



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

NLP深度学习新进展

New progress in NLP deep learning

康铠、雷沛珂、杨子研、杨曼芝、马弋洋

2021/09/18



01

面向标注的解决方案

主动学习

02

面向输入的解决方案

GNN 多模态

03

面向数据的解决方案

自监督学习 提示学习 课程学习

04

面向应用的解决方案

知识蒸馏、模型压缩、安全（对抗）、可解释性

05

面向未来的解决方案

迁移学习(元学习) 多任务学习
终身学习 MRC框架





01

面向标注的解决方案

主动学习



深度学习：大量标注数据-->优化海量参数-->提取高质量特征

数据集	训练集样本数量	验证集样本数量
MNIST	50,000+	10,000
CIFAR10	50,000	10,000
ImageNet	15 million	

(模型的性能依赖于大量的标注样本)

	X-Ray	CT
时间成本	4-5 min/image	15-20 min/image
金钱成本	20-30 RMB/image	50-70 RMB/person

(标注样本很贵!)

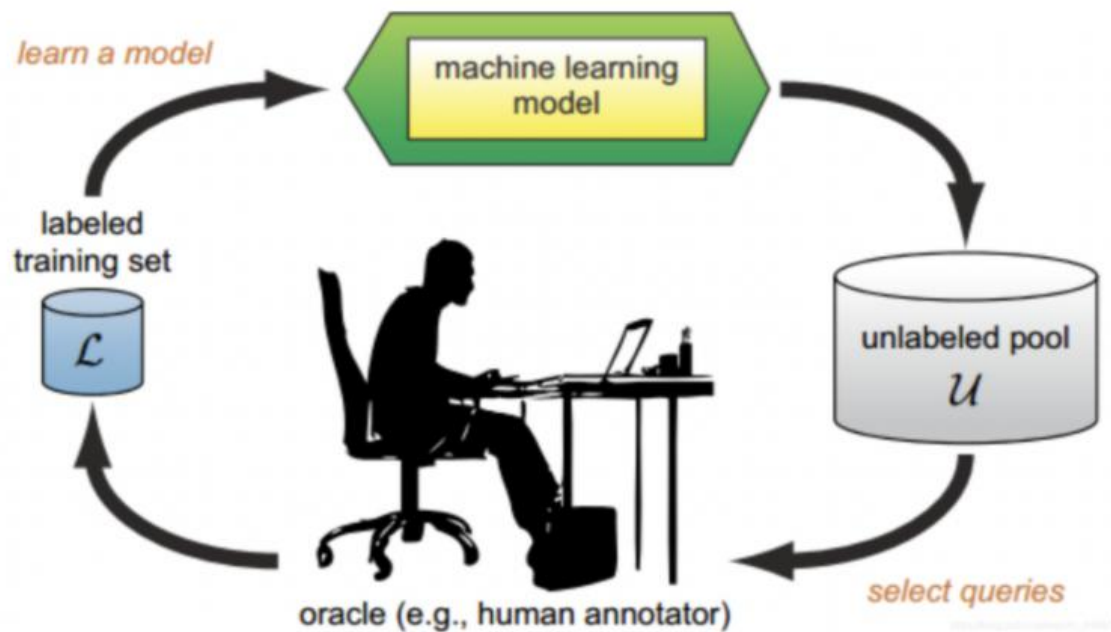
如何在模型达到性能的前提下，尽可能地减少标注成本？



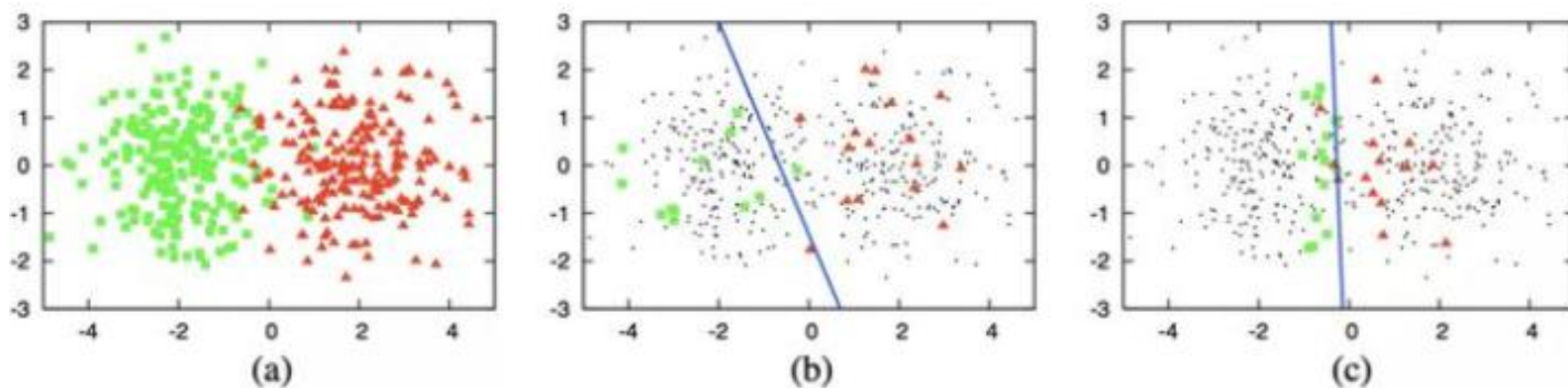
主动学习 Active learning, AL

⑩ 解决问题：降低标注成本。通过标记最少量的样本使得模型的性能收益最大化。

⑩ select queries



主动学习案例：二分类问题



(a) 400个样本；(每个类别200个样本)

(b) **随机选择**30个标注样本，验证精度70%；

(c) 通过**主动学习**策略选择30个标注样本，验证精度90%。

AL在NLP领域的前沿应用:

1. Sentiment Analysis

- ⑩ Hasan M S *et al.* (IEEE, 2020) [Truth or Lie: Pre-emptive Detection of Fake News in Different Languages Through Entropy-based Active Learning and Multi-model Neural Ensemble.](#)

2. Knowledge graph

- ⑩ (S. Seo *et al.* IEEE 2021) [Active Learning for Knowledge Graph Schema Expansion](#)

3. Speech and audio

- ⑩ Abdelwahab *et al.* (IEEE, 2019) [Active learning for speech emotion recognition using deep neural network.](#)

.....



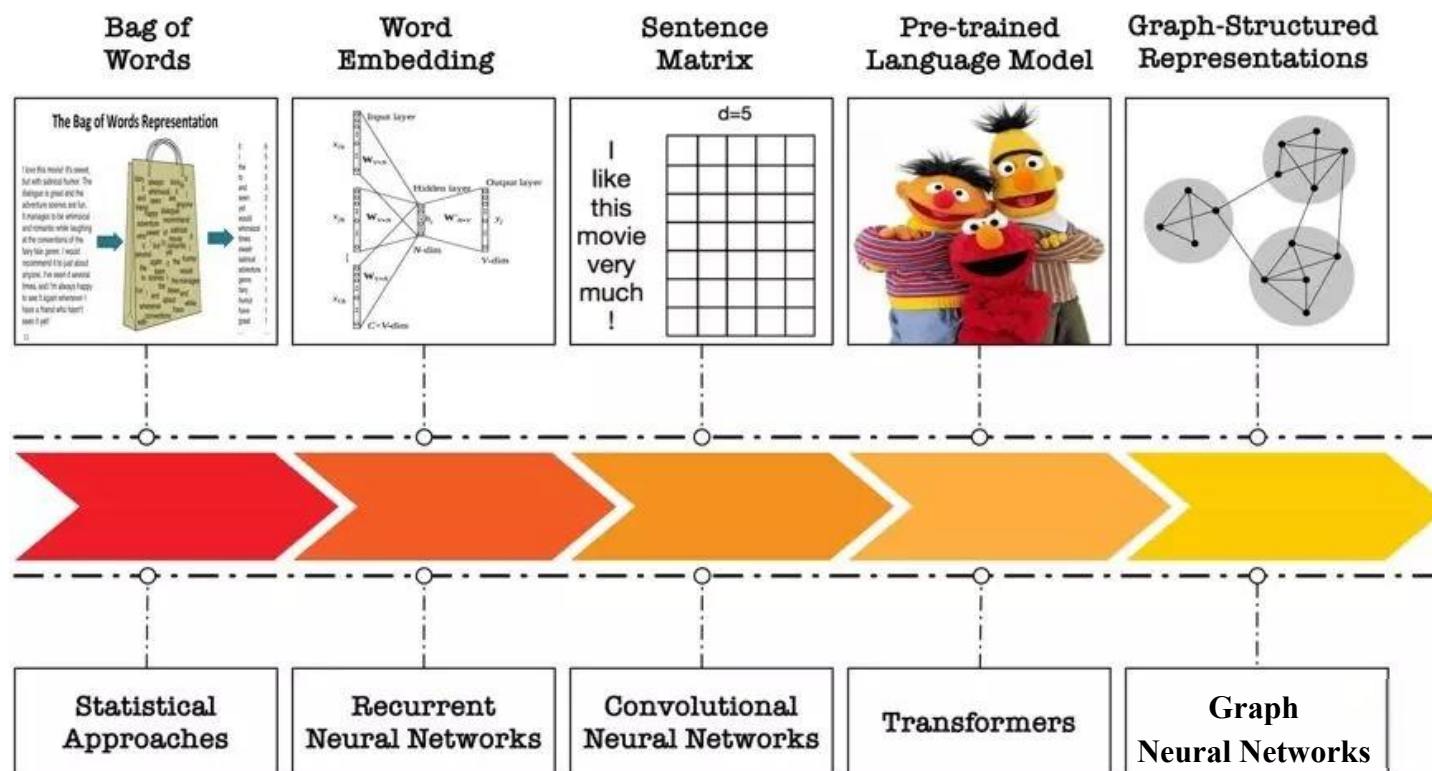
02

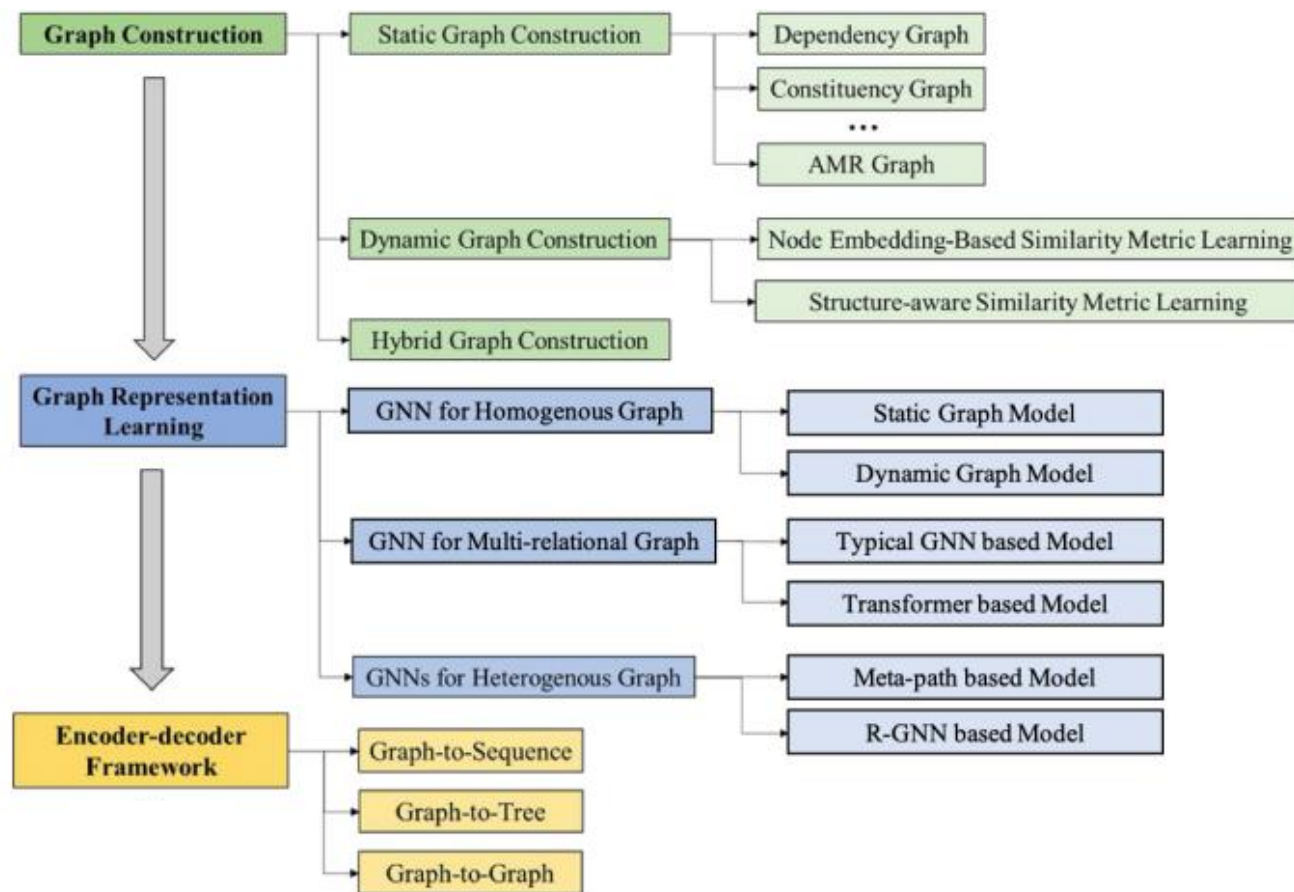
面向输入的解决方案

—
GNN、多模态



自然语言表示方式:





研究分类体系:

- ⑩ 图的构建
- ⑩ 图表示学习
- ⑩ 基于图的编码器-解码器模型



GNN在NLP领域的应用:

- ⑩ 自然语言生成
- ⑩ 机器阅读理解
- ⑩ 文本分类
- ⑩ 话题建模
- ⑩ 情感分类
- ⑩ 知识图谱
- ⑩ 实体抽取
- ⑩ 语义和句法解析
- ⑩

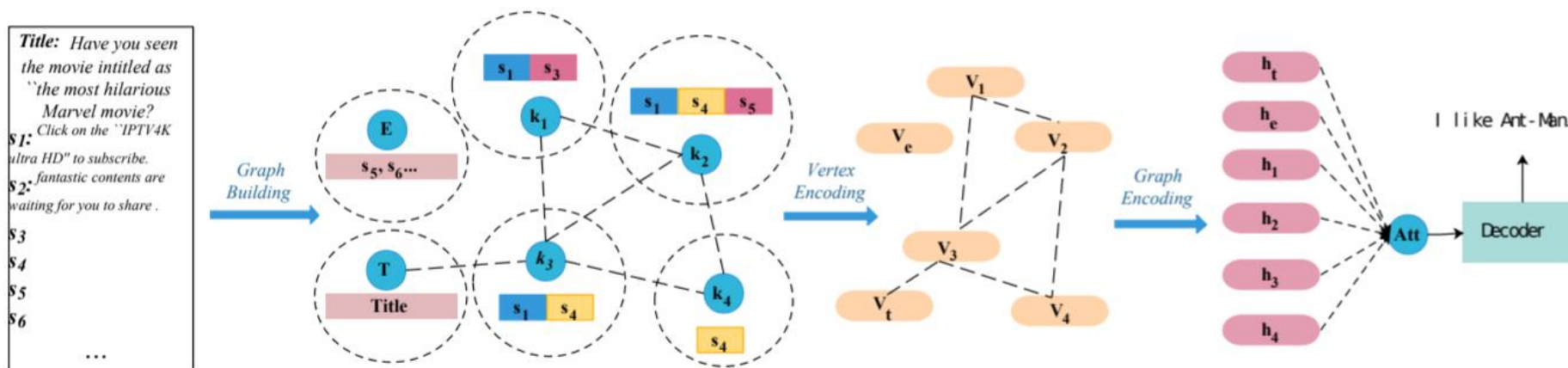
Application	Task	Evaluation	References
NLG	Neural Machine Translation	BLEU	Bastings et al. (2017); Beck et al. (2018); Cai and Lam (2020c); Guo et al. (2019c); Marcheggiani et al. (2018); Shaw et al. (2018); Song et al. (2019); Xiao et al. (2019); Xu et al. (2020c); Yin et al. (2020)
	Summarization	ROUGE	Xu et al. (2020a); Wang et al. (2019c); Li et al. (2020b); Fernandes et al. (2019); Wang et al. (2020a); Cui et al. (2020b); Jia et al. (2020); Zhao et al. (2020a); Jin et al. (2020b); Yasunaga et al. (2017); LeClair et al. (2020)
	Structural-data to Text	BLEU, METEOR	Bai et al. (2020); Jin and Gildea (2020); Xu et al. (2018a); Beck et al. (2018); Cai and Lam (2020b); Zhu et al. (2019c); Cai and Lam (2020c); Ribeiro et al. (2019b); Song et al. (2020); Wang et al. (2020f); Yao et al. (2018); Zhang et al. (2020d)
	Natural Question Generation	BLEU, METEOR, ROUGE	Chen et al. (2020g); Liu et al. (2019b); Pan et al. (2020); Wang et al. (2020d); Sachan et al. (2020); Su et al. (2020)
MRC and QA	Machine Reading Comprehension	F1, Exact Match	De Cao et al. (2018); Cao et al. (2019b); Chen et al. (2020d); Qiu et al. (2019); Schlichtkrull et al. (2018); Tang et al. (2020c); Tu et al. (2019b); Song et al. (2018b); Fang et al. (2020b); Zheng and Kordjamshidi (2020)
	Knowledge Base Question Answering	F1, Accuracy	Feng et al. (2020b); Sorokin and Gurevych (2018b); Santoro et al. (2017); Yasunaga et al. (2021)
	Open-domain Question Answering	Hits@1, F1	Han et al. (2020); Sun et al. (2019b, 2018a)
	Community Question Answering	nDCG, Precision	Hu et al. (2019b, 2020b)
Dialog Systems	Dialog State Tracking	Accuracy	Chen et al. (2018b, 2020a)
	Dialog Response Generation	BLEU, METEOR, ROUGE	Hu et al. (2019d)
	Next Utterance Selection	Recall@K	Liu et al. (2021c)
Text Classification		Accuracy	Chen et al. (2020e); Defferrard et al. (2016); Henaff et al. (2015); Huang et al. (2019); Hu et al. (2020c); Liu et al. (2020)
Text Matching		Accuracy, F1	Chen et al. (2017c); Liu et al. (2019a)
Topic Modeling		Topic Coherence Score	Long et al. (2020); Yang et al. (2020); Zhou et al. (2020a); Zhu et al. (2018)
Sentiment Classification		Accuracy, F1	Zhang and Qian (2020); Pouran Ben Veyseh et al. (2020); Chen et al. (2020c); Tang et al. (2020a); Sun et al. (2019c); Wang et al. (2020b); Zhang et al. (2019a); Ghosal et al. (2020); Huang and Carley (2019)
Knowledge Graph	Knowledge Graph Completion	Hits@N	Malaviya et al. (2020); Nathani et al. (2019a); Teru et al. (2020); Bansal et al. (2019); Schlichtkrull et al. (2018); Shang et al. (2019); Wang et al. (2019a,g); Zhang et al. (2020g)
	Knowledge Graph Alignment		Cao et al. (2019c); Li et al. (2019); Sun et al. (2020a); Wang et al. (2018, 2020b); Ye et al. (2019); Xu et al. (2019a); Wu et al. (2019a)
Information Extraction	Named Entity Recognition	Precision, Recall, F1	Luo and Zhao (2020); Ding et al. (2019b); Gui et al. (2019); Jin et al. (2019); Sui et al. (2019)
	Relation Extraction		Qu et al. (2020); Zeng et al. (2020); Sahu et al. (2019); Guo et al. (2019b); Zhu et al. (2019a)
	Joint Learning Models		Fu et al. (2019); Luan et al. (2019); Sun et al. (2019a)
Parsing	Syntax-related	Accuracy	Do and Rehbein (2020); Ji et al. (2019); Yang and Deng (2020)
	Semantics-related		Bai et al. (2020); Zhou et al. (2020b); Shao et al. (2020); Bogin et al. (2019a,b)
Reasoning	Math Word Problem Solving	Accuracy	Li et al. (2020a); Lee et al. (2020); Wu et al. (2020b)
	Natural Language Inference		Zhang et al. (2020b); Ferreira and Freitas (2020)
	Commonsense Reasoning		Kapanipathi et al. (2020); Wang et al. (2019f); Zhou et al. (2018a); Lin et al. (2019b,a)
Semantic Role Labelling		Precision, Recall, F1	Marcheggiani and Titov (2020); Xia et al. (2020); Zhang et al. (2020a); Li et al. (2018c); Marcheggiani and Titov (2017); Fei et al. (2020)



应用示例:

ACL 2019---Coherent comment generation for chinese articles with a graph-to-sequence model

将输入的新闻建模为主题交互图，表达文章内部结构与主题之间的联系。



多模态 Multimodal: 文本、图像、视频、音频等

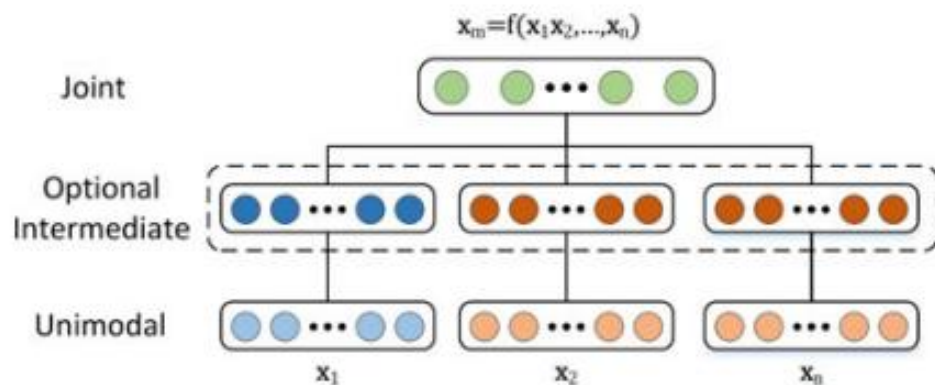
五个方向:

- ⑩ Representation
- ⑩ Translation
- ⑩ Alignment
- ⑩ Fusion
- ⑩ Co-learning

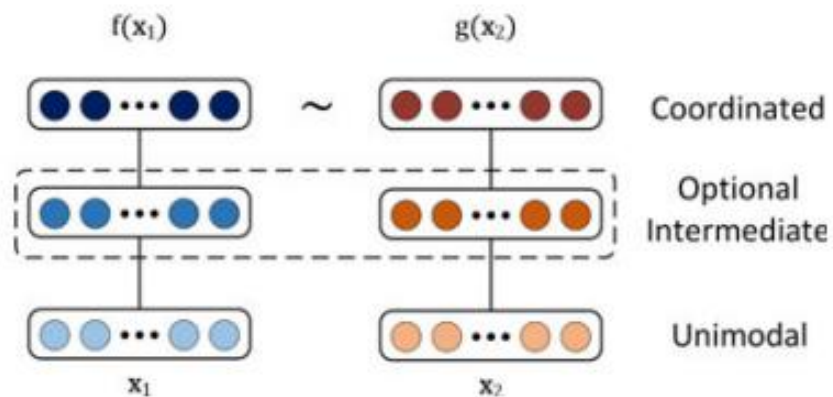


Representation (表示)

将多个模态信息抽象为实值向量。



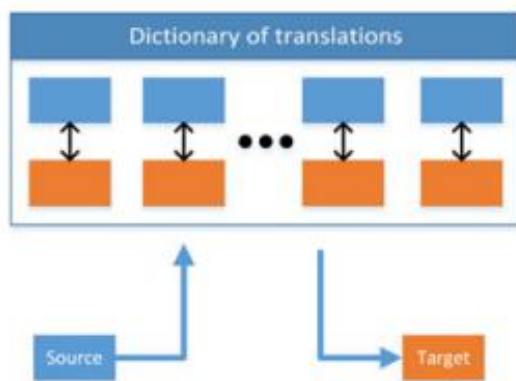
(a) Joint representation



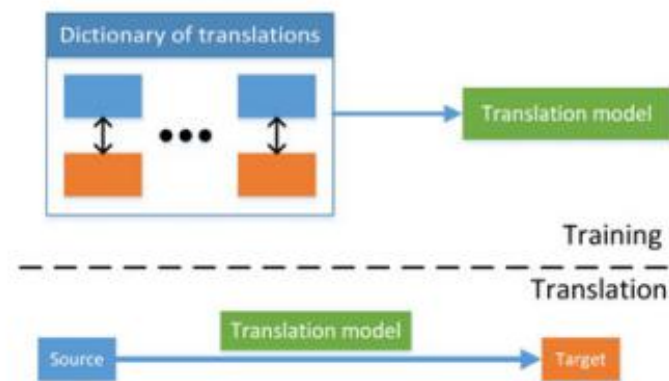
(b) Coordinated representations

Translation (映射)

给定一种模态下的表示，转换成其他模态下的表示。



(a) Example-based

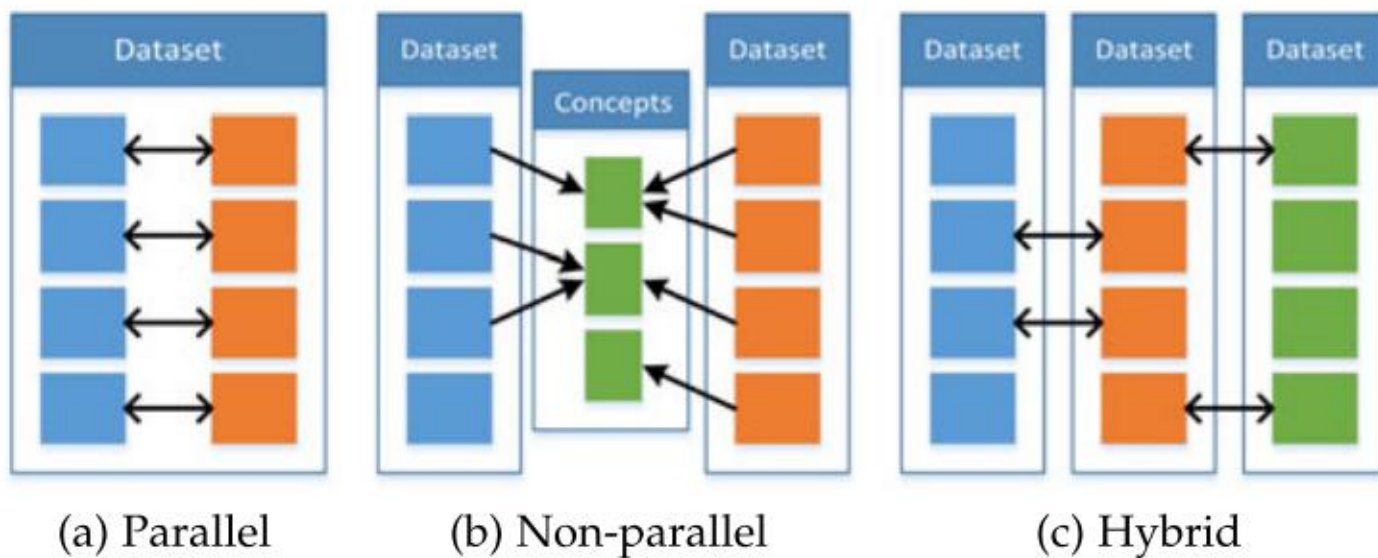


(b) Generative

Alignment (对齐) : 多种模态子成分之间的关系和对应。

Fusion (融合) : 将多个模态的信息整合到一起预测结果。

Co-learning (联合) : 通过数据源丰富的模态来辅助数据源稀少的模态进行学习。



多模态的应用:

⑩ 医疗问答

MH Vu et al. (IEEE 2020) [A question-centric model for visual question answering in medical imaging](#)

⑩ 假新闻识别

P Qi et al. (IEEE 2019) [Exploiting multi-domain visual information for fake news detection.](#)

⑩ 情感分类

H Deng (IEEE 2021) [Dense Fusion Network with Multimodal Residual for Sentiment Classification](#)

⑩ 摘要生成

M Li (ACL 2019) [Keep meeting summaries on topic: Abstractive multi-modal meeting](#)



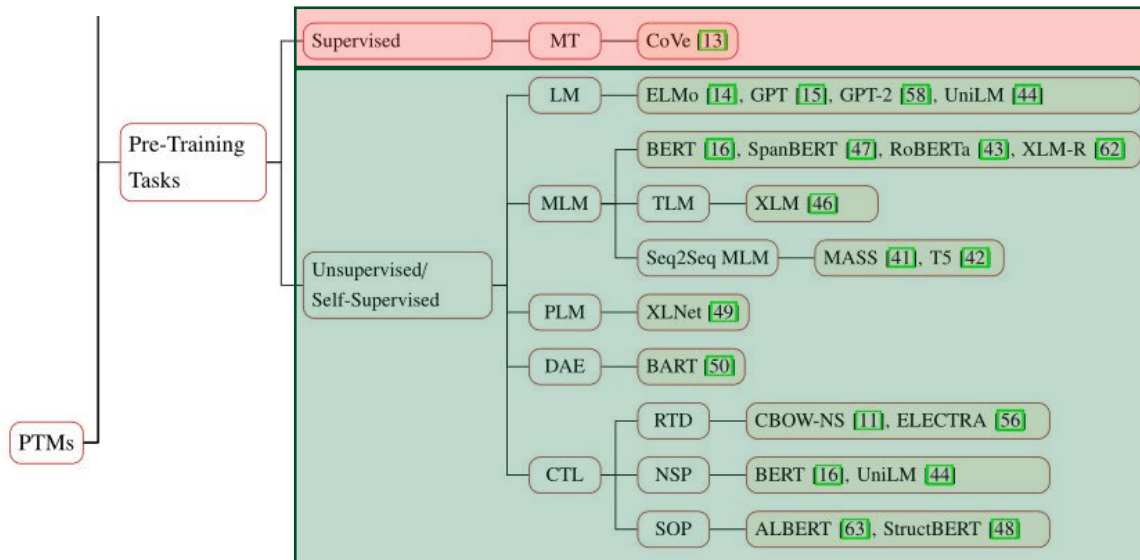
03

面向数据的解决方案

—
自监督学习 提示学习 课程学习



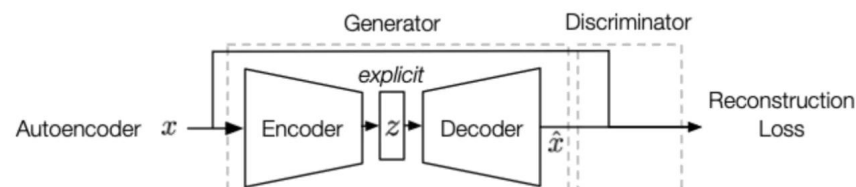
自监督学习：摆脱人类监督、回归数据自我监督



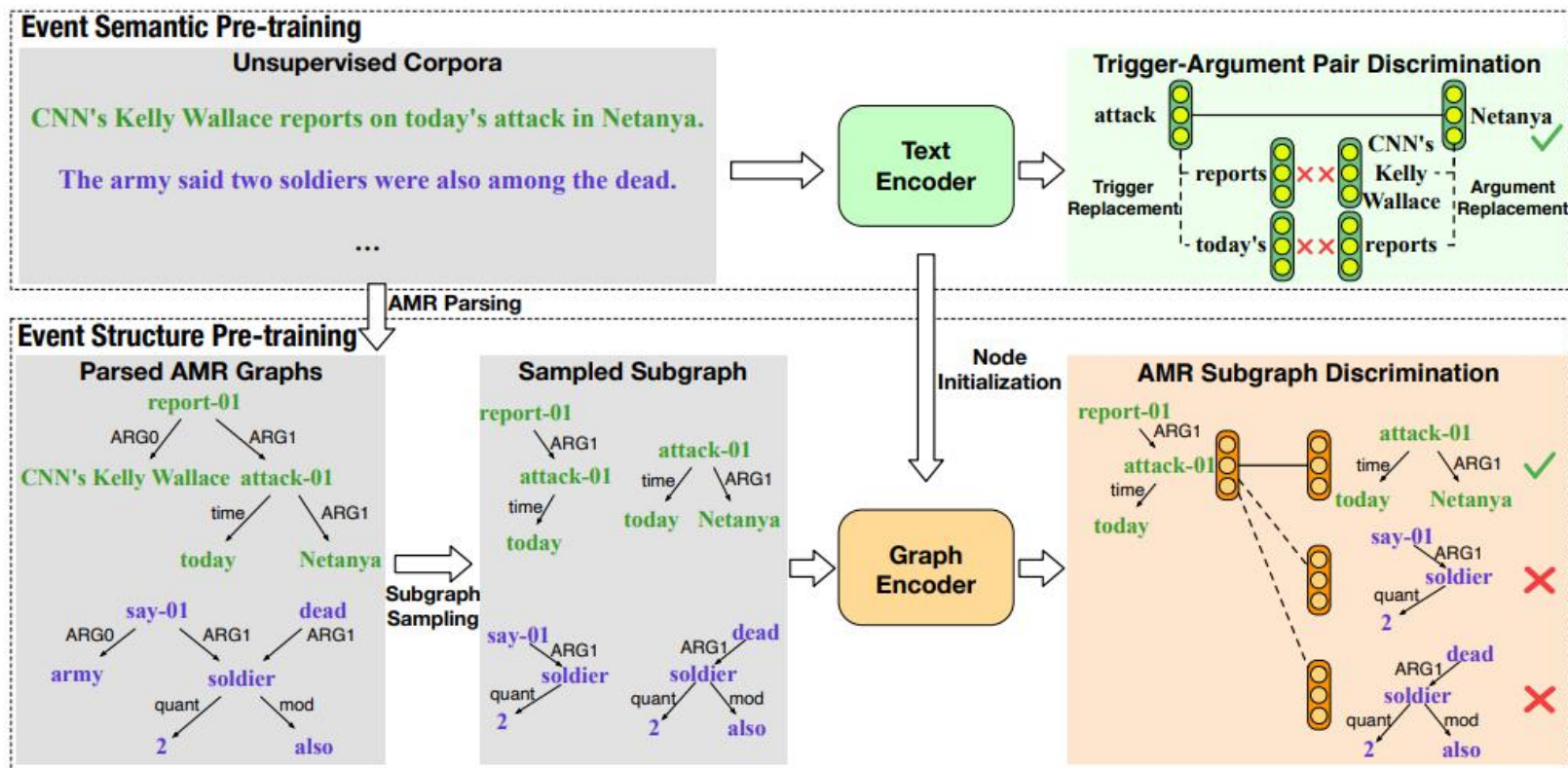
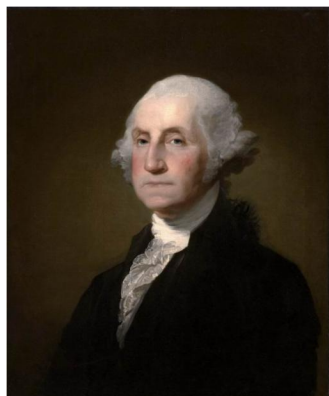
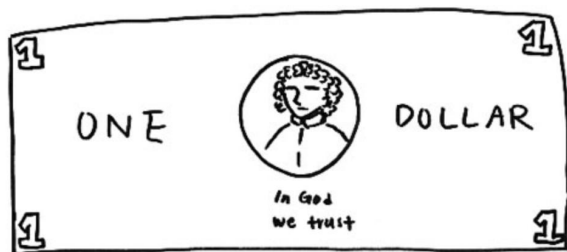
生成式：自己预测自己

Operation	Element	Original Text	Corrupted Text
Mask	one token	Jane will move to New York .	Jane will [Z] to New York .
	two tokens	Jane will move to New York .	Jane will [Z] [Z] New York .
	one entity	Jane will move to New York .	Jane will move to [Z] .
Replace	one token	Jane will move to New York .	Jane will move [X] New York .
	two tokens	Jane will move to New York .	Jane will move [X] [Y] York .
	one entity	Jane will move to New York .	Jane will move to [X] .
Delete	one token	Jane will move to New York .	Jane move to New York .
	two token	Jane will move to New York .	Jane to New York .
Permute	token	Jane will move to New York .	New York . Jane will move to
Rotate	none	Jane will move to New York .	to New York . Jane will move
Concatenation	two languages	Jane will move to New York .	Jane will move to New York . [/s] 简将搬到纽约。

Table 4: Detailed examples for different noising operations.



对比学习：相同的近，不同的远。

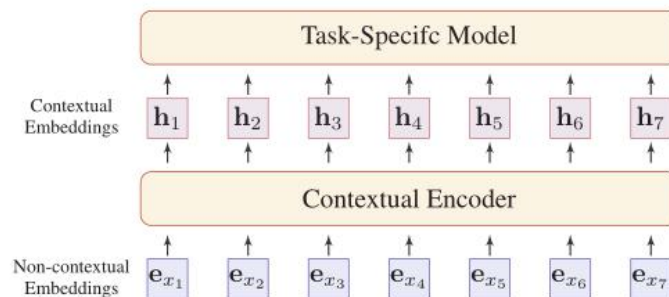
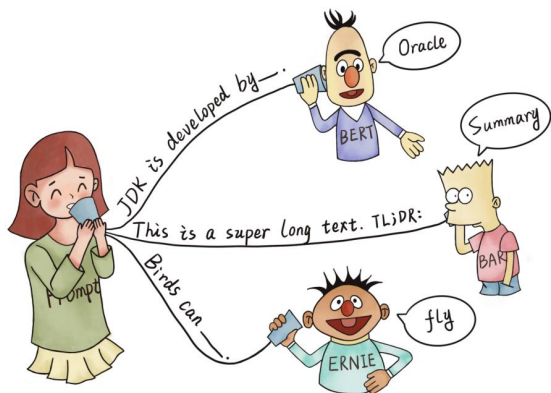


为事件的多样化语义和复杂结构找到合理的自我监督信号

对比学习：相同的近，不同的远。

Task	Loss Function	Description
LM	$\mathcal{L}_{\text{LM}} = - \sum_{t=1}^T \log p(x_t \mathbf{x}_{<t})$	$\mathbf{x}_{<t} = x_1, x_2, \dots, x_{t-1}$.
MLM	$\mathcal{L}_{\text{MLM}} = - \sum_{\hat{x} \in m(\mathbf{x})} \log p(\hat{x} \mathbf{x}_{\setminus m(\mathbf{x})})$	$m(\mathbf{x})$ and $\mathbf{x}_{\setminus m(\mathbf{x})}$ denote the masked words from $\mathbf{x}$ and the rest words respectively.
Seq2Seq MLM	$\mathcal{L}_{\text{S2SMLM}} = - \sum_{i=1}^J \log p(x_i \mathbf{x}_{\setminus \mathbf{x}_{i:j}}, \mathbf{x}_{i:t-1})$	$\mathbf{x}_{i:j}$ denotes an masked n-gram span from i to j in $\mathbf{x}$.
PLM	$\mathcal{L}_{\text{PLM}} = - \sum_{t=1}^T \log p(z_t \mathbf{z}_{<t})$	$\mathbf{z} = \text{perm}(\mathbf{x})$ is a permutation of $\mathbf{x}$ with random order.
DAE	$\mathcal{L}_{\text{DAE}} = - \sum_{t=1}^T \log p(x_t \hat{\mathbf{x}}, \mathbf{x}_{<t})$	$\hat{\mathbf{x}}$ is randomly perturbed text from $\mathbf{x}$.
DIM	$\mathcal{L}_{\text{DIM}} = s(\hat{\mathbf{x}}_{i:j}, \mathbf{x}_{i:j}) - \log \sum_{\tilde{\mathbf{x}}_{i:j} \in \mathcal{N}} s(\hat{\mathbf{x}}_{i:j}, \tilde{\mathbf{x}}_{i:j})$	$\mathbf{x}_{i:j}$ denotes an n-gram span from i to j in $\mathbf{x}$, $\hat{\mathbf{x}}_{i:j}$ denotes a sentence masked at position i to j , and $\tilde{\mathbf{x}}_{i:j}$ denotes a randomly-sampled negative n-gram from corpus.
NSP/SOP	$\mathcal{L}_{\text{NSP/SOP}} = - \log p(t \mathbf{x}, \mathbf{y})$	$t = 1$ if $\mathbf{x}$ and $\mathbf{y}$ are continuous segments from corpus.
RTD	$\mathcal{L}_{\text{RTD}} = - \sum_{t=1}^T \log p(y_t \hat{\mathbf{x}})$	$y_t = \mathbf{1}(\hat{x}_t = x_t)$, $\hat{\mathbf{x}}$ is corrupted from $\mathbf{x}$.

提示学习：动机+定义



Name	Notation	Example	Description
Input	x	I love this movie.	One or multiple texts
Output	y	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(x)$	[X] Overall, it was a [Z] movie.	A function that converts the input into a specific form by inserting the input x and adding a slot [Z] where answer z may be filled later.
Prompt	x'	I love this movie. Overall, it was a [Z] movie.	A text where [X] is instantiated by input x but answer slot [Z] is not.
Filled Prompt	$f_{\text{fill}}(x', z)$	I love this movie. Overall, it was a bad movie.	A prompt where slot [Z] is filled with any answer.
Answered Prompt	$f_{\text{fill}}(x', z^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot [Z] is filled with a true answer.
Answer	z	“good”, “fantastic”, “boring”	A token, phrase, or sentence that fills [Z]



Generate Options

Learn more in [the docs](#).

Length to generate ?

100

☐ Start at beginning ?

[Advanced Settings »](#)

Type some text here and a neural network will generate more.

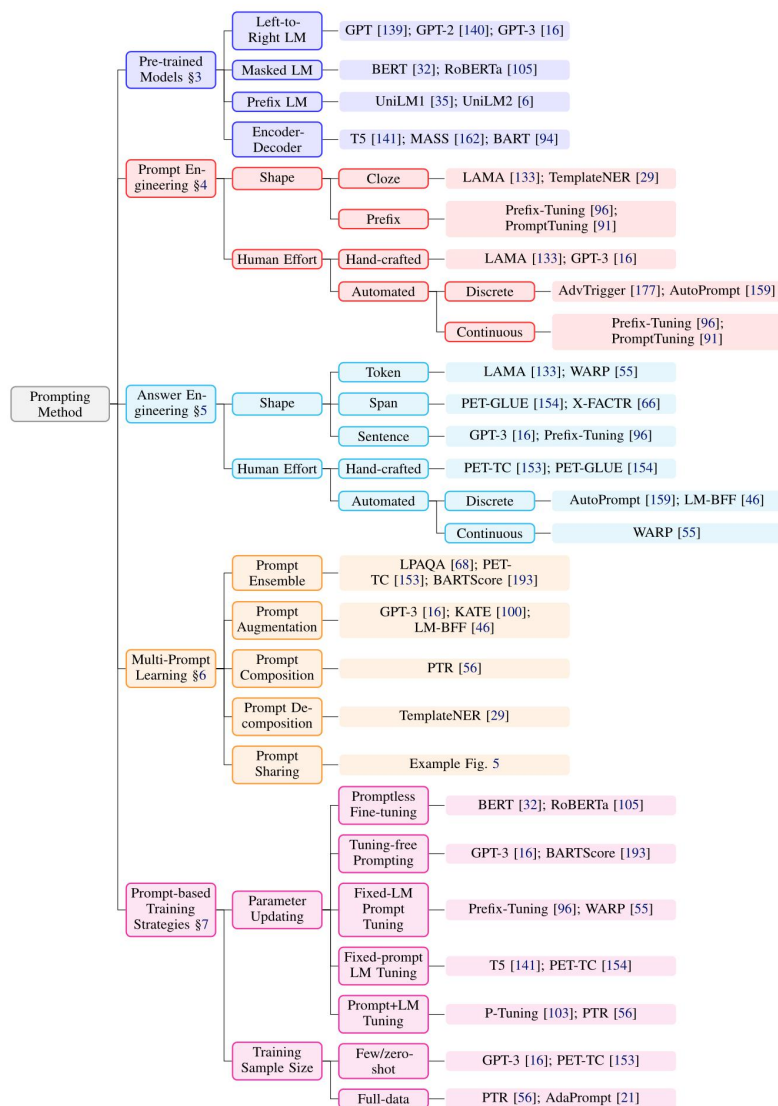
[Try an example](#)

Press tab at any point to generate more text, and esc to stop or revert.

Generate Text



提示学习：方法+应用



知识挖掘

分类任务：文本分类

信息抽取：命名实体识别、关系抽取

常识推理、数学推理

问答

文本生成

多模态

.....



Generate Options

Learn more in [the docs](#).

Length to generate ?

100

☐ Start at beginning ?

[Advanced Settings »](#)

Tom: I graduated from the experimental high school attached to Beijing Normal University. And I am studying in Beijing Institute of Technology

Tony: Which is your university?

Tom: In

Generate Text

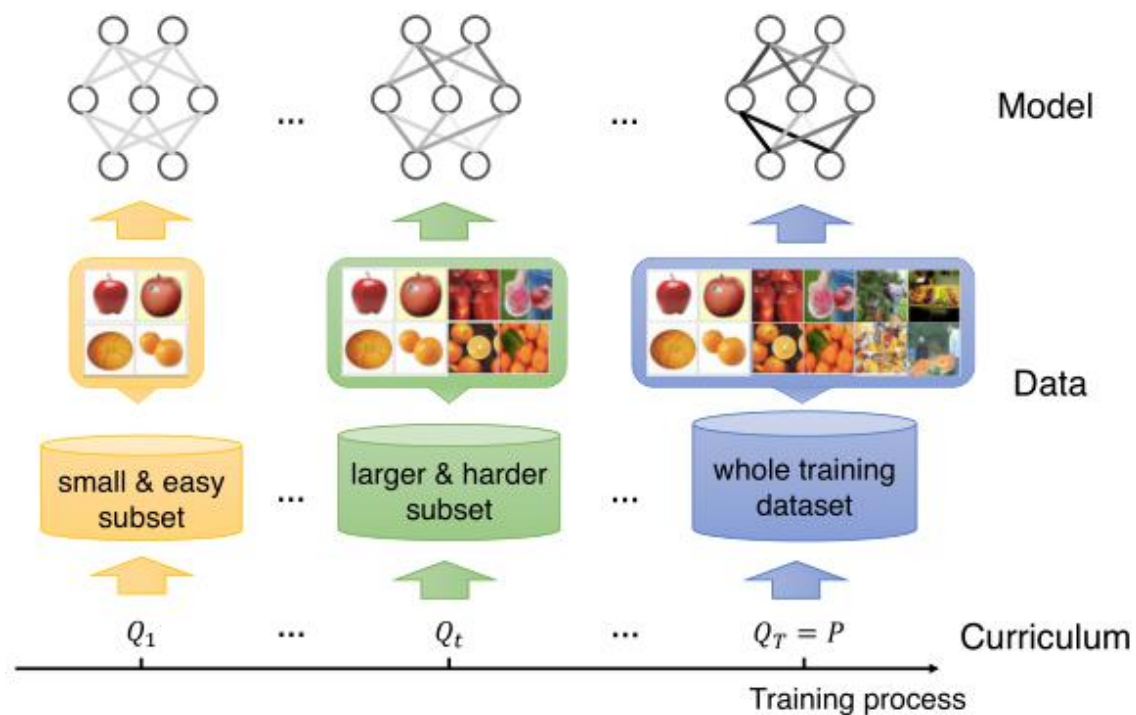


• 定义

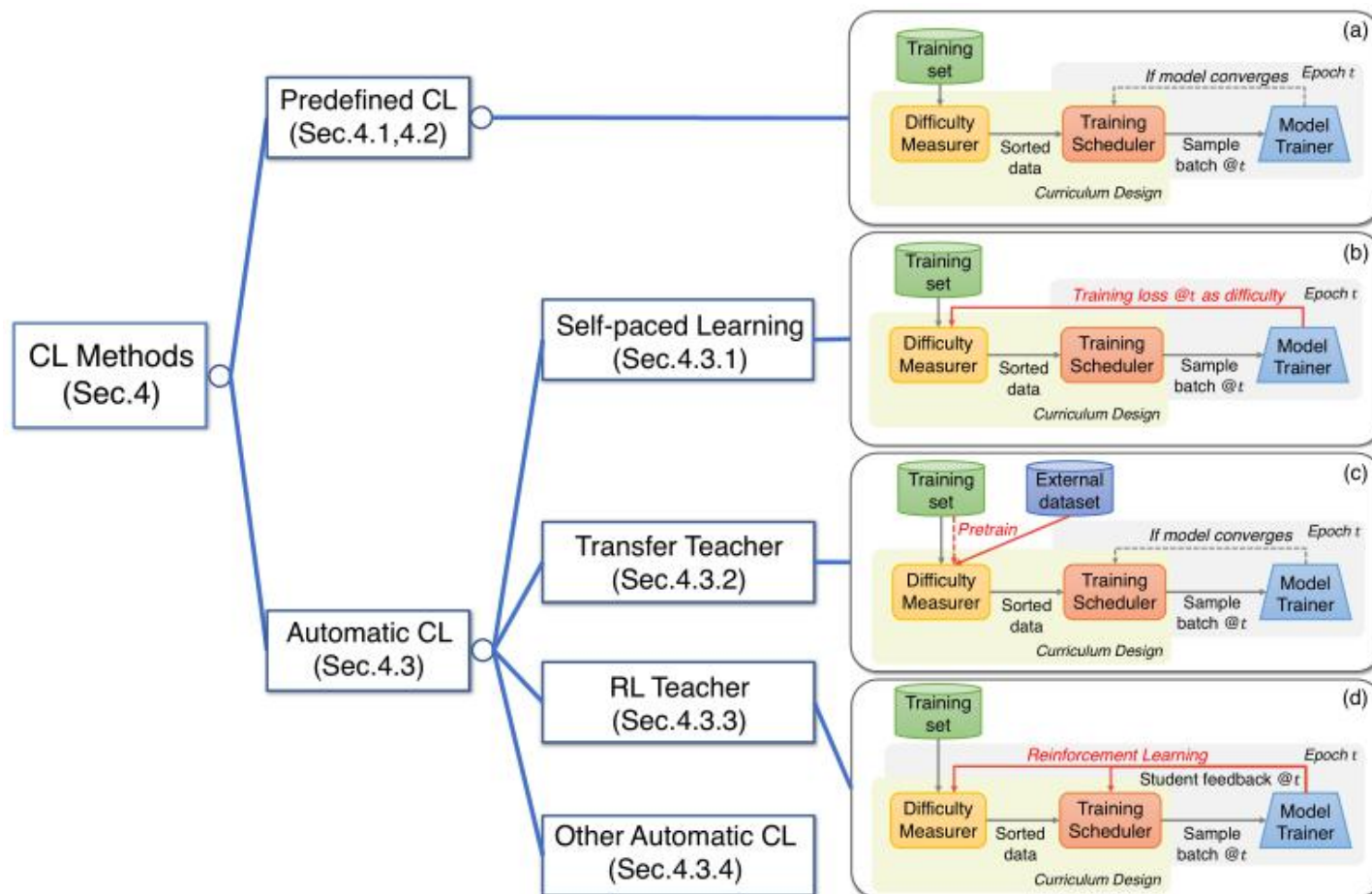
- 课程学习是一种训练策略。
- 模仿人类的学习过程，让模型从简单样本开始，逐渐进阶到复杂样本和知识。

• 动机

- 引导 去噪
- 提高泛化 加速收敛



- 难度评估器
- 训练调度器





04

面向应用的解决方案

知识蒸馏、模型压缩
安全（对抗）、可解释性



在训练过程中，通常使用复杂的模型与大量的计算资源，以从大量、高度冗余的数据集中提取出所需数据，导致如下问题：

1、推断的速度慢

2、对部署设备的要求高

为了解决以上两个问题，针对复杂模型对其进行“蒸馏压缩”。知识蒸馏属于模型压缩的一部分。

理论依据： 使用Teacher-Student模型。

Teacher模型简称Net-T，为较大规模模型，根据输入X都能对应输出Y。

Student模型简称Net-S，为小规模模型，同样能根据输入X都能对应输出Y。

soft target:

在softmax层中引入“温度”参数 T ，使softmax层的映射曲线更加平滑，概率映射更为集中，因此被称为“soft target”。

soft target中梯度方差更小，Net-S学习所需要的数据量大大减少，并且可以使用更高的学习率。

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

步骤:

- 1、训练出一个大模型 (即Net-T)
- 2、在transfer set上利用Net-T对样本生成对应的soft target
- 3、在transfer set上对小模型 (即Net-S) 进行训练, 损失函数包括两部分, 一部分拟合soft target, 另一部分拟合hard target
- 4、保留Net-S, 去掉soft target部分

应用:

对Bert进行蒸馏

Zhao S , R Gupta, Song Y , et al. Extremely Small BERT Models from Mixed-Vocabulary Training[J]. 2019.

目前的训练模型计算复杂度高且存储密集，难以在低内存设备等部署。为了优化模型以便于迁移，在不降低模型的性能的前提下进行模型的压缩和加速。

1、参数修剪与共享

2、低秩分解

3、迁移/压缩卷积滤波器

4、知识蒸馏

参数修剪与共享：

- 1、量化和二值化 减少每个权重所占存储空间来压缩原始网络
- 2、剪枝和共享 在预先训练的CNN模型中修剪冗余的，非信息量的权重
- 3、设计结构化矩阵 使用更少的参数用于描述矩阵，减少全连接层的参数冗余

低秩分解：

由于卷积需要大量的计算量，因此减少卷积的层数可以对模型进行压缩和加速。

将较大的卷积核分解为两个级联的行卷积核与列卷积核

迁移/压缩卷积滤波器:

设计特殊结构的卷积滤波器来减少存储和计算的复杂度

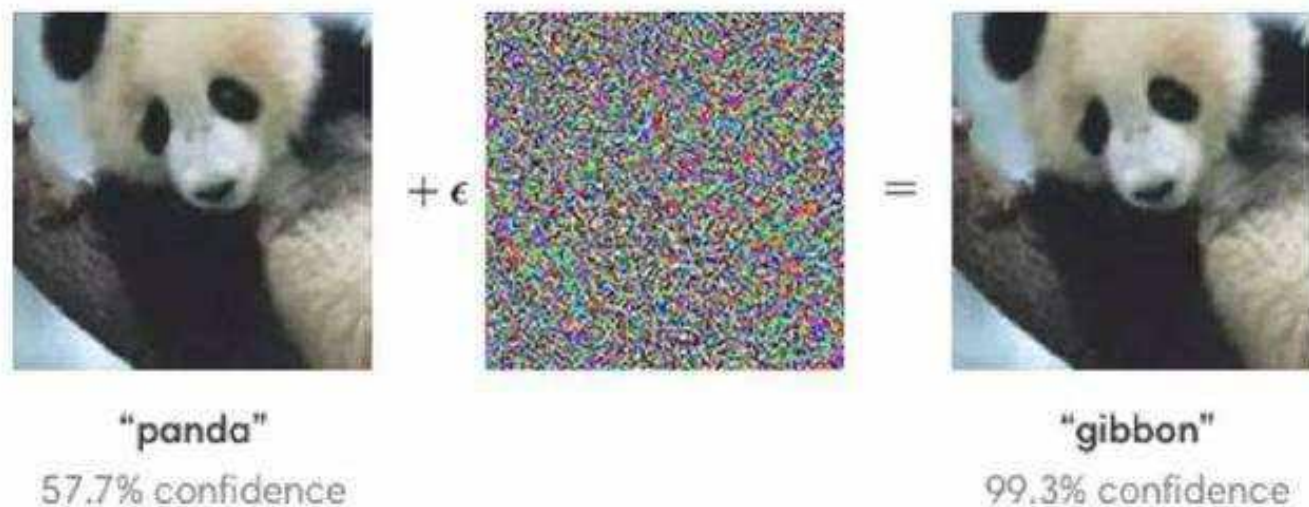
理论依据: 等变群理论 $\mathcal{T} \Phi(\mathbf{x}) = \Phi(\mathcal{T}\mathbf{x})$

使用变换矩阵 $\mathcal{T}(\cdot)$ 转换输入 $\mathbf{x}$, 然后将其传送至网络或层 $\Phi(\cdot)$, 其结果和先将 $\mathbf{x}$ 映射到网络再变换映射后的表征结果一致。

应用：

论文	剪枝	因子	蒸馏	权重共享	量化	预训练	下游任务
《压缩 BERT:研究权重剪枝对迁移学习的影响》	☑					☑	☑
《16 个 Head 真的比 1 个 head 好吗?》	☑						☑
《基于 BERT 剪枝的问答模型》	☑						☑
《通过结构化丢弃按需减少 Transformer 深度》	☑					☑	
《大规模语言表示的重加权近端剪枝》	☑					☑	
《增强拉格朗日模型的结构化剪枝》		☑					☑
《ALBERT：语言表示的自监督学习 Lite BERT》		☑		☑		☑	
《具有最优词语和分享投影的极限语言模型压缩》			☑			☑	
《DistilBERT，蒸馏版本的 BERT：更小、更快、更便宜、更轻量级》			☑			☑	
《从 BERT 中将特定任务知识蒸馏为简单神经网络》			☑				☑
《利用无标签迁移数据将 Transformers 蒸馏为简单神经网络》			☑				☑
《当专注的学生遇到多任务老师：对预训练模型的知识蒸馏改进》			☑			多任务	☑
《BERT 模型压缩的耐心知识蒸馏》			☑				☑
《TinyBERT：用于自然语言理解的蒸馏 BERT》			☑			☑	☑
《MobileBERT：基于渐进知识迁移的 BERT 无偏任务压缩》			☑			☑	
《Q8BERT：8 比特 BERT 量化》					☑		☑
《Q-BERT：基于 Hessian 的超低精度量化 BERT》					☑		☑

深度神经网络在解决复杂问题中具有较好的效果，但是它们很容易受到外界的攻击，例如对输入的扰动，导致最终模型输出不正确的输出。



神经网络的两个特性（《Intriguing properties of neural networks》）

- 1、神经网络中包含语义信息的部分并不在每个独立的神经元，而是整个空间
- 2、神经网络学习到的从输入到输出的映射很大程度上是不连续的

该论文证明了图片经适当修改后能够欺骗神经网络的可能性

一种欺骗神经网络的方法：快速梯度符号法 (Fast Gradient Sign Method, FGSM)

The Fast Gradient Sign Method

$$J(\tilde{x}, \theta) \approx J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x).$$

Maximize

$$J(x, \theta) + (\tilde{x} - x)^\top \nabla_x J(x)$$

subject to

$$\|\tilde{x} - x\|_\infty \leq \epsilon$$

$$\Rightarrow \tilde{x} = x + \epsilon \text{sign}(\nabla_x J(x)).$$

防御对抗攻击的方法：

- 1、修改训练样本，通过添加更多的对抗样本到训练集中可以有效避免一部分攻击。
- 2、修改训练网络，对训练网络做出一定调整，包括在最后一层使用更加非线性的激活函数，但这种方式又会导致训练效率和效果下降。修改训练网络的方法分为完全抵抗和仅检测两种方式。
- 3、附加网络，在不改变原有模型的情况下使用额外的网络进行辅助，其中最有效的一种方式生成式对抗网络——GAN。同样分为完全抵抗和仅检测两种方式。

应用:

通过对抗训练进行文本生成

Zhang Y , Gan Z , Carin L . Generating Text via Adversarial Training.

对于多个领域，神经网络有着较好的效果。然而，例如图像识别领域，在经过对输入图像的干扰后，神经网络会输出错误的识别结果。因此，这需要对神经网络的潜在机制进行分析，即需要具有一定的可解释性

可解释性的定义：

可解释性是用可理解的术语向人类提供解释的能力。

- 1、理想情况下，解释应符合逻辑决策规则（if-then规则），或者可以转换为逻辑规则。
- 2、可理解的术语应该来自与任务相关的领域知识（或根据任务的常识）。

可解释性的重要性：

- 1、高可靠性要求。由于神经网络会出现不可避免的意想不到的失败。可解释性可以协助解决这些故障，使系统更加可靠。
- 2、道德及法律要求。其中包括算法歧视以及解释权。
- 3、科学的使用方法。为了实现模型的优化，即为了提高模型的性能，需要要求模型具有一定的可解释性，更利于对其进行性能方面的优化。

对训练过的网络的被动解释：

- 1、被动，规则作为解释：If P, then Q
- 2、被动，隐藏语义作为解释：寻找输入中的特征点
- 3、被动，归因作为解释：构建输入的特殊模型
- 4、被动，通过举例作为解释：相似的例子会产生相同的输出

对训练中的网络的主动干预性解释：

- 1、主动，规则作为解释：树正则化，能够实现可解释性
- 2、主动，隐藏语义作为解释：对输入中的特征点利用模板进行映射
- 3、主动，归因作为解释：可解释性正则化，具有与先验相似的归因
- 4、主动，通过举例作为解释：将原型层合并到网络，根据输入与学习模型的接近程度进行预测

应用:

对RNN进行可解释性试验

Karpathy A , Johnson J , Li F F . Visualizing and Understanding Recurrent Networks[J]. 2015.



05

面向未来的解决方案

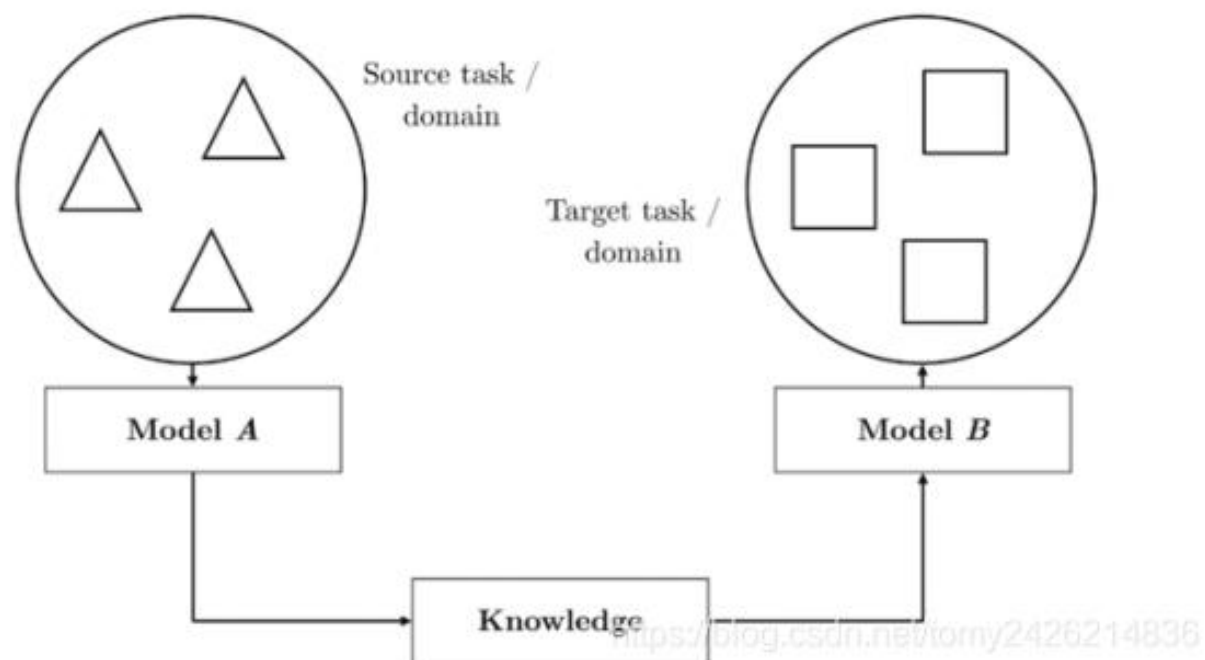
元学习 多任务学习
终身学习 MRC框架



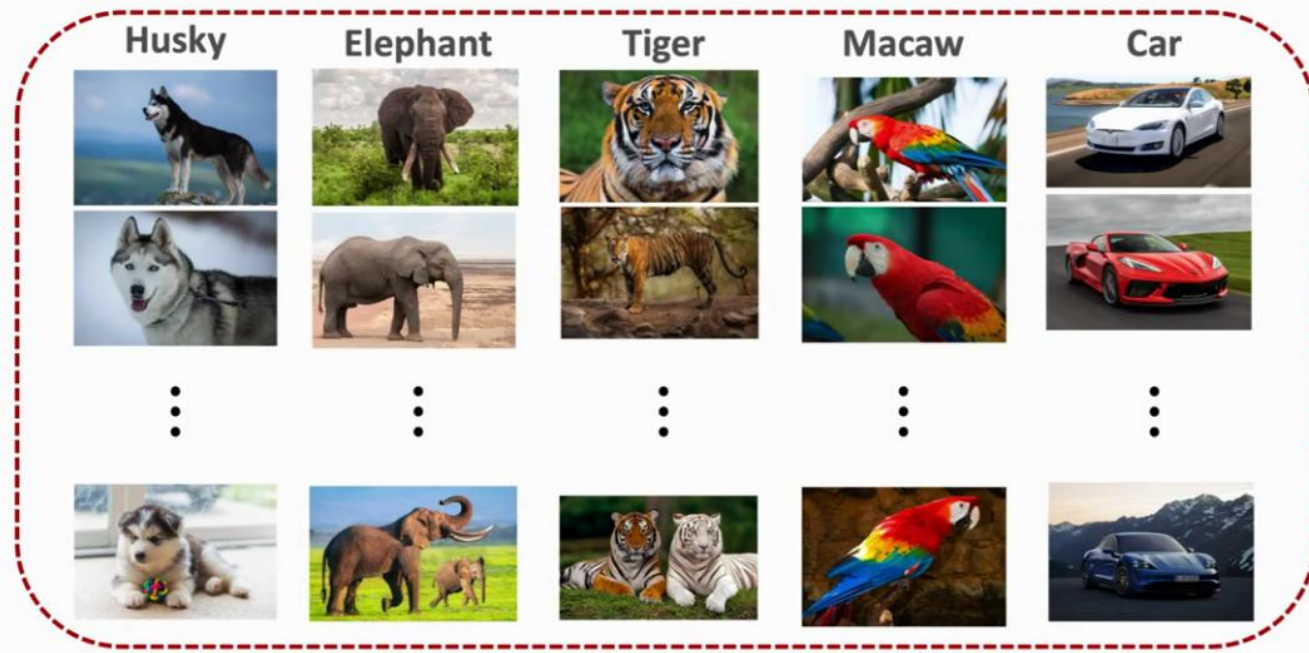
- ⑩ 知道如何骑自行车，学习如何骑摩托车。
- ⑩ 知道如何打乒乓球，学习如何打羽毛球。
- ⑩ 知道如何弹钢琴，学习如何拉小提琴



迁移学习的基本模型



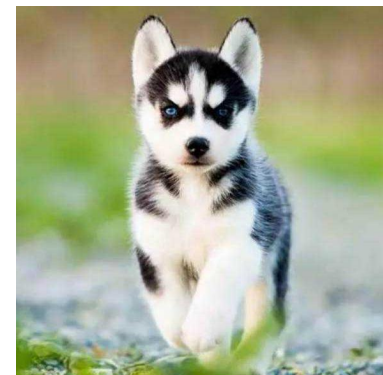
Training Set



大量数据

QUERY1

Husky✓



QUERY2

?



Support Set:

Fox



Squirrel



Rabbit



Hamster



Otter



Beaver



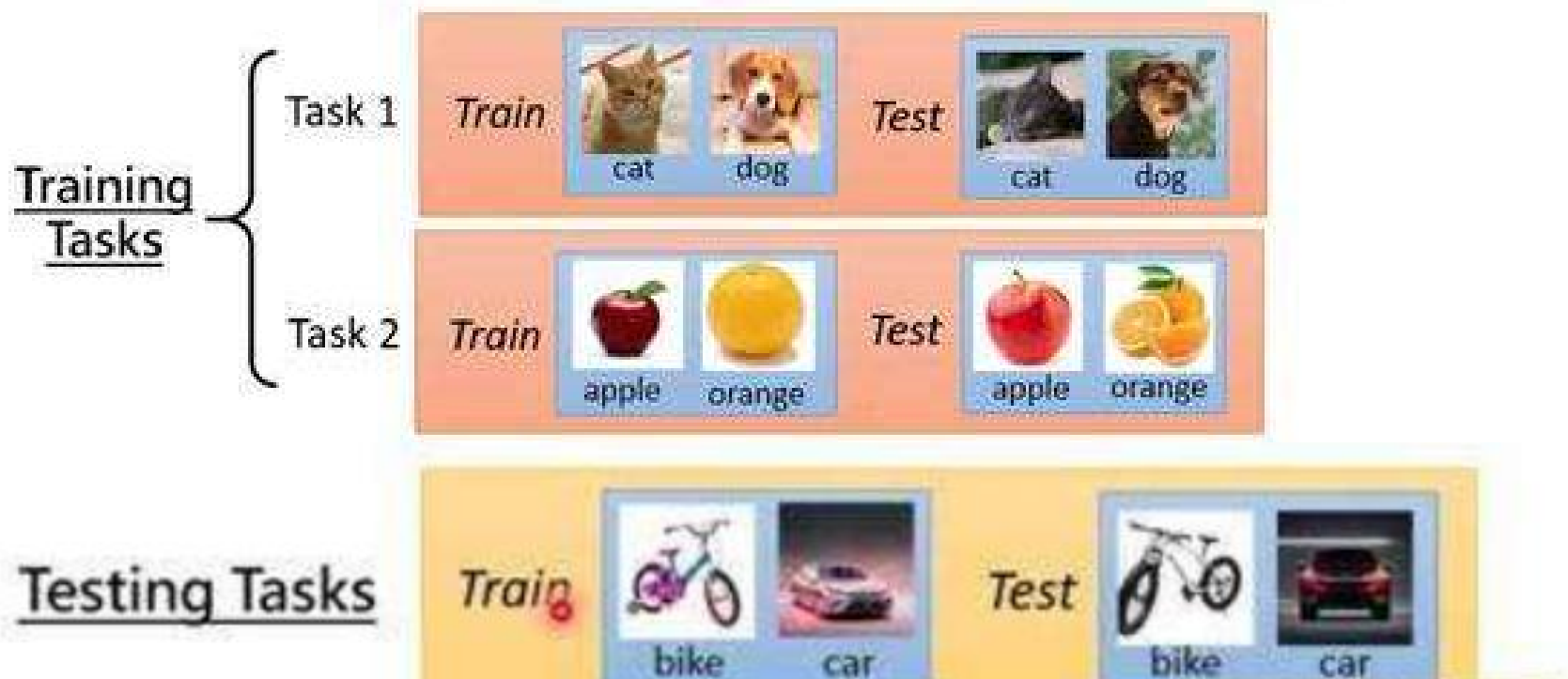
少量数据

QUERY

Squirrel
✓



元学习 *Meta Learning*, 含义为学会学习, 即 *learn to learn*, 就是带着这种对人类这种“学习能力”的期望诞生的。*Meta Learning* 希望使得模型获取一种“学会学习”的能力, 使其可以在获取已有“知识”的基础上快速学习新的任务。



针对NLP问题，元学习的应用可以分为以下两类：

(i) 同一问题不同领域的元学习，学习其他领域，应用于新领域。

- ⑩ 将一个分类作为一个任务。

- ⑩ *Attentive task-agnostic meta-learning for few-shot text classification*

- ⑩ 将一个领域作为一个任务。

- ⑩ *Diverse few-shot text classification with multiple metrics*

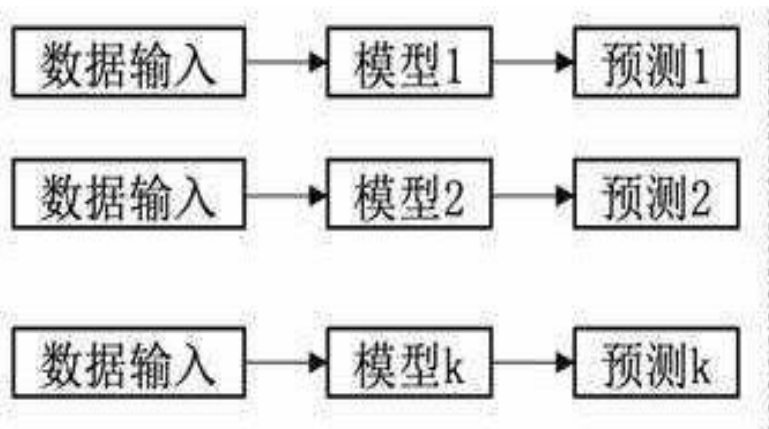
- ⑩ *Induction networks for few-shot text classification*

(ii) 不同问题的元学习，学习其他问题，将其应用于解决新问题。

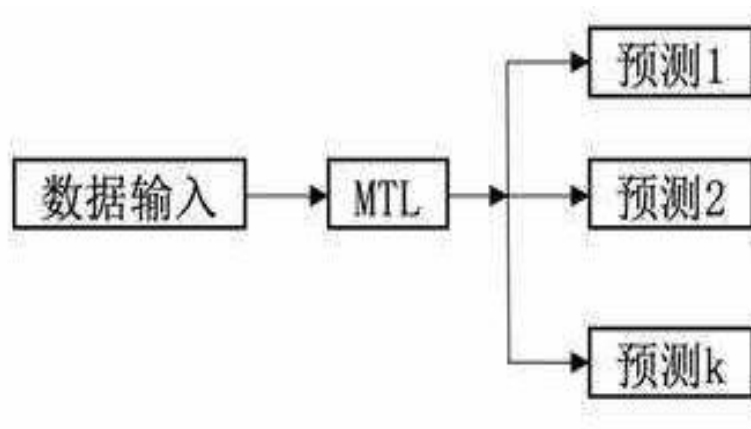
- ⑩ *Meta-learning for low-resource neural machine translation*

- ⑩ *Learning to few-shot learn across diverse natural language classification tasks*

多任务学习 (multi task learning) : 简单来说有多个目标函数loss同时学习的就算多任务学习。

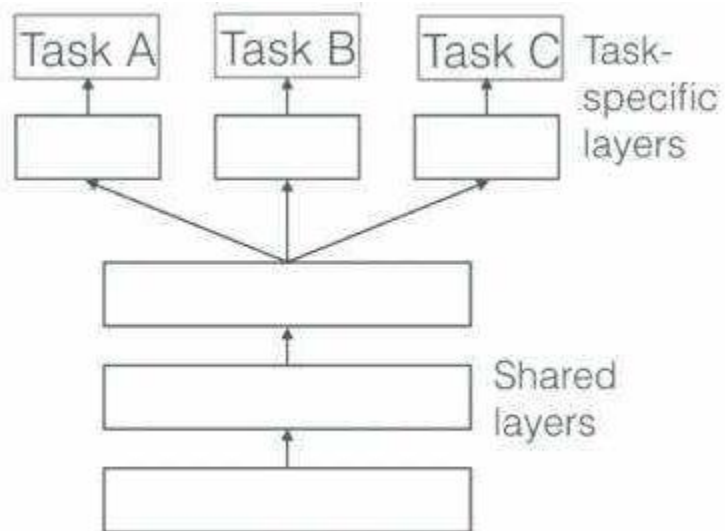


单任务学习

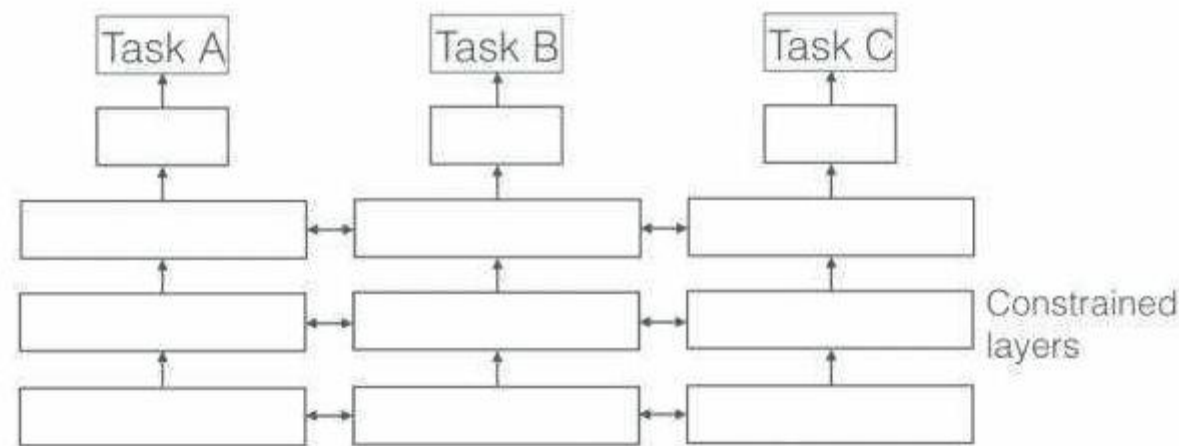


多任务学习

通常将多任务学习方法分为：hard parameter sharing和soft parameter sharing



hard parameter sharing



soft parameter sharing

设计更合理的**辅助任务**。很多时候，我们的仅仅只考虑优化一个目标，看似和多任务学习无关，但其实也可以通过添加合适的辅助任务，来帮助得到更好的结果

目前做的最多的工作的方向也是比较多热门论文的方向主要在于辅助任务的设计上

常见的辅助任务类型如下：

related tasks（具有相关性的任务）

Adversarial（对抗任务）

Hints（带有提示性质的任务）

Focusing attention

Quantization smoothing（量化平滑）

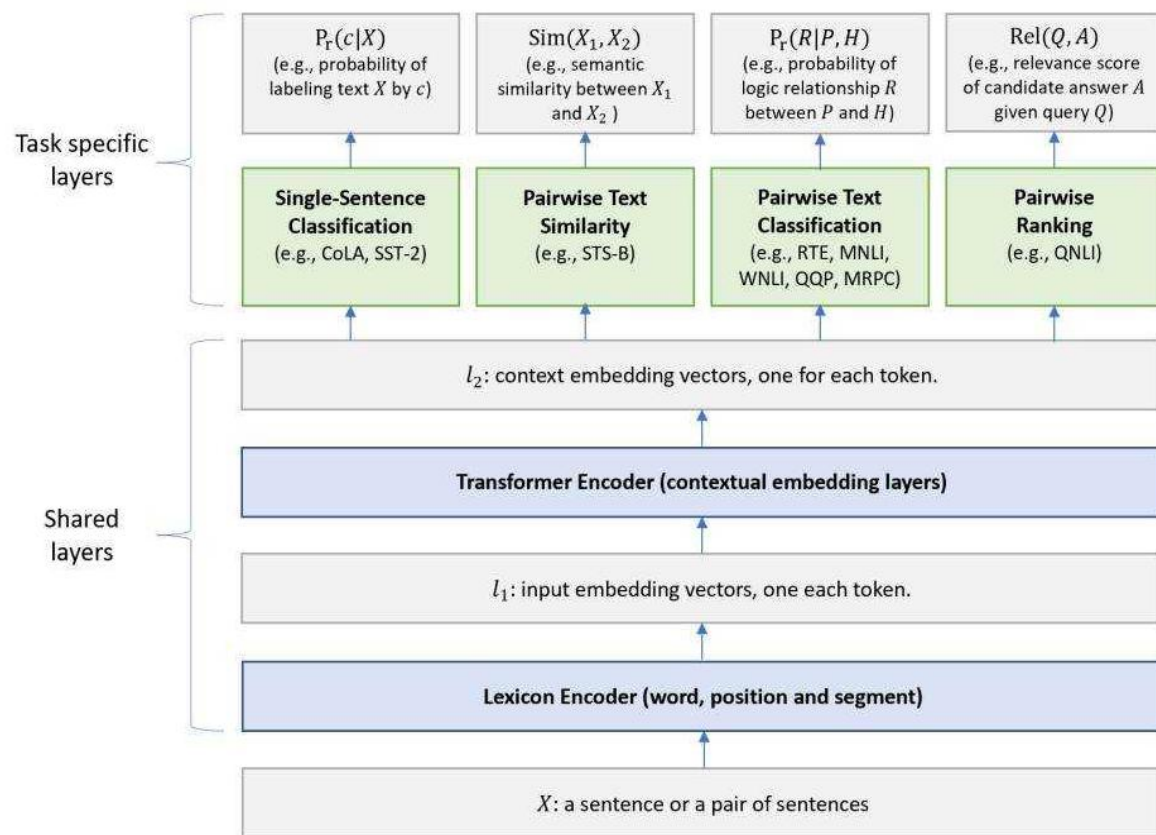
Predicting inputs（预测输入）

using the future to predict the present（用未来预测现在）

Representation learning（表示学习）

微软 MT-DNN模型

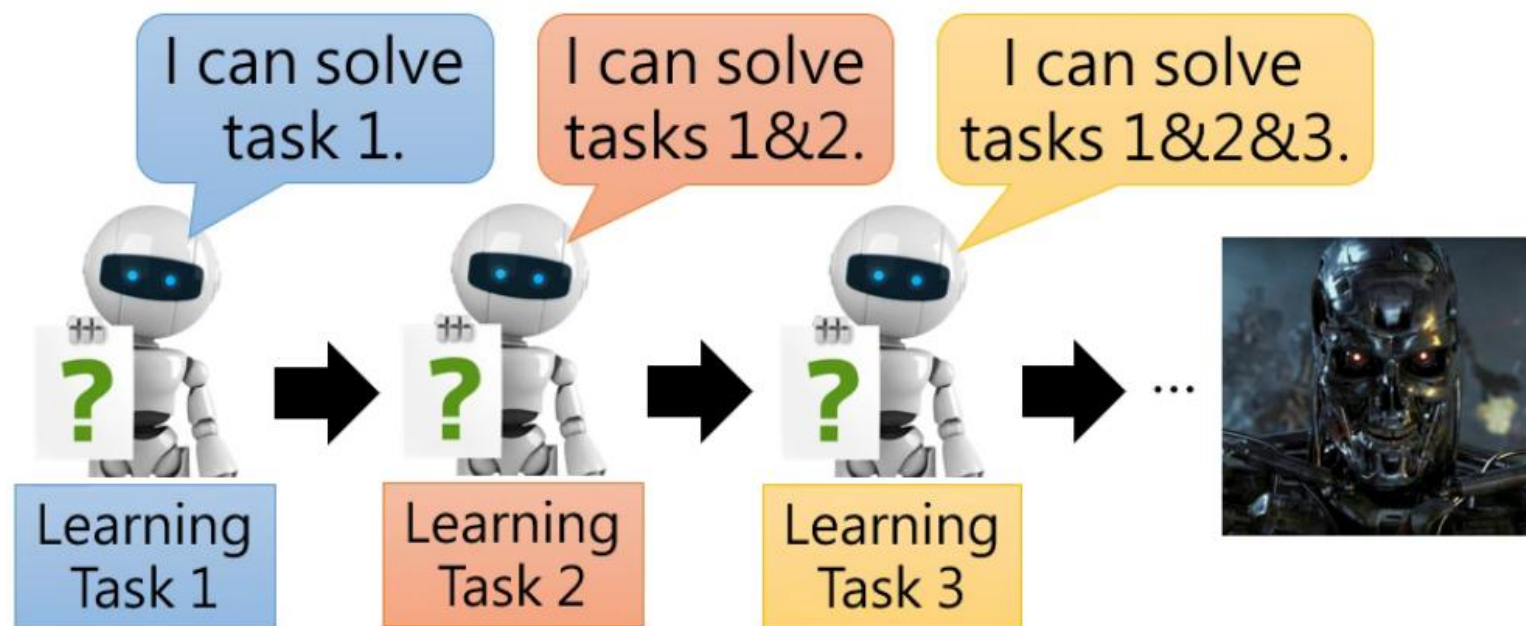
Multi-task deep neural networks for natural language understanding

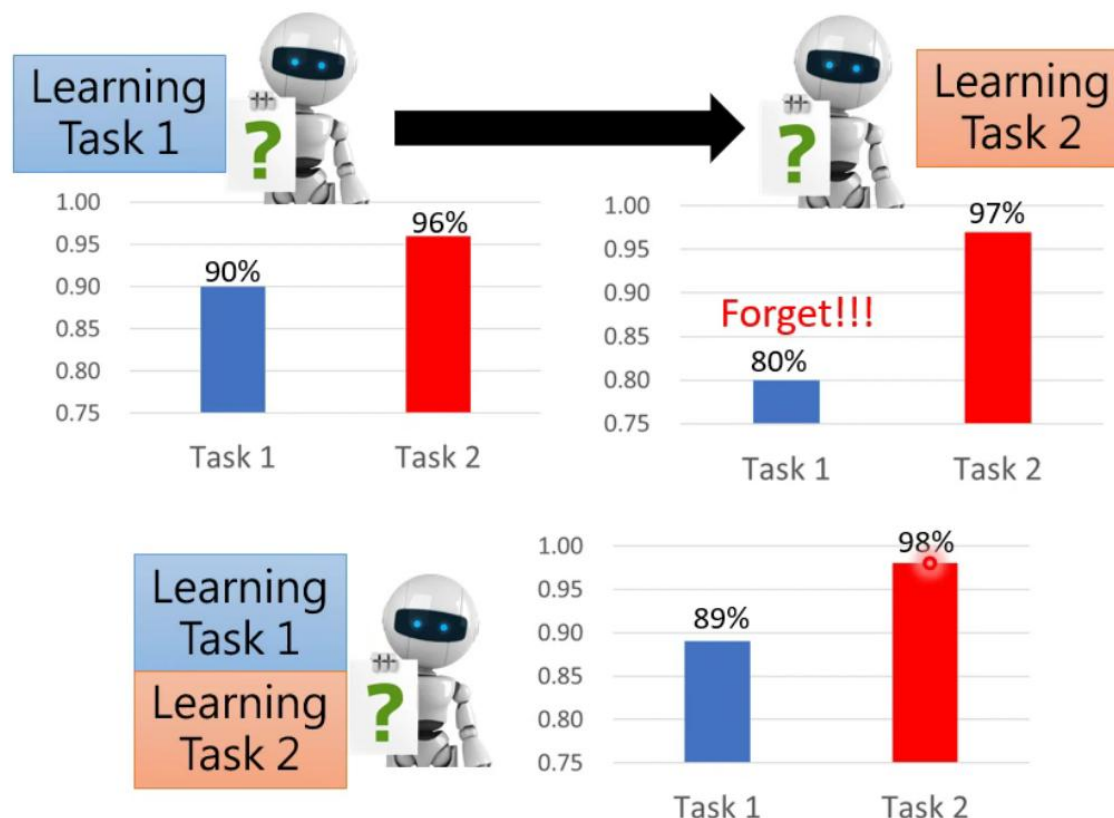
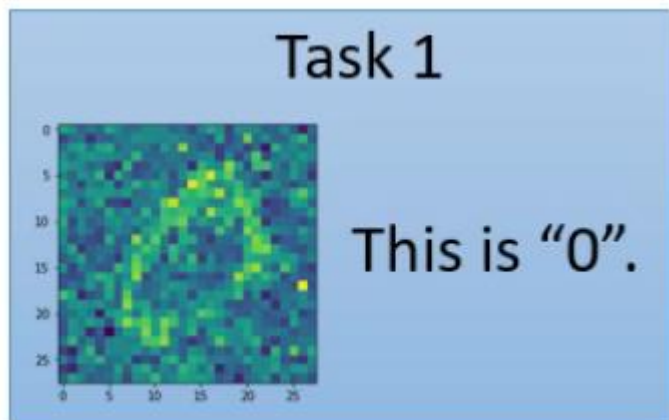


MT-DNN模型结合了四种语言理解（NLU）任务：

- 单句分类
- 两两文本分类
- 文本相似性评分
- 相关性排序

Life Long Learning=Continuous Learning
=Incremental Learning





QA tasks: 给定一段文本，根据文本来回答问题。

Task 5: Three Argument Relations

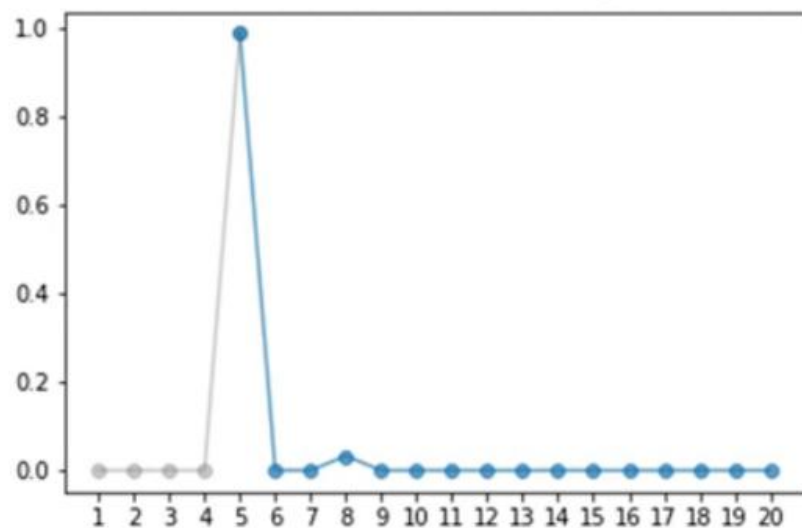
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.

Who gave the cake to Fred?
A: Mary

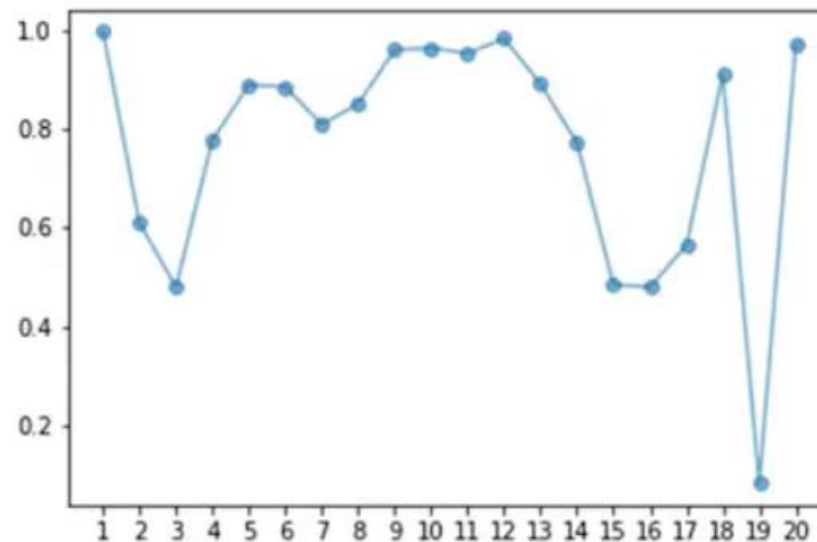
Who did Fred give the cake to?
A: Bill

There are 20 QA tasks in bAbi corpus.
Train a QA model through the 20 tasks.

采用两种方式进行训练：

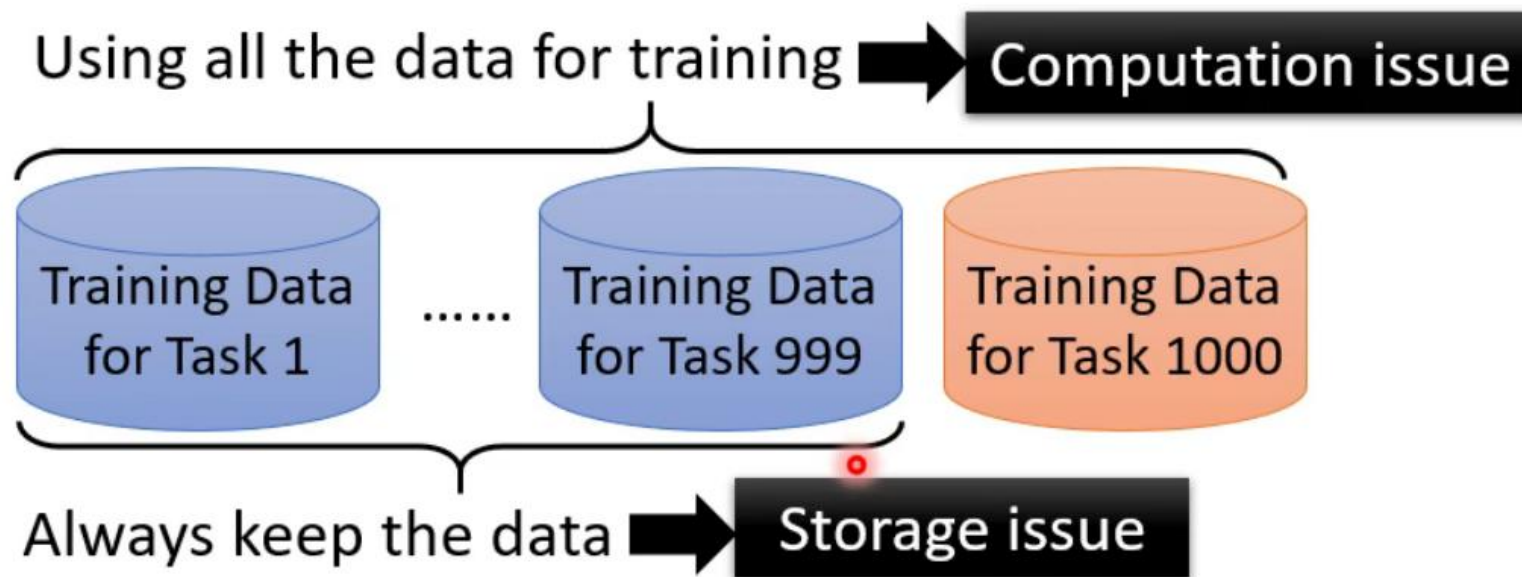


按顺序依次学习20个任务
Task 5的准确率



同时学习20个任务
Task 1-20的准确率

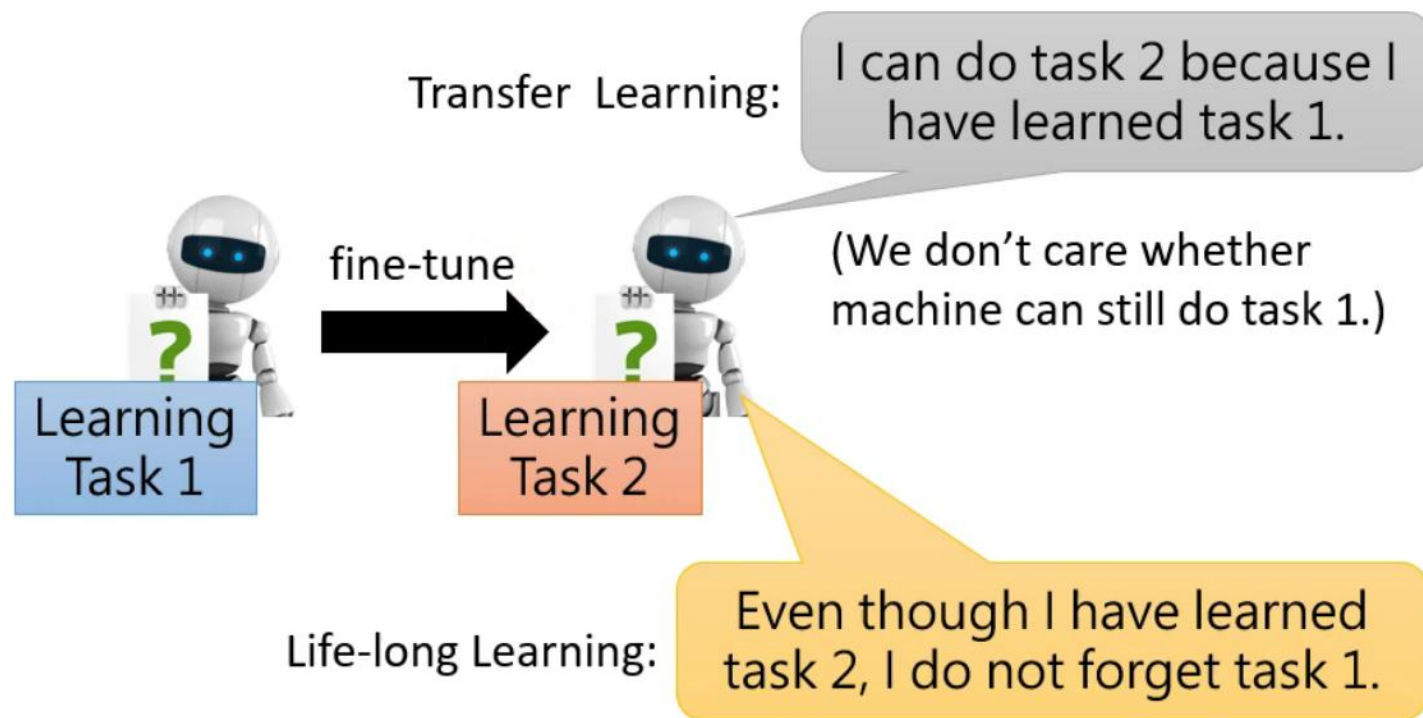
终身学习和多任务学习的对比



Multi-task training可以解决灾难性遗忘的问题，但是需要所有的数据全都放在一起训练，将会占用巨量算力和存储空间。

虽然Multi-task training可以让机器学会多个任务，但它不是Life Long Learning的最终解决方案。它通常被视为Life Long Learning的upper bound。

终身学习和迁移学习的对比



Transfer Learning更关注新任务的完成情况，而Life Long Learning更关注旧任务完成的如何。

MRC使用三元组

C = The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

Q = Where do water droplets collide with ice crystals to form precipitation?

A = within a cloud

尝试将一些NLP任务转换为QA任务，从而MRC框架下解决：

1. NER 命名实体识别
2. Relation Extraction 关系抽取
3. Text Classification 文本分类
4. Co-reference 共指消解

核心问题：如何构造query

生成query的方法：

1.手工定义。

2.在文本分类中对不同的文本生成不同的问题，然后用强化学习的方法进行优化。

这样做的优势：

1.在这些questions中包含了先验知识，这些先验知识可以更明确地告诉模型哪些信息需要注意。

X: at the Chinese embassy in France.....

Q: find all geographical regions, such as a government, a populace, and a geographic location. (Annotation Guidelines)

2.领域自适应能力

3.小样本学习能力

4.灵活性

谢谢大家
敬请批评指正

Thanks for your listening



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

康铠、雷沛珂、杨子研、杨曼芝、马弋洋

2021/09/18