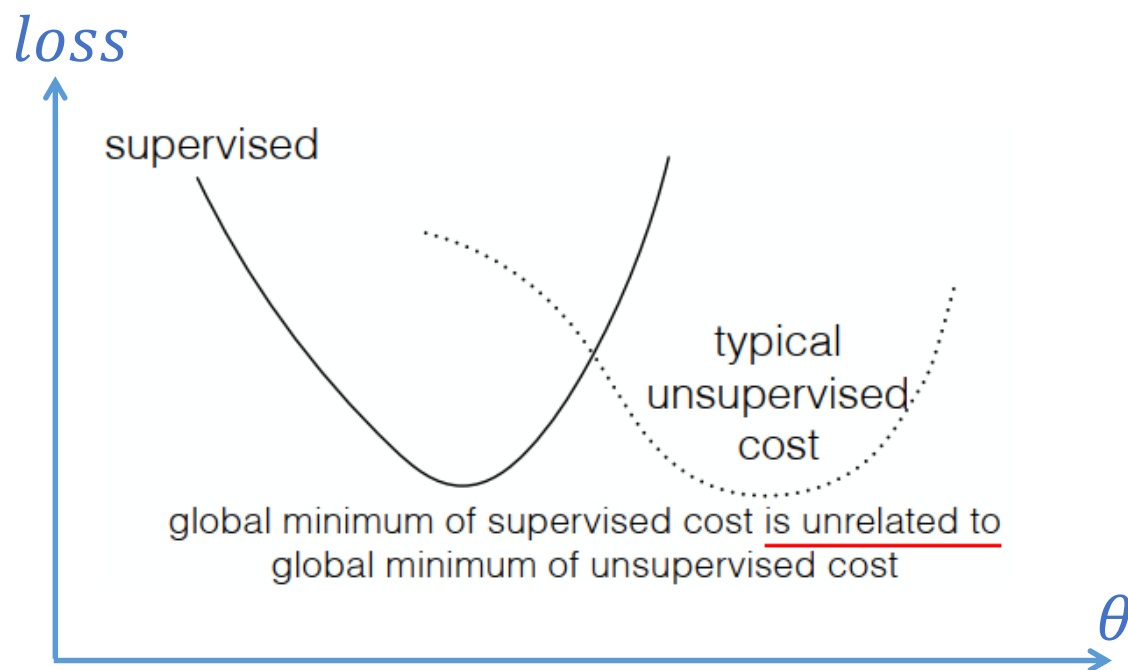


TOWARDS PRINCIPLED UNSUPERVISED LEARNING

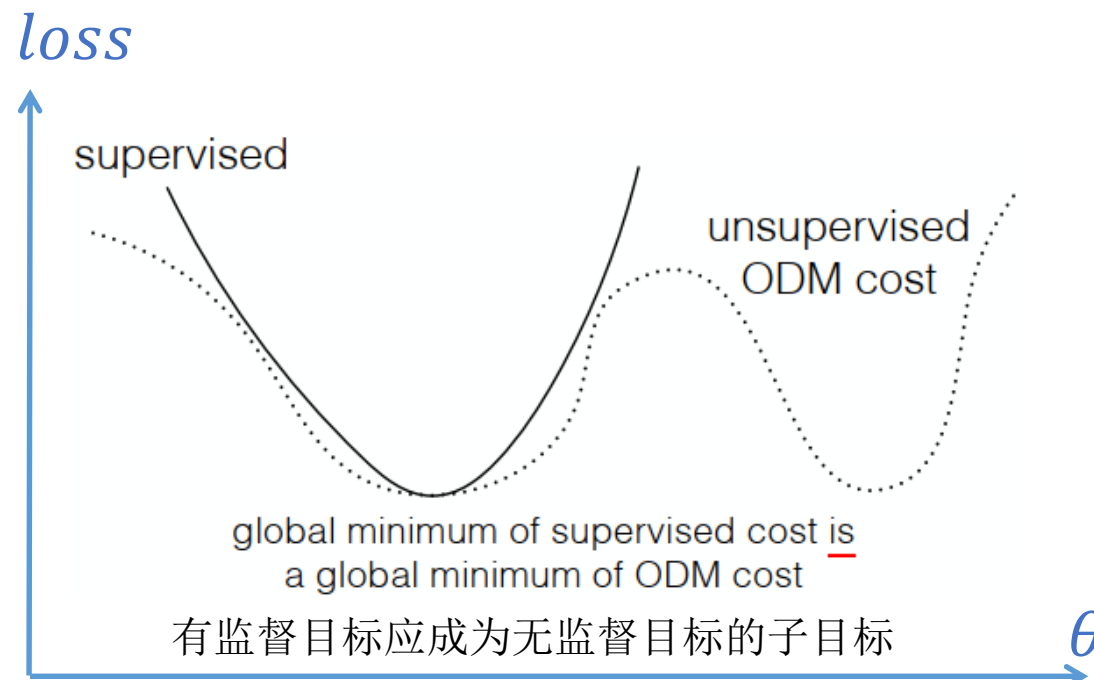
Ilya Sutskever¹, Rafal Jozefowicz¹, Karol Gregor², Danilo Rezende², Tim Lillicrap², Oriol Vinyals¹

Google Brain¹ and Google DeepMind²

{ilyasu, rafalj, karolg, danilor, countzero, vinyals}@google.com



朴素的无监督学习



有原则的无监督学习

Workshop

Large Language Models and Transformers

Date

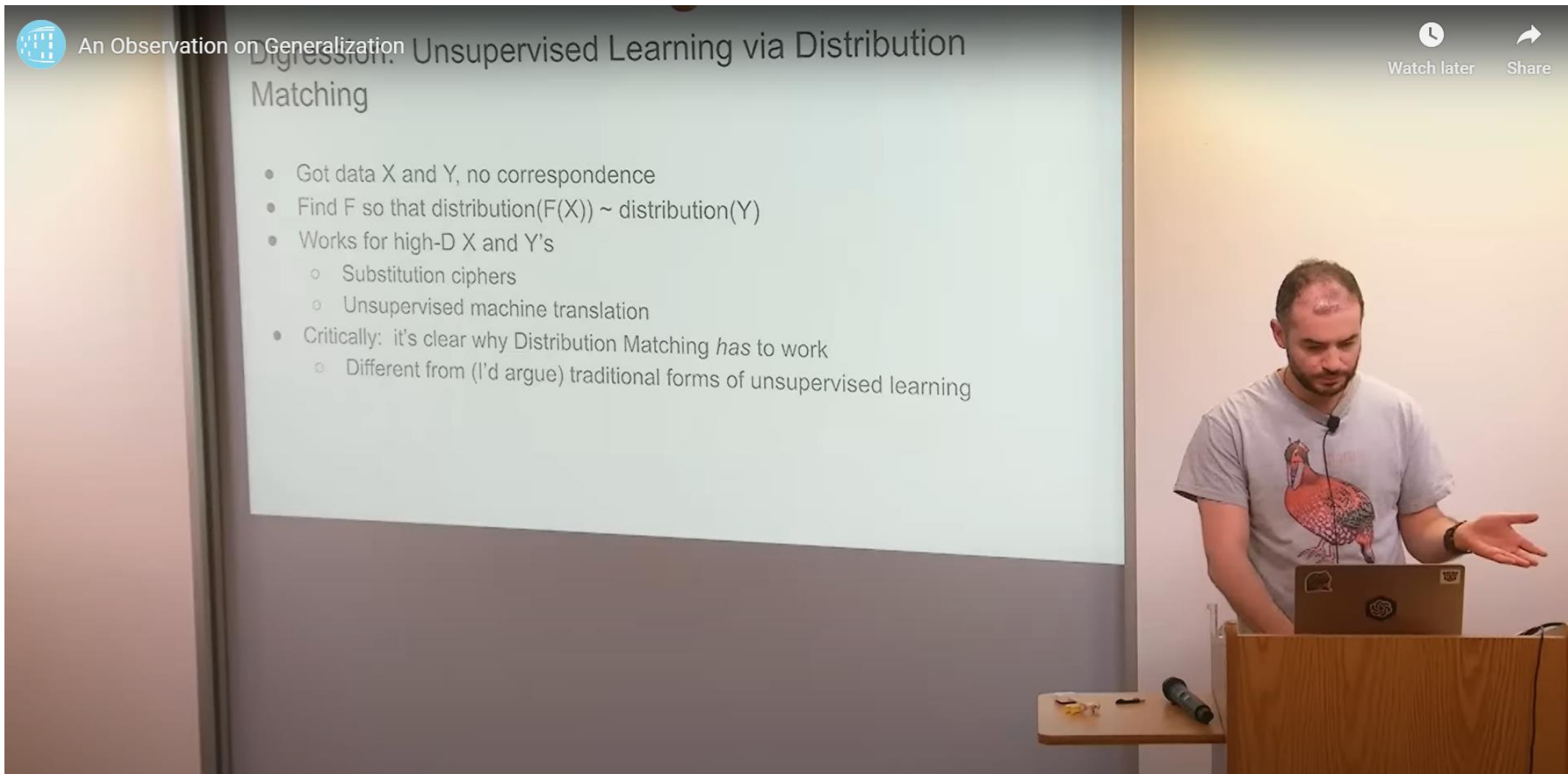
Monday, Aug. 14, 2023

Speaker(s)

Ilya Sutskever (OpenAI)

Time

3 – 4 p.m. PT



GPT-2 (2019)

让机器去学习执行一个自然语言理解任务，本质是去估计条件分布

$$P(\text{output} \mid \text{input})$$

以GPT-2为代表的创新做法是，*task*、*input*、*output*都用自然语言来表述成符号序列，**这样模型 $P(\text{output} \mid \text{task}, \text{input})$ 就归结为一个语言模型**——给定上文，递归生成下一个符号。不同任务的训练数据都统一组织成

$$\text{task}, \text{input}, \text{output}$$

比如，

(translate to french, english text, french text)

(answer the question, document, question, answer)

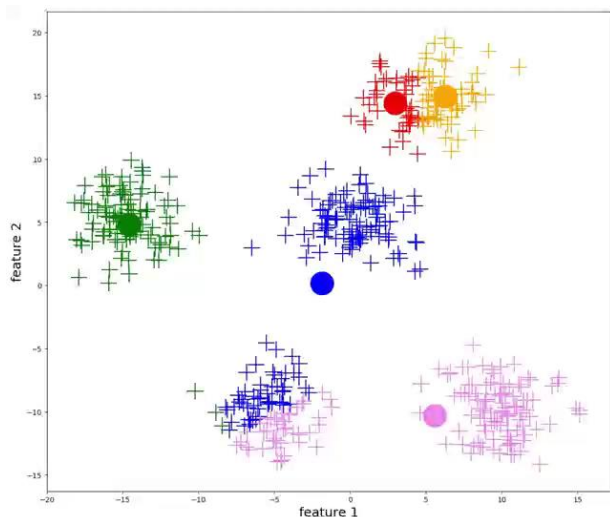
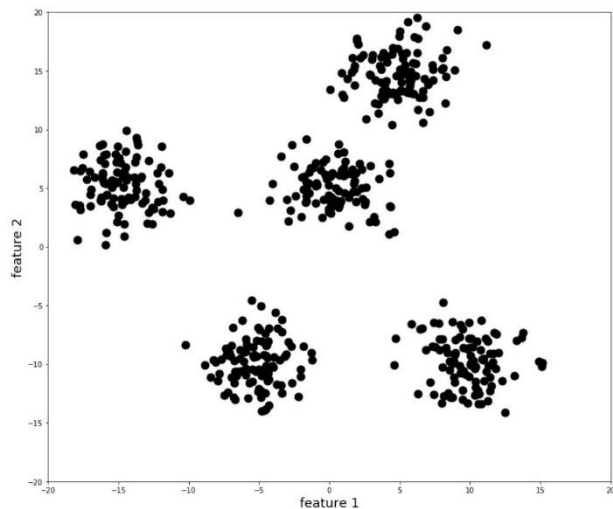
GPT实现了
principled unsupervised learning

dicted. Since the supervised objective is the the same as the unsupervised objective but **only evaluated on a subset** of the sequence, the global minimum of the unsupervised objective is also the global minimum of the supervised objective. In this slightly toy setting, the concerns with density estimation as a principled training objective discussed in (Sutskever et al., 2015) are side stepped. The problem instead becomes

无监督学习：解决只有 x 时的模型训练

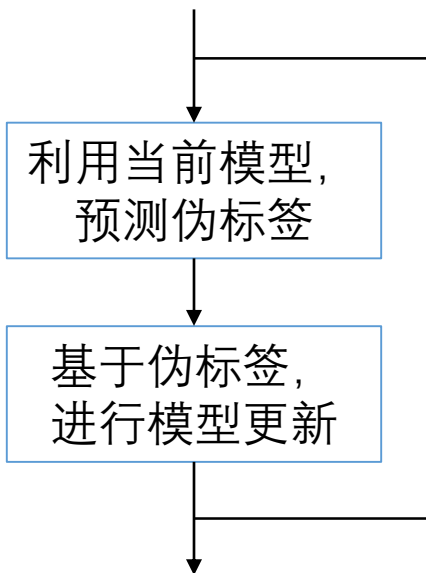


朴素：K-means聚类算法



高级：基于概率模型

- 只有 x ，如何构建出 $p_{\theta}(y|x)$ ？
- 升维打击：建模 $p_{\theta}(x, y)$ ，然后通过最大化边缘似然 $\log p_{\theta}(x)$ 去估计 $p_{\theta}(y|x)$
- 最大似然参数估计：在模型容量足够大时，无监督学习能找到Oracle模型 $p^*(x, y)$ ，也自然找到了 $p^*(y|x)$ 。

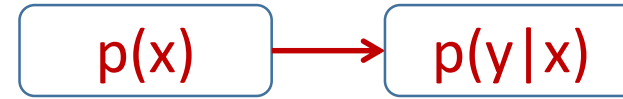


有监督目标是无监督优化时的子目标，由此实现了principled unsupervised learning.

$$\begin{array}{ccc} & p_{\theta}(x, y) & \\ \swarrow & & \searrow \\ p_{\theta}(x) & & p_{\theta}(y|x) \end{array}$$

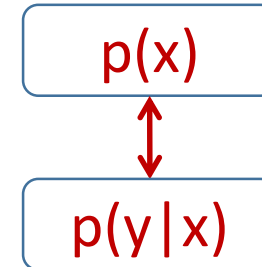
Generative SSL - Two Different Approaches

- **Pre-training (serial collaboration)**



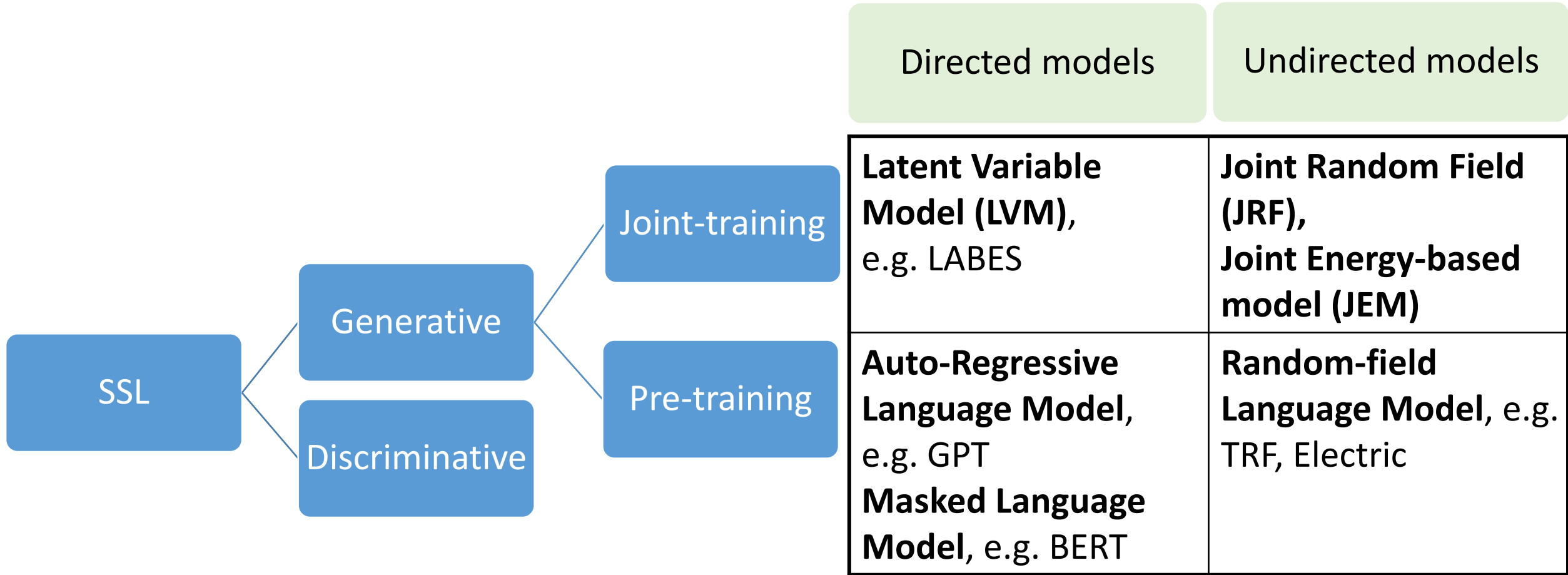
- Only defines $p(x)$ without y .
- Perform unsupervised representation learning (called **pre-training**) on unlabeled data, followed by supervised training (called **fine-tuning**) on labeled data.
- This manner of pre-training followed by fine-tuning has received increasing application in Natural Language Processing.

- **Joint-training (parallel collaboration)**



- A joint model of $p(x,y)$ is defined.
- When we have label y , we maximize $p(y|x)$ (the supervised objective), and when the label is unobserved, we marginalize it out and maximize $p(x)$ (the unsupervised objective).
- Semi-supervised learning over a mix of labeled and unlabeled data is formulated as maximizing the (weighted) sum of $\log p(y|x)$ and $\log p(x)$.

There are many open questions in designing semi-supervised methods for particular tasks !



[LABES] Y. Zhang, Z. Ou, et al. A Probabilistic End-To-End Task-Oriented Dialog Model with Latent Belief States towards Semi-Supervised Learning. EMNLP, 2020.

[JRF] Y. Song, Z. Ou, et al. Upgrading CRFs to JRFs and its benefits to sequence modeling and labeling. ICASSP, 2020.

[JEM] S. Zhao, J.H. Jacobsen, et al. Joint energy-based models for semi-supervised classification. ICML Workshop on Uncertainty and Robustness in Deep Learning, 2020.

[TRF] B. Wang, Z. Ou. Improved training of neural trans-dimensional random field language models with dynamic noise-contrastive estimation. SLT, 2018.

[Electric] K. Clark, M.T. Luong, et al. Pre-Training Transformers as Energy-Based Cloze Models. EMNLP, 2020.

Joint-training of an EBM for semi-supervised image classification

- **Joint modeling** of observation $x \in \mathbb{R}^d$ and class label $y \in \{1, \dots, K\}$:

$$p_{\theta}(x, y) = \frac{1}{Z(\theta)} \exp[u_{\theta}(x, y)]$$

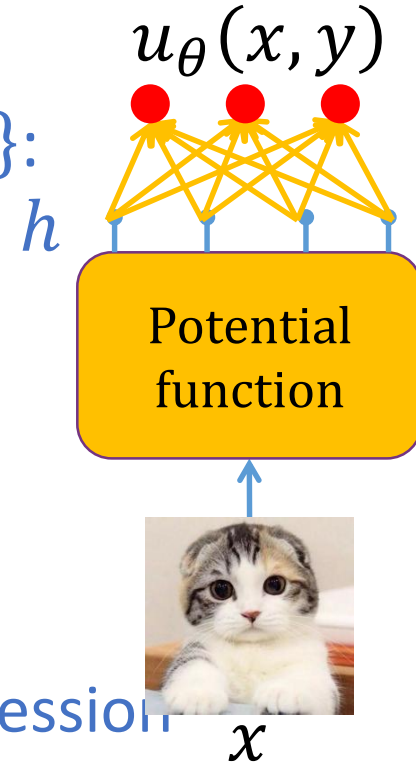
- Consider a NN $\Psi_{\theta}(x): \mathbb{R}^d \rightarrow \mathbb{R}^K$ and define:

$$u_{\theta}(x, y) = \Psi_{\theta}(x)[y]$$

- **Classifier:** $p_{\theta}(y|x) = \frac{p_{\theta}(x, y)}{p_{\theta}(x)} = \frac{\exp[u_{\theta}(x, y)]}{\sum_y \exp[u_{\theta}(x, y)]}$, like a K -class logistic regression

Marginal density: $p_{\theta}(x) = \frac{1}{Z(\theta)} \exp[u_{\theta}(x)]$, where $u_{\theta}(x) \triangleq \log \sum_y \exp[u_{\theta}(x, y)]$

$$\max_{\theta} \sum_{x_i \sim \text{unlabeled}} \log p_{\theta}(x_i) + \sum_{(x_j, y_j) \sim \text{labeled}} \log p_{\theta}(y_j | x_j)$$



Learning Neural Random Fields with Inclusive Auxiliary Generators

Yunfu Song, Zhijian Ou

In this paper we develop Neural Random Field learning with Inclusive-divergence minimized Auxiliary Generators (NRF-IAG), which is underappreciated in the literature. The contributions are two-fold. First, we rigorously apply the stochastic approximation algorithm to solve the joint optimization and provide theoretical justification. The new approach of learning NRF-IAG achieves superior unsupervised learning performance competitive with state-of-the-art deep generative models (DGMs) in terms of sample generation quality. Second, semi-supervised learning (SSL) with NRF-IAG gives rise to strong classification results comparable to state-of-art DGM-based SSL methods, and simultaneously achieves superior generation. This is in contrast to the conflict of good classification and good generation, as observed in GAN-based SSL.

Published as a conference paper at ICLR 2020

YOUR CLASSIFIER IS SECRETLY AN ENERGY BASED MODEL AND YOU SHOULD TREAT IT LIKE ONE

Will Grathwohl

University of Toronto & Vector Institute
Google Research
wgrathwohl@cs.toronto.edu

Kuan-Chieh Wang* & Jörn-Henrik Jacobsen*

University of Toronto & Vector Institute
wangkual@cs.toronto.edu
j.jacobsen@vectorinstitute.ai

David Duvenaud

University of Toronto & Vector Institute
duvenaud@cs.toronto.edu

Kevin Swersky & Mohammad Norouzi

Google Research
{kswersky, mnorouzi}@google.com

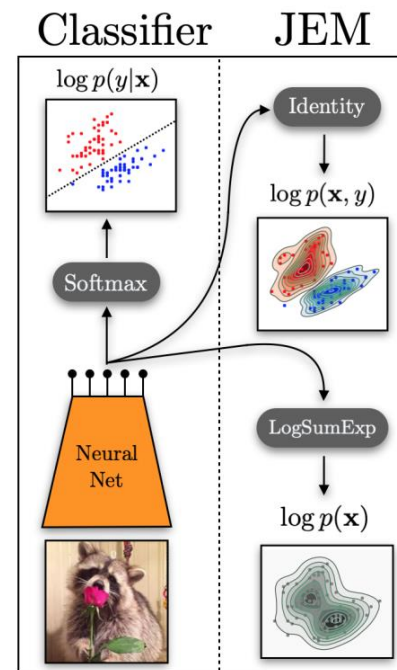
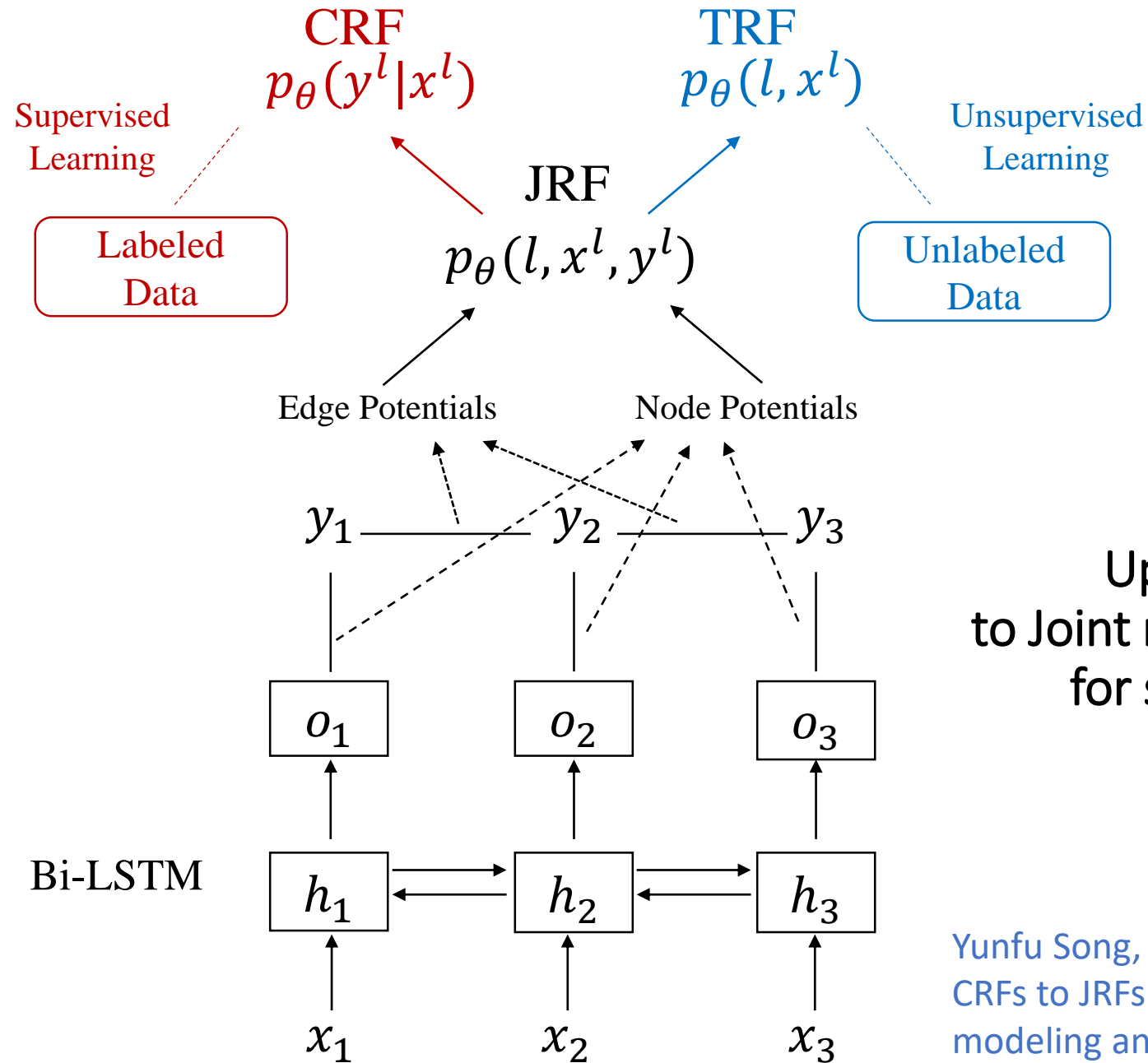


Figure 1: Visualization of our method, JEM, which defines a joint EBM from classifier architectures.

Joint-training of an EBM for semi-supervised natural language labeling



Upgrading CRFs
to Joint random fields (JRFs)
for sequential data

Table 2. SSL for image classification over CIFAR-10 with 4,000 labels. The upper/lower blocks show generative/discriminative SSL methods respectively. The means and standard deviations are calculated over ten independent runs with randomly sampled labels.

Methods	error (%)
CatGAN [30]	19.58±0.46
Ladder network [31]	20.40±0.47
Improved-GAN [32]	18.63±2.32
BadGAN [33]	14.41±0.30
Sobolev-GAN [34]	15.77±0.19
Supervised baseline	25.72±0.44
Pre-training+fine-tuning EBM	21.40±0.38
Joint-training EBM	15.12±0.36
Results below this line cannot be directly compared to those above.	
VAT small [1]	14.87
Temporal Ensembling [2]	12.16±0.31
Mean Teacher [3]	12.31±0.28

Joint-training EBMs outperform pre-training+fine-tuning EBMs by a large margin in this task.

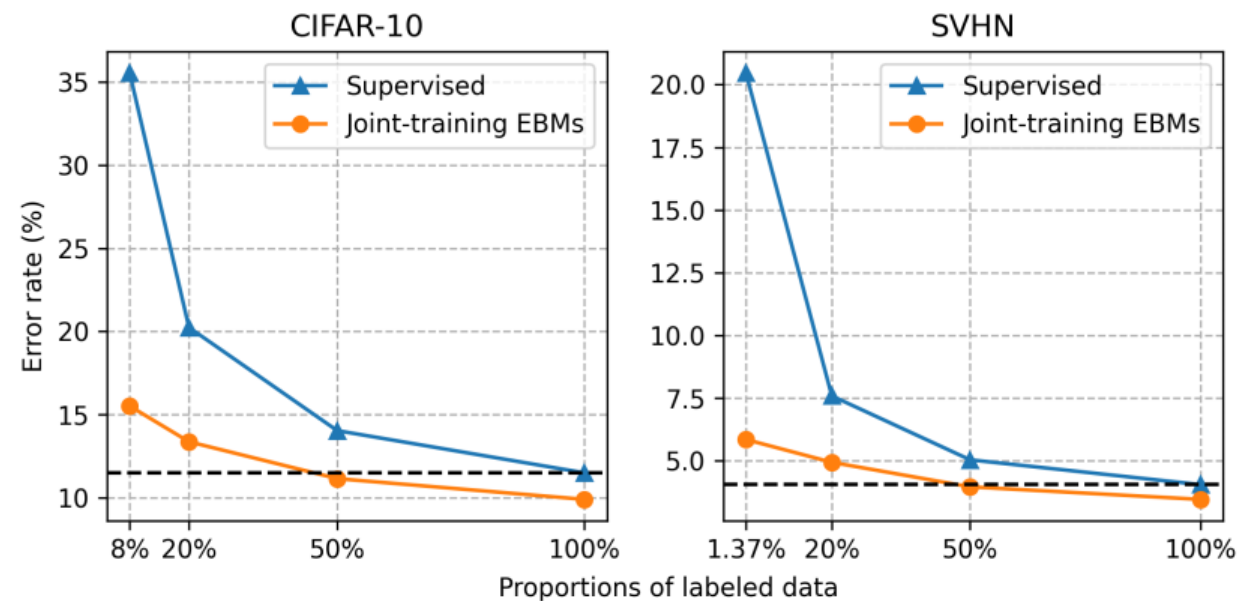


Fig. 1. Error rates of supervised baseline and joint-training EBMs as the amount of labels varies on SVHN and CIFAR-10 datasets. The dash line is the supervised result trained with 100% labeled data.

Can reduce 50% of labels without losing performance.

Table 3. Natural language labeling results. The evaluation metric is accuracy for POS and F_1 for chunking and NER. “Labeled” denotes the amount of labels in terms of the proportions w.r.t. the full set of labels. “U/L” denotes the ratio between the amount of unlabeled and labeled data. “U/L=0” denotes the supervised baseline. “pre.” and “joint” denote the results by pre-training+fine-tuning EBMs and joint-training EBMs, respectively.

Labeled	U/L	POS tagging		Chunking		NER	
		pre.	joint	pre.	joint	pre.	joint
2%	0	95.57		78.73		78.19	
	50	95.72	95.92	81.62	82.24	76.74	77.61
	250	95.96	96.13	82.10	82.26	78.49	78.51
	500	96.08	96.24	83.10	83.05	79.47	79.17
10%	0	96.81		90.06		86.93	
	50	96.87	96.99	91.60	91.85	86.37	87.05
	250	96.88	97.00	91.09	91.93	86.86	86.77
	500	96.92	97.08	91.93	92.23	87.57	87.06
100%	0	97.41		94.77		90.74	
	50	97.40	97.49	95.05	95.31	91.24	91.34
	250	97.45	97.54	95.12	95.48	91.19	91.51
	500	97.46	97.57	95.19	95.50	91.30	91.52

Table 4. Relative improvements by joint-training EBMs compared to the supervised baseline (abbreviated as sup.) and pretraining+fine-tuning EBMs respectively. Refer to Table 3 for notations.

Labeled	U/L	joint over sup.			joint over pre.		
		POS	Chunking	NER	POS	Chunking	NER
2%	50	7.9	16.5	-2.7	4.7	3.4	3.7
	250	12.6	16.6	1.5	4.2	0.9	0.1
	500	15.1	20.3	4.5	4.1	-0.3	-1.5
10%	50	5.6	18.0	0.9	3.8	3.0	5.0
	250	6.0	18.3	-1.2	3.8	9.4	-0.7
	500	8.5	21.8	1.0	5.2	3.7	-4.1
100%	50	3.1	10.3	6.5	3.5	5.3	1.1
	250	5.0	13.6	8.3	3.5	7.4	3.6
	500	6.2	14.0	8.4	4.3	6.4	2.5

- Joint-training EBMs outperform pre-training EBMs in 23 out of the 27 settings marginally but nearly consistently.
- A possible explanation: the optimization of **joint-training** is directly related to the targeted task, while **pre-training** does not.

内容安排



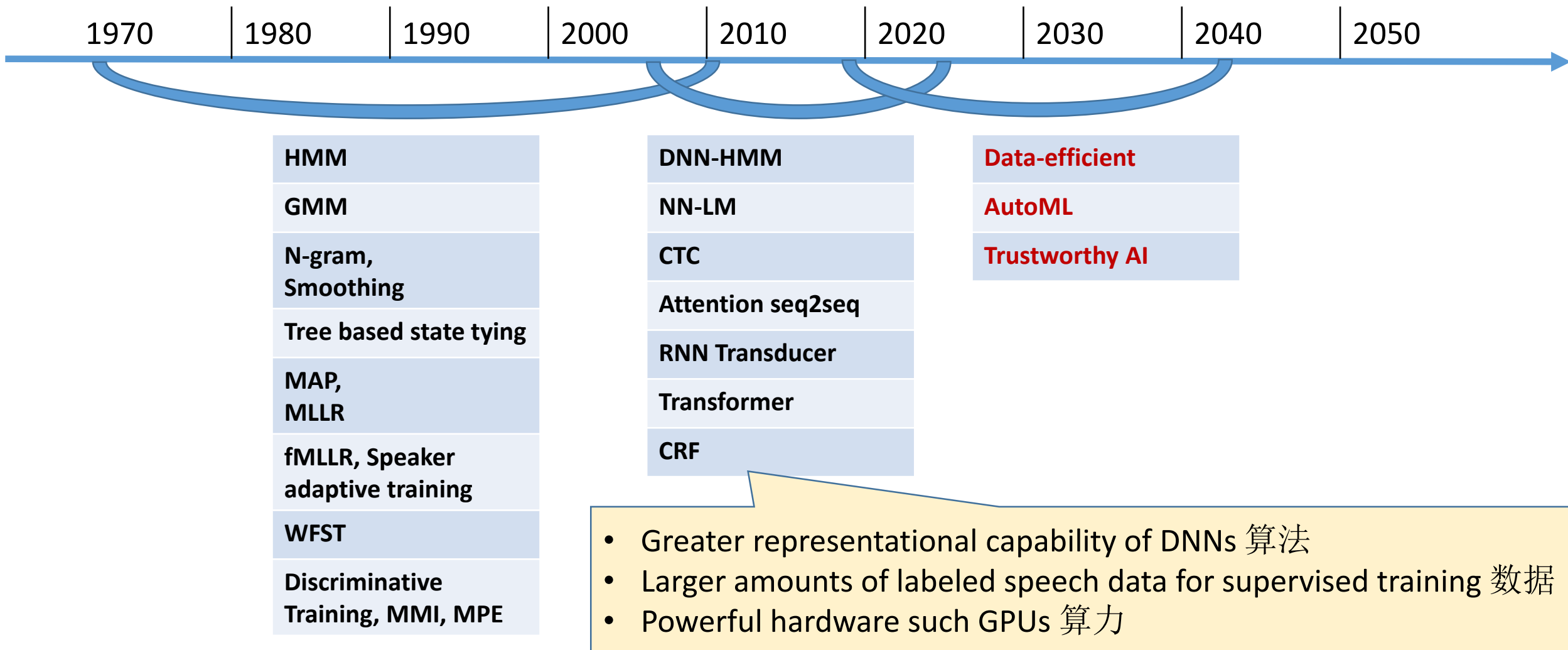
一、引言

二、语音大模型的若干思考 from first principles

- Principled **unsupervised learning**, yes/no, how?
- End-to-end is all you need (for supervised learning)?
 - **AM and LM fusion**, yes/no, how?
 - **Multi-lingual** ASR needs phonetic knowledge or not, how?

三、总结

New-generation ASR



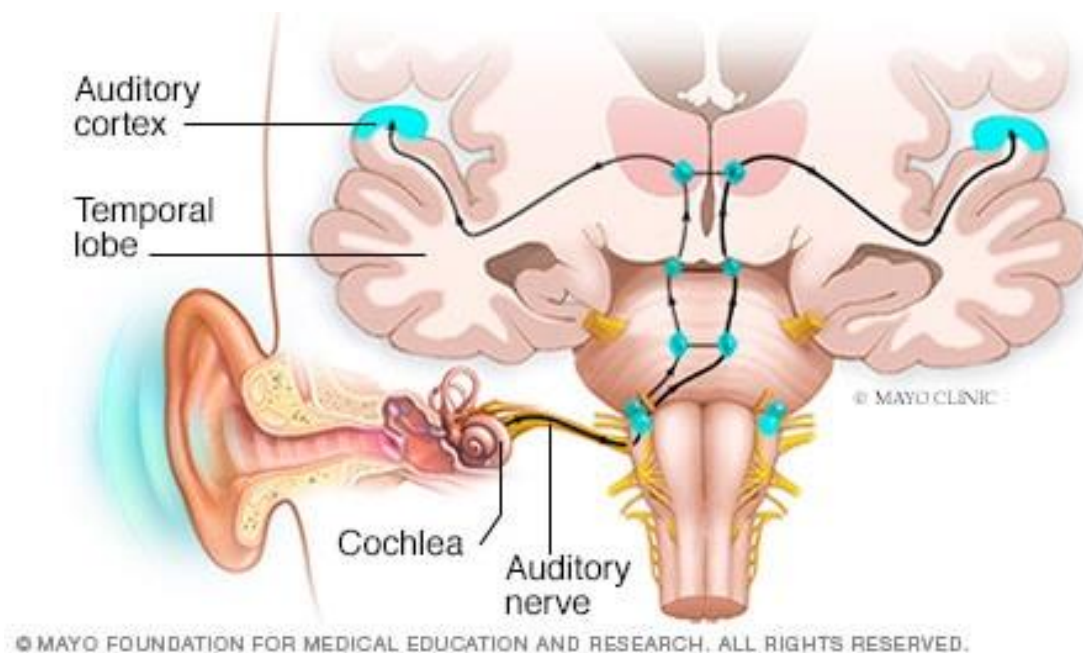
研究背景

语音信号
 $X = x_1 x_2 \dots$

声学模型 语言模型

$$P(Y|X) = \frac{\overbrace{P(X|Y)} \overbrace{P(Y)}}{\cancel{P(X)}}$$

欢迎参观清华电子系
词序列
 $Y = y_1 y_2 \dots$



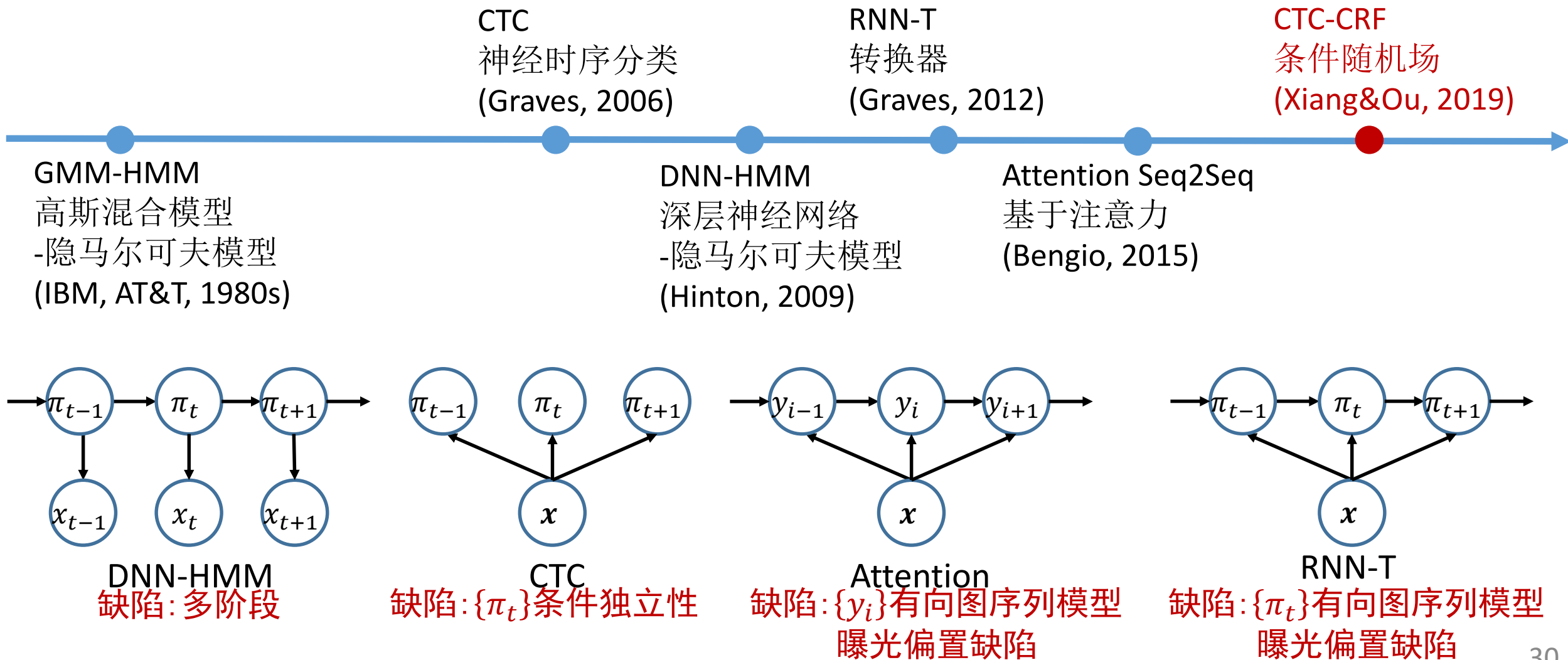
题海

- 当前技术依赖
N种声学场景 * M种语言领域
大量标注下有监督训练
- 适度模块化实现高效学习,
保留声学模型、语言模型的必要分解

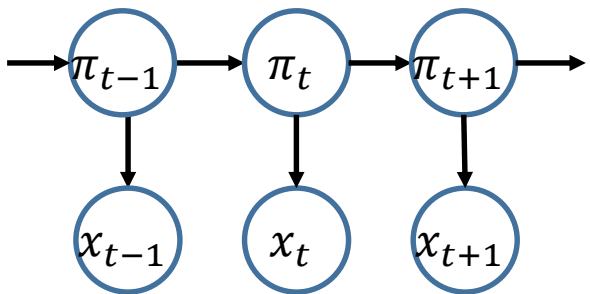
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

技术挑战

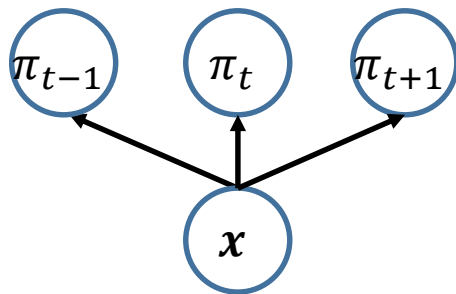
语音识别模型 $P(X|Y)$ 发展历史：具有不同的图结构，不断进步



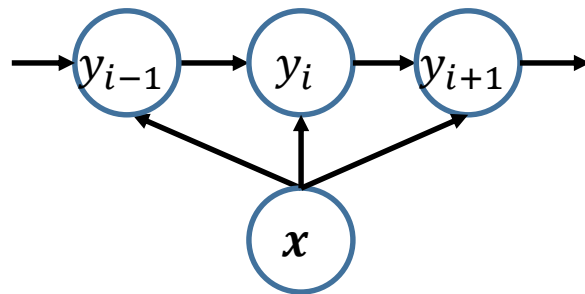
基于条件随机场的声学模型



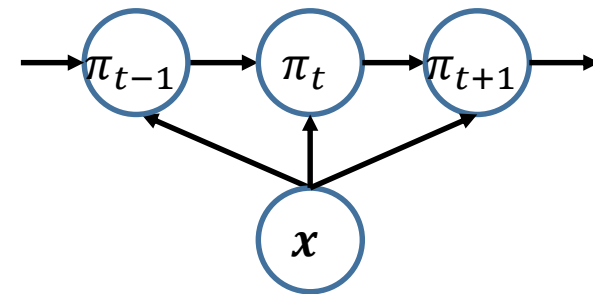
DNN-HMM
缺陷: 多阶段



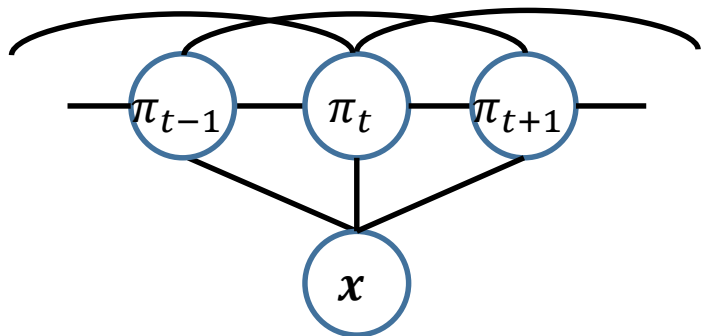
CTC
缺陷: $\{\pi_t\}$ 条件独立性



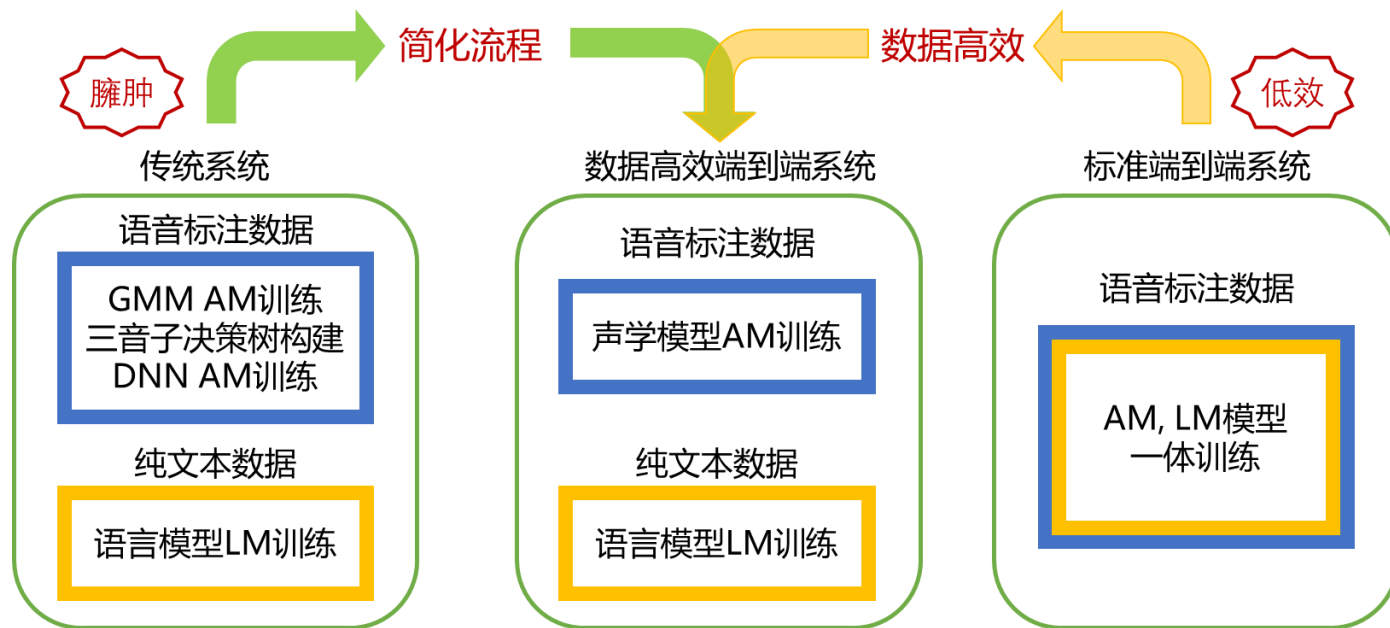
Attention
缺陷: $\{y_i\}$ 有向图序列模型



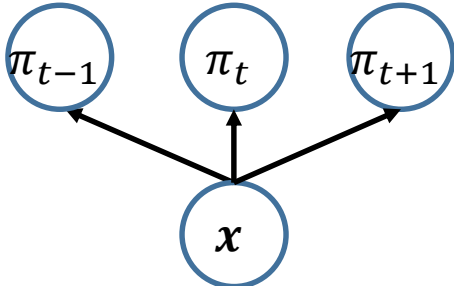
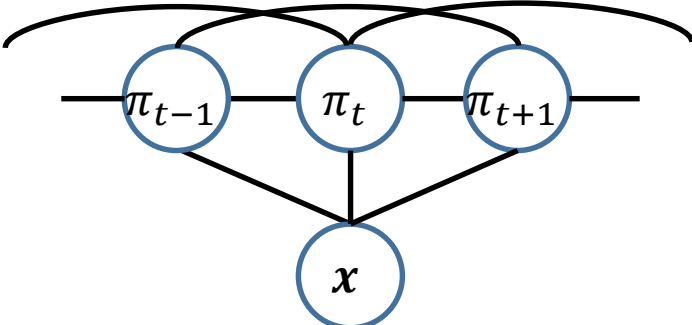
RNN-T
缺陷: $\{\pi_t\}$ 有向图序列模型



提出 CTC-CRF, 占有独特位置, 克服了历史上各类模型的缺陷, 助力数据高效,



CTC vs CTC-CRF

CTC	CTC-CRF
$p(\mathbf{y} \mathbf{x}) = \sum_{\boldsymbol{\pi}:\mathcal{B}(\boldsymbol{\pi})=\mathbf{y}} p(\boldsymbol{\pi} \mathbf{x}), \text{ using CTC topology } \mathcal{B}$	
<p>State Independence</p> $p(\boldsymbol{\pi} \mathbf{x}; \boldsymbol{\theta}) = \prod_{t=1}^T p(\pi_t \mathbf{x})$	$p(\boldsymbol{\pi} \mathbf{x}; \boldsymbol{\theta}) = \frac{e^{\phi(\boldsymbol{\pi}, \mathbf{x}; \boldsymbol{\theta})}}{\sum_{\boldsymbol{\pi}'} e^{\phi(\boldsymbol{\pi}', \mathbf{x}; \boldsymbol{\theta})}}$ <p style="text-align: right; color: red;">Node potential, by NN</p> $\phi(\boldsymbol{\pi}, \mathbf{x}; \boldsymbol{\theta}) = \sum_{t=1}^T \left(\log p(\pi_t \mathbf{x}) + \log p_{LM}(\mathcal{B}(\boldsymbol{\pi})) \right)$ <p style="text-align: right; color: red;">Edge potential, by n-gram denominator LM of labels, like in LF-MMI</p>
$\frac{\partial \log p(\mathbf{y} \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi} \mathbf{y}, \mathbf{x}; \boldsymbol{\theta})} \left[\frac{\partial \log p(\boldsymbol{\pi} \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]$	$\frac{\partial \log p(\mathbf{y} \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{\pi} \mathbf{x}, \mathbf{y}; \boldsymbol{\theta})} \left[\frac{\partial \phi(\boldsymbol{\pi}, \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] - \mathbb{E}_{p(\boldsymbol{\pi}' \mathbf{x}; \boldsymbol{\theta})} \left[\frac{\partial \phi(\boldsymbol{\pi}', \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]$
	

Section Conclusion

- The CTC-CRF framework inherits the **data-efficiency** of the hybrid approach and the **simplicity** of the end-to-end approach.
- CTC-CRF significantly **outperforms** regular CTC on a wide range of benchmarks, and is **on par with** other state-of-the-art end-to-end models.
 - English WSJ-80h, Switchboard-300h, Librispeech-1000h; Mandarin Aishell-170h; ...
- **Flexibility**
 - Streaming ASR <- INTRESPEECH 2020
 - Neural Architecture Search <- SLT 2021
 - Children Speech Recognition <- SLT 2021
 - Wordpieces, Conformer architectures
 - Multilingual and Crosslingual <- ASRU2021
 - CUSIDE: streaming ASR <- INTERSPEECH 2022
 - LODR: LM integration <- INTERSPEECH 2022
 - Integrating energy-based language model <- INTERSPEECH 2023



<https://github.com/thu-spmi/cat>

Global Normalization for Streaming Speech Recognition in a Modular Framework

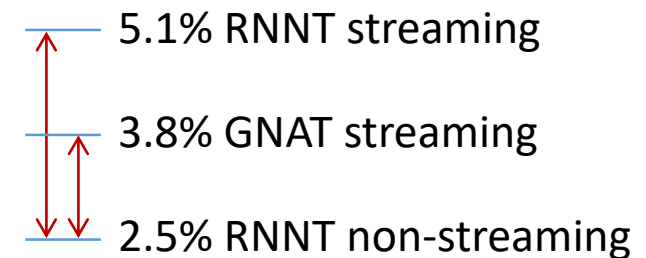
Ehsan Variani, Ke Wu, Michael Riley, David Rybach, Matt Shannon, Cyril Allauzen

Google Research

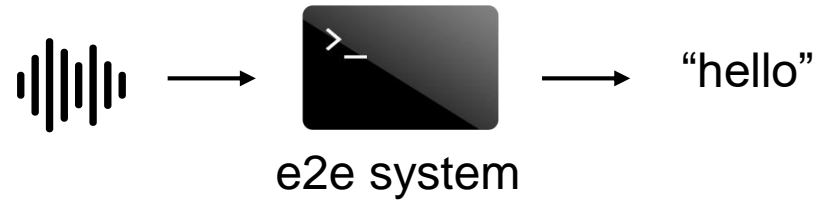
Abstract

We introduce the Globally Normalized Autoregressive Transducer (GNAT) for addressing **the label bias problem in streaming speech recognition**. Our solution admits a tractable exact computation of the denominator for the sequence-level normalization. Through theoretical and empirical results, we demonstrate that by switching to a globally normalized model, the word error rate **gap between streaming and non-streaming** speech-recognition models can be greatly reduced (by more than 50% on the Librispeech dataset). This model is developed in a modular framework which encompasses all the common neural speech recognition models. The modularity of this framework enables controlled comparison of modelling choices and creation of new models. A JAX implementation of our models has been open sourced. [\[1\]](#)

Librispeech Test-clean
WER%



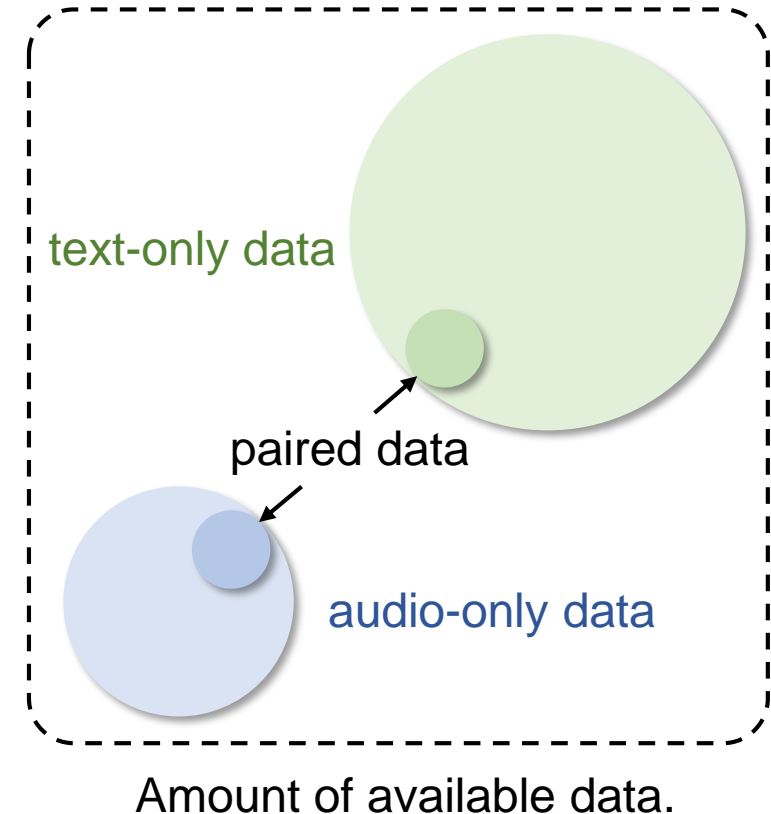
“Data efficiency” in speech recognition: towards utilizing the text-only data



- End-to-end (e2e) speech recognition is “**data hungry**”, whose performance relies on the amount of paired speech-text data.
- **Text-only** & **audio-only** data are more easily available, compared to paired ones (a.k.a. the labeled data).

How to utilize the text?

Language Model (LM) integration!



LM integration in Transducer: some intuition and heuristic experience

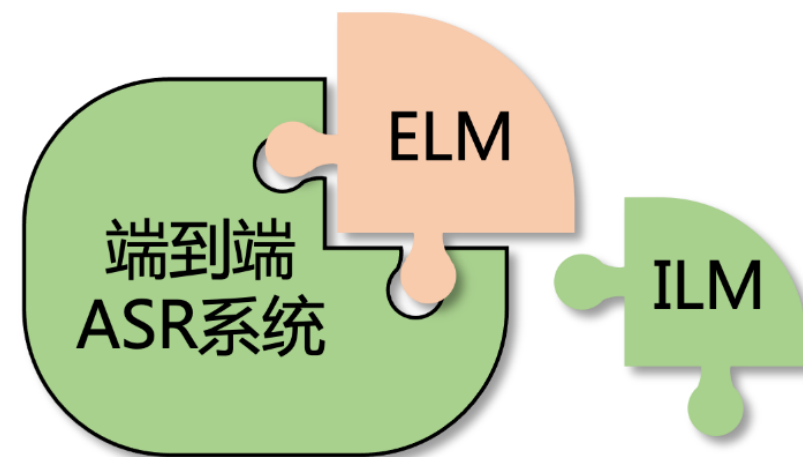
X: speech data, Y: corresponding label sequence.

Hybrid model (e.g., DNN-HMM):

$$\hat{Y} = \arg \max_Y [P_{AM}(X|Y)P_{ELM}(Y)]$$

E2E model (e.g., RNNT, AED):

$$\hat{Y} = \arg \max_Y \left[\frac{P_{RNN-T}(Y|X)}{P_{ILM}(Y)} P_{ELM}(Y) \right]$$



[1] A. Graves, "Sequence transduction with recurrent neural networks," arXiv preprint arXiv:1211.3711, 2012.

[2] Z. Meng, and et al., "Internal language model estimation for domain-adaptive end-to-end speech recognition," SLT 2021.

Related work:

shallow fusion, density ratio and ILME

1. shallow fusion (SF):

$$Y^* = \arg \max_Y (\log P_{\text{RNNT}}(Y|X) + \lambda_1 \log P_{\text{ELM}}(Y) + \beta|Y|)$$

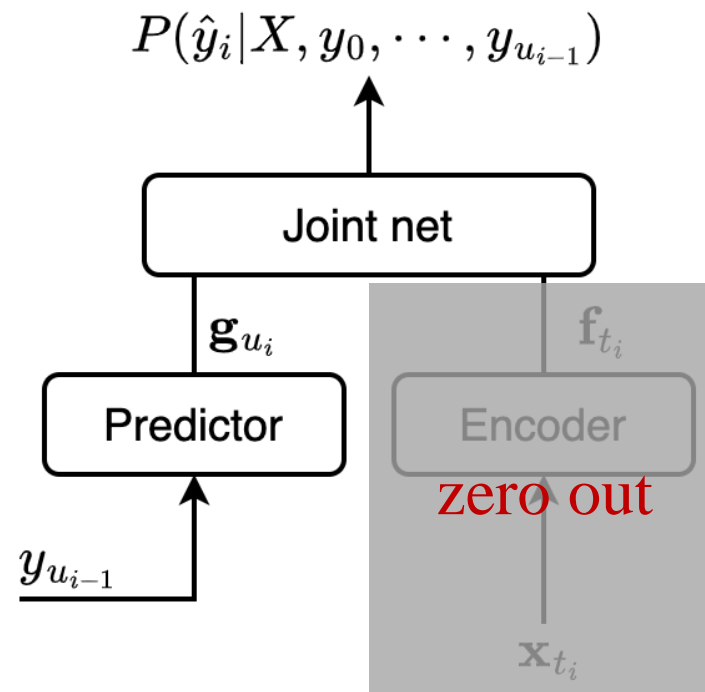
2.1 density ratio (DR):

$$Y^* = \arg \max_Y (\log P_{\text{RNNT}}(Y|X) + \lambda_0 \log P_{\text{ILM}}(Y) + \lambda_1 \log P_{\text{ELM}}(Y) + \beta|Y|)$$

ILM is approximated via a separate NN LM trained with the same linguistic information as RNN-T (transcript of the audio data).

2.2 ILME (Internal Language Model Estimation):

$$\begin{aligned} \text{linear approximation } J(\mathbf{g}_u, \mathbf{f}_t) &\approx J(\mathbf{g}_u, \mathbf{0}) + J(\mathbf{0}, \mathbf{f}_t) \\ \longrightarrow P_{\text{ILM}}(y_{u+1}|y_{0:u}) &\propto \exp(J(\mathbf{g}_u, \mathbf{0})) \end{aligned}$$



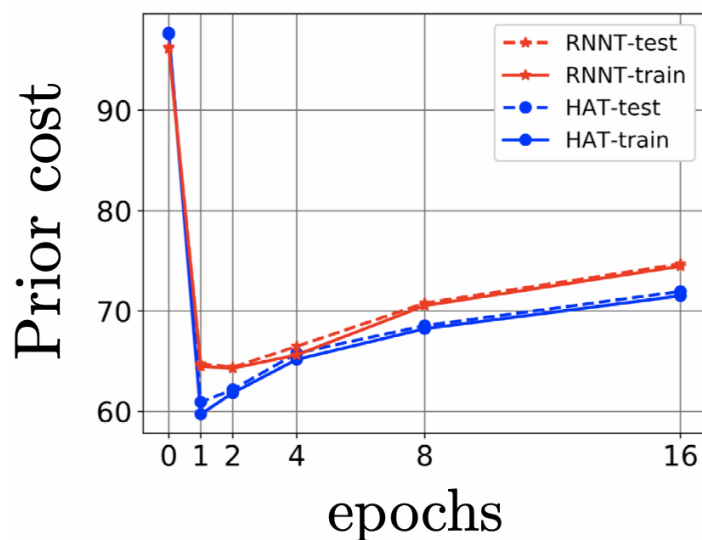
- [1] E. McDermott, and et al., "A density ratio approach to language model fusion in end-to-end automatic speech recognition," ASRU 2019.
- [2] Z. Meng, and et al., "Internal language model estimation for domain-adaptive end-to-end speech recognition," SLT 2021.

A brief summary of observation about the Predictor.

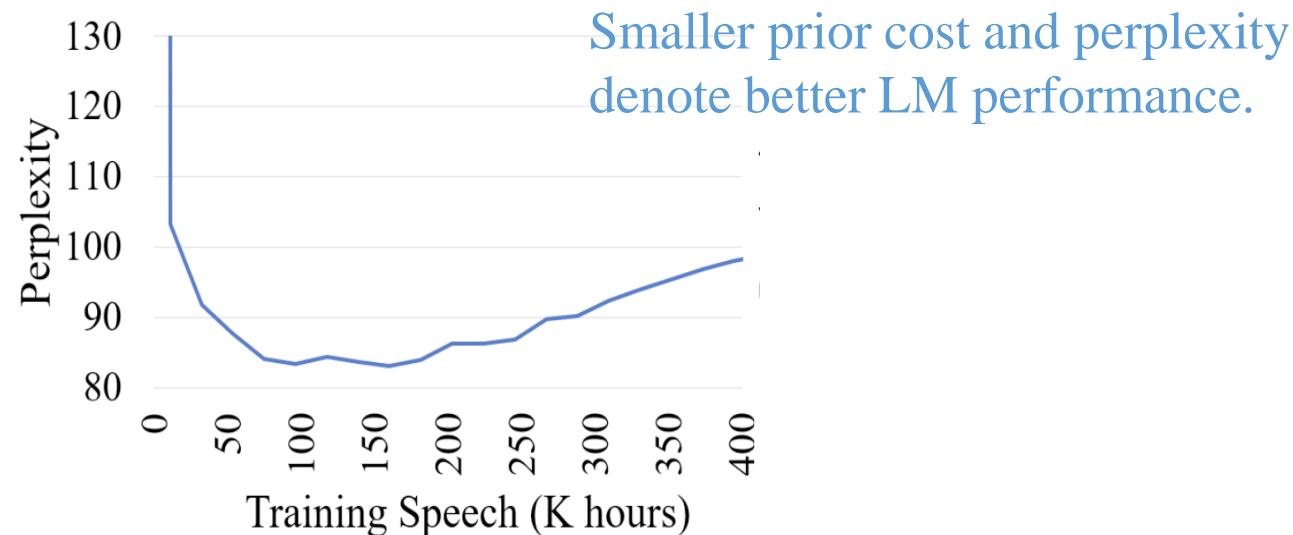
1. The Predictor is commonly very shallow neural network. (e.g. 1x LSTM);
2. The Predictor only makes use of limited context (Table 1);
3. The ILM estimated from Predictor performs poorly when evaluated as normal LM.

Table 1. Effect of limited context history [1].

Context	0	1	2	4	∞
1st-pass WER	8.5	7.4	6.6	6.6	6.6
posterior cost	34.6	5.6	5.2	4.7	4.6



(a) Prior cost of estimated ILM from HAT [1];
The “prior cost” measures the $-\log P(Y)$.

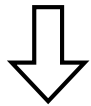
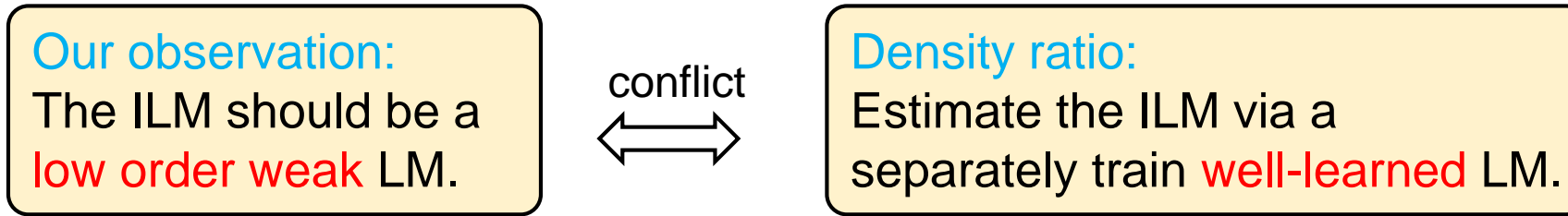


(b) Perplexity of estimated ILM from ILME [2].
A “normal” LM trained on the transcript has a perplexity of 30.1

[1] E. Variani, and et al, “Hybrid autoregressive transducer (HAT),” in ICASSP 2020.

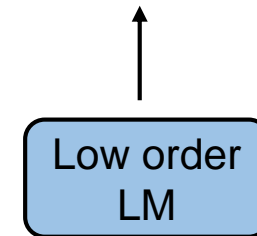
[2] Z. Meng, and et al., “Internal language model estimation for domain-adaptive end-to-end speech recognition,” in SLT 2021.

Low Order Density Ratio (LODR)



Low Order Density ratio:
Estimate the ILM via **2-gram** model.

$$Y^* = \arg \max_Y (\log P_{\text{RNNT}}(Y|X) + \lambda_0 \log P_{\text{ILM}}(Y) + \lambda_1 \log P_{\text{ELM}}(Y) + \beta|Y|)$$



In practice, we obtain the ILM as follows:

1. Prepare the training corpus: we use the transcript only;
2. Train a 2-gram LM on the corpus using KenLM with some prunes if required*.

* The size of context could be different according to the granularity of the modeling units.

[1] <https://github.com/kpu/kenlm>

Experiments: in-domain evaluation with large amount of text corpus

Table 3. Performance of LM integration methods, measured by WER % on LibriSpeech and CER % on WenetSpeech. The perplexity (PPL) of the ILM is computed on the transcript of each dataset. “Rel %” measures the relative reduction of WER (CER) compared to “No LM” setup.

Method	ILM PPL	λ_0	λ_1	β	LibriSpeech					
					dev		test		avg.	Rel %
					clean	other	clean	other		
No LM	-	-	-	-	2.18	5.33	2.40	5.42	3.81	-
SF	-	-	0.625	1.0	1.82	4.06	1.96	4.42	3.04	20.2
DR	24.72	-0.125	0.75	0.5	1.79	4.00	1.97	4.31	3.00	21.3
ILME	50.21	-0.125	0.75	1.0	1.78	3.99	1.92	4.35	2.99	21.5
LODR	100.94	-0.125	0.75	0.75	1.83	4.00	1.94	4.34	3.01	21.0

Method	ILM PPL	λ_0	λ_1	β	WenetSpeech				
					dev	test		avg.	Rel %
						net	meeting		
No LM	-	-	-	-	11.14	12.75	20.88	14.05	-
SF	-	-	0.25	3.125	9.19	11.73	18.36	12.37	12.0
DR	37.89	0.0	0.25	3.125	9.19	11.73	18.36	12.37	12.0
ILME	94.32	-0.125	0.375	3.0	9.10	11.56	18.26	12.25	12.8
LODR	79.33	-0.125	0.375	3.125	9.07	11.54	18.23	12.22	13.0

Size of extra corpus:

English: 800 million words
(9.4M words in transcript)

Chinese: 200 million chars
(17M chars in transcript)

All methods subtracting ILM perform better than the shallow fusion consistently.

Experiments: cross-domain evaluation and discussion

Table 4. Performance of LM integration methods evaluated on cross-domain scenarios.

Method	λ_0	λ_1	β	LibriSpeech \rightarrow Tedlium-2			
				dev	test	avg.	Rel %
No LM	-	-	-	11.67	11.41	11.51	-
SF	-	0.625	1.5	10.26	10.05	10.13	12.0
DR	-0.125	0.625	1.5	10.21	9.85	9.99	13.2
ILME	-0.125	0.5	1.0	10.23	9.87	10.01	13.0
LODR	-0.125	0.625	1.5	10.25	9.97	10.08	12.4

Method	λ_0	λ_1	β	WenetSpeech \rightarrow AISHELL-1			
				dev	test	avg.	Rel %
No LM	-	-	-	6.32	7.22	6.63	-
SF	-	0.5	1.375	5.11	5.56	5.26	20.7
DR	-0.125	0.5	1.375	5.10	5.65	5.28	20.4
ILME	-0.125	0.5	1.125	4.99	5.55	5.18	21.9
LODR	-0.375	0.625	0.375	4.76	5.33	4.95	25.3

Size of extra corpus:
English (Tedlium-2):
2.2M words (9.4M words in transcript)

Chinese (AISHELL-1):
1.7M chars (17M chars in transcript)

Librispeech (960 hours). Streaming encoder + stateless Transducer.

Decoding method	λ_1	λ_2	test-clean	WERR	test-other	WERR
Modified beam search	-	-	2.73	-	7.15	-
+ SF	0.3	-	2.42	11.4%	6.46	9.7%
+ ILME	0.3	-0.05	2.36	13.6%	6.23	12.9%
+ LODR (bi-gram)	0.3	-0.16	2.28	16.5%	5.94	16.9%

Librispeech + Gigaspeech (10k hours). Non-streaming encoder + pruned & stateless Transducer.

Decoding method	λ_1	λ_2	test-clean	WERR	test-other	WERR
Modified beam search	-	-	2.00	-	4.63	-
+ SF	0.3	-	1.96	2.0%	4.18	9.7%
+ ILME	0.3	-0.05	1.82	9.0%	4.10	11.4%
+ LODR (bi-gram)	0.4	-0.14	1.83	8.5%	4.03	13.0%

在K2实验中，LODR表现优秀！

*Results are reported on icefall, a repo maintained by the K2 team.

[1] <https://github.com/k2-fsa/k2>

[2] <https://github.com/k2-fsa/icfall>

内容安排



一、引言

二、语音大模型的若干思考 from first principles

- Principled **unsupervised learning**, yes/no, how?
- End-to-end is all you need (for supervised learning)?
 - **AM and LM** fusion, yes/no, how?
 - **Multi-lingual ASR** needs phonetic knowledge or not, how?

三、总结

Motivation

- There are more than 7100 languages in the world, and most of them are low-resourced languages.
- Multilingual speech recognition
 - Training data from a number of languages (**seen languages**) are merged to train a multilingual AM.
- Crosslingual speech recognition
 - The target language is **unseen** in training the multilingual AM.
 - In **few-shot** setting , the AM can be finetuned on limited target language data.
 - In **zero-shot** setting , the AM is directly used without finetuning*.

* Suppose that text corpus from the target language are available.

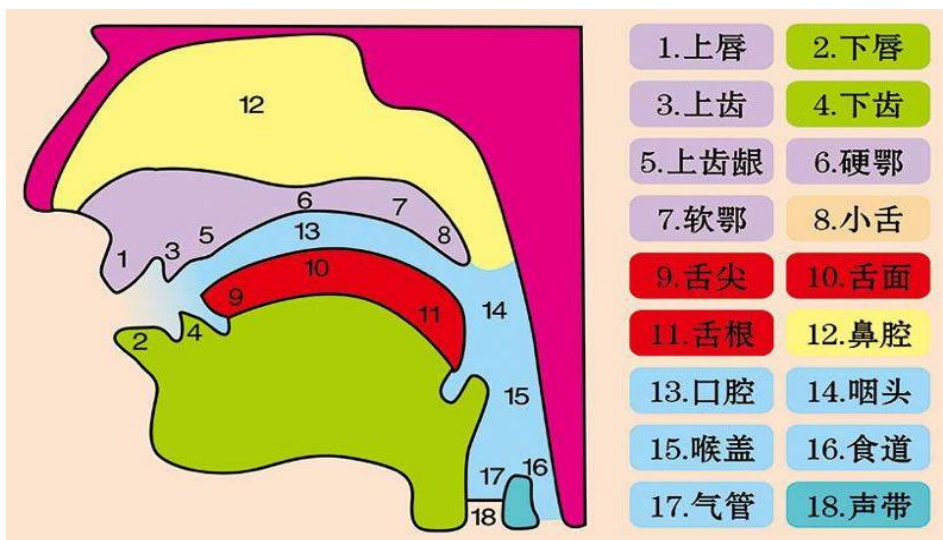
Intuitively, the key to successful multilingual and crosslingual recognition is to promote the information sharing in multilingual training and maximize the knowledge transferring from the well trained multilingual model to the model for recognizing the utterances in the new language.

Universal Phone Set

国际音标 (修订至 2005 年) since 1888

中文版 © 2007 中国语言学会语音学分会

无论哪种人类语言，都是人类的一套发音器官发出来的音



辅音 (肺部气流)

	双唇	唇齿	齿	龈	龈后	卷舌	硬腭	软腭	小舌	咽	喉
爆发音	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
鼻音	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
颤音	ʙ			r					ʀ		
拍音或闪音		ɸ		ɾ		ɽ					
擦音	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
边擦音				ɬ ɮ							
近音		ʋ		ɹ		ɻ	ɰ	ɰ			
边近音				ɭ		ɮ	ʎ	ʎ			

成对出现的音标, 右边的为浊辅音。阴影区域表示不可能产生的音。

辅音 (非肺部气流)

喷音	浊内爆音	喷音
⊙ 双唇音	ɓ 双唇音	ʼ 例如:
齿音	ɗ 齿音/龈音	pʼ 双唇音
! 龈(后)音	ɟ 硬腭音	tʼ 齿音/龈音
≡ 腭龈音	ɠ 软腭音	kʼ 软腭音
龈边音	ɢ 小舌音	sʼ 龈擦音

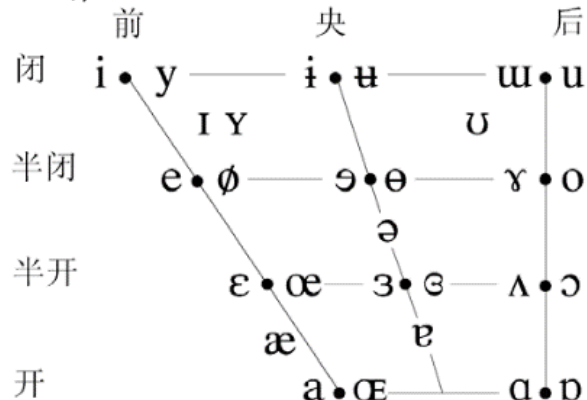
其他符号

- ʌ 唇-软腭清擦音
- ʷ 唇-软腭浊近音
- ɥ 唇-硬腭浊近音
- ʜ 会厌清擦音
- ʕ 会厌浊擦音
- ʔ 会厌爆发音
- ɕ ɟ 龈-腭擦音
- ɺ 龈边浊闪音
- ɧ 同时发 ʃ 和 x

若有必要, 塞擦音及双重调音可以用连音符连接两个符号, 如:

kp̚ ts̚

元音



成对出现的音标, 右边的为圆唇元音。

超音段

- ˈ 主重音
- ˌ 次重音
- ː 长

founəˈtʃən
eː

Joining of Acoustics and Phonology (JoinAP)

- The JoinAP method

- DNN based acoustic feature extraction (bottom-up) and phonology driven phone embedding (top-down) are joined to calculate the **logits**.

- JoinAP-Linear

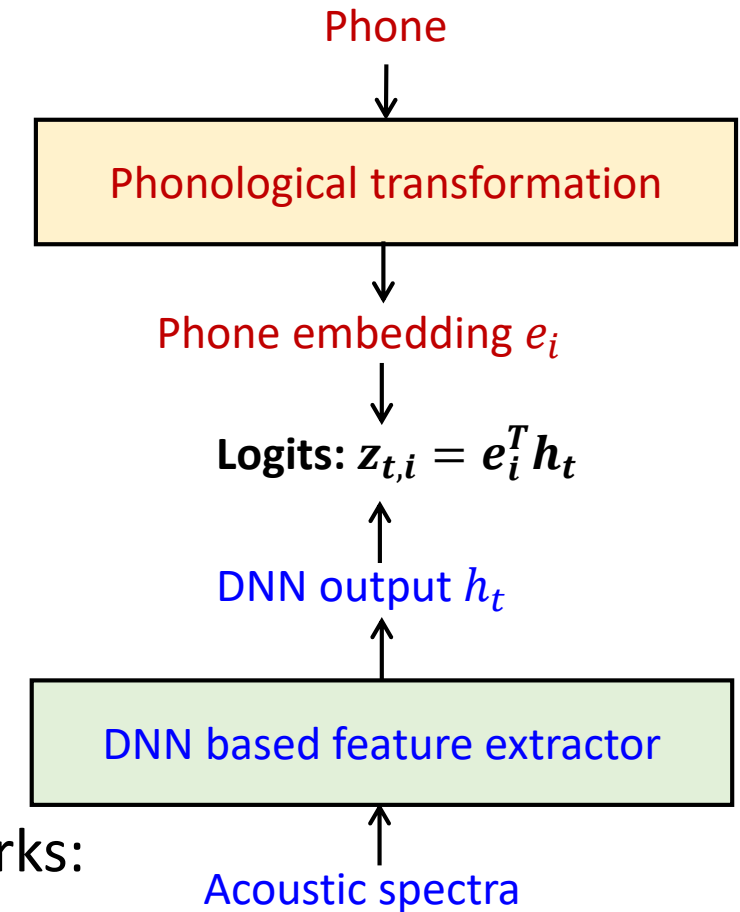
- Linear transformation of phonological-vector p_i to define the embedding vector for phone i :

$$e_i = Ap_i \in \mathbb{R}^H$$

- JoinAP-Nonlinear

- Apply nonlinear transformation, multilayered neural networks:

$$e_i = A_2 \sigma(A_1 p_i) \in \mathbb{R}^H$$



结合声学 (Acoustic) 和音韵学 (Phonology) , 促进多语言信息共享与迁移

Phonological features

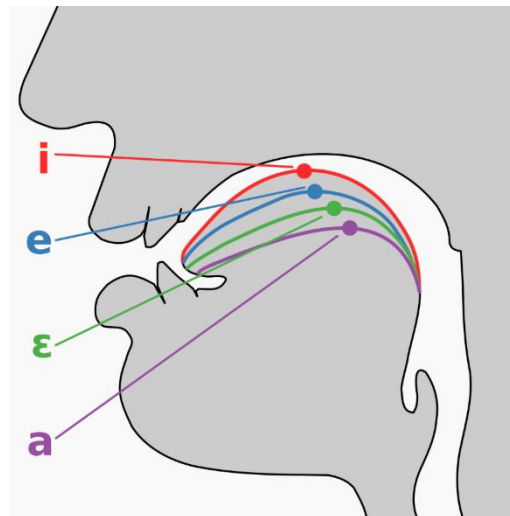
- Often **phones** are seen as being the “atoms” of speech.
- But it is now widely accepted in phonology that phones are decomposable into smaller, more fundamental units, sharable across all languages, called **phonological (distinctive) features**.
- Describe phones by phonological features

- Vowels

- vowel height
 - vowel backness

- Consonants

- Place of articulation
 - Manner of articulation



Phonological feature	d	ε	ð	ə	i	ɖ	kʲ
syllabic	-	+	-	+	+	-	-
sonorant	-	+	-	+	+	-	-
consonantal	+	-	+	-	-	+	+
continuant	-	+	+	+	+	-	-
delayed release	-	-	-	-	-	+	-
lateral	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-
strident	0	0	0	0	0	0	0
voice	+	+	+	+	+	+	-
spread glottis	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-
anterior	+	0	+	0	0	-	-
coronal	+	-	+	-	-	+	-
distributed labial	-	0	+	0	0	+	0
labial	-	-	-	-	-	-	-
high	-	-	-	-	+	+	+
low	-	-	-	-	-	-	-
back	-	-	-	+	-	-	-
round	-	-	-	-	-	-	-
velaric	-	-	-	-	-	-	-
tense	0	-	0	-	+	0	0
long	-	-	-	-	-	-	-
hitone	0	0	0	0	0	0	0
hireg	0	0	0	0	0	0	0

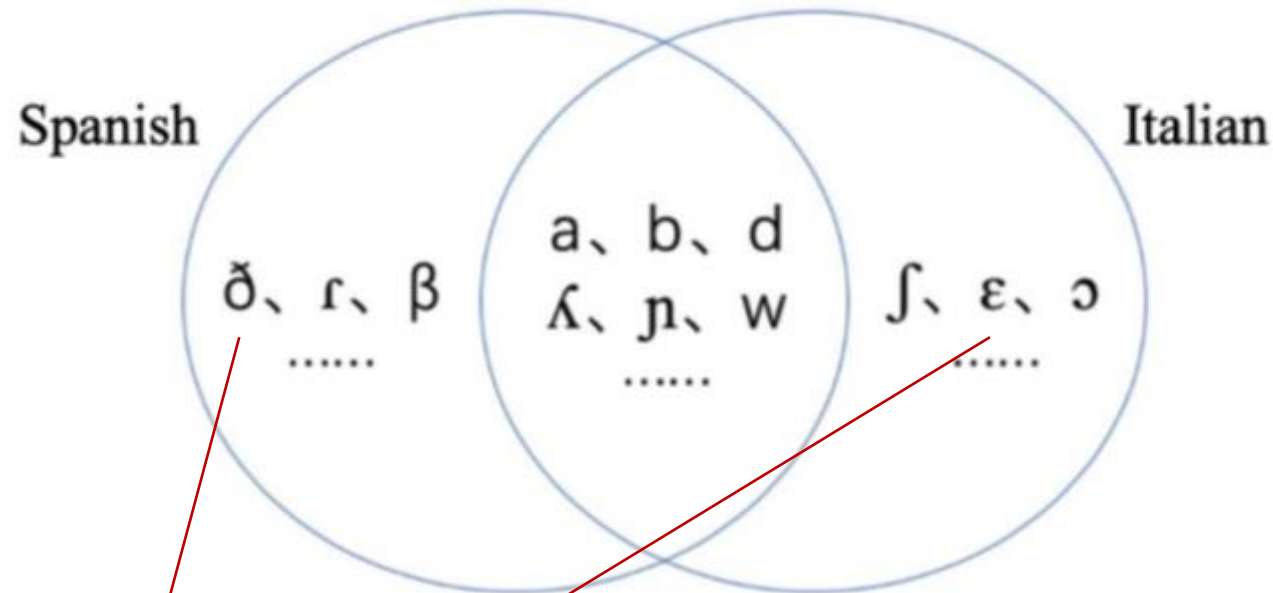
Phonological features: micro-decomposition of phones

- Like atoms could be split into nucleus and electrons, phones can be expressed by phonological features.

物质 Matter	语音 Speech
元素 Atoms	音素 Phones
元素周期表 Periodic table of elements	国际音标表 IPA table
原子核、电子 Nucleus, electrons	音韵特征 Phonological features

Phonological features: promote information sharing

- Even language-specific phones are connected by using phonological features.



ð : -, +, +, -, -, -, 0, +, -, -, +, +, +, -, -, -, -, -, -, 0, -, -, 0, 0
ε : +, +, -, +, -, -, -, 0, +, -, -, 0, -, 0, -, -, -, +, -, -, +, -, -, 0, 0

内容安排

一、引言

二、语音大模型的若干思考

三、总结



<https://docs.qq.com/form/page/DQnFzdm50Z2RNZnFX#/edit>

Summary

语音大模型的若干思考 from first principles

- Principled **semi-supervised learning**
- End-to-end is **NOT** all you need (for supervised learning)
 - **AM and LM** fusion
 - **Multi-lingual** ASR needs phonetic knowledge

First Principle 指引

感谢关注！ 欢迎交流合作！

<http://oa.ee.tsinghua.edu.cn/ouzhijian>
ozj@tsinghua.edu.cn

Talk videos can be found [here](#) at bilibili; Chinese blogs [here](#) at zhihu; Code [here](#) at github; News here  Follow @ZhijianOu

感谢宋运福、向鸿雨、安柯宇、郑华焕、刘红、朱程睿、赛尔、马特...

THU-SPMI, TasiTech is hiring!

