華東師範大學
EAST CHINA NORMAL UNIVERSITY

Unsupervised Representation Learning

李鑫

2019.12.3

Representation Learning：指学习对观测样本X有效的表示，表示的好坏取决于下游任务的性能提升。
例： Pre-training – Fine-tune

Unsupervised/Self-supervised Learning：采用没有人为干预的自然存在的(监督)信号进行辅助学习。
主要方法：构造辅助任务，例如：图像上色

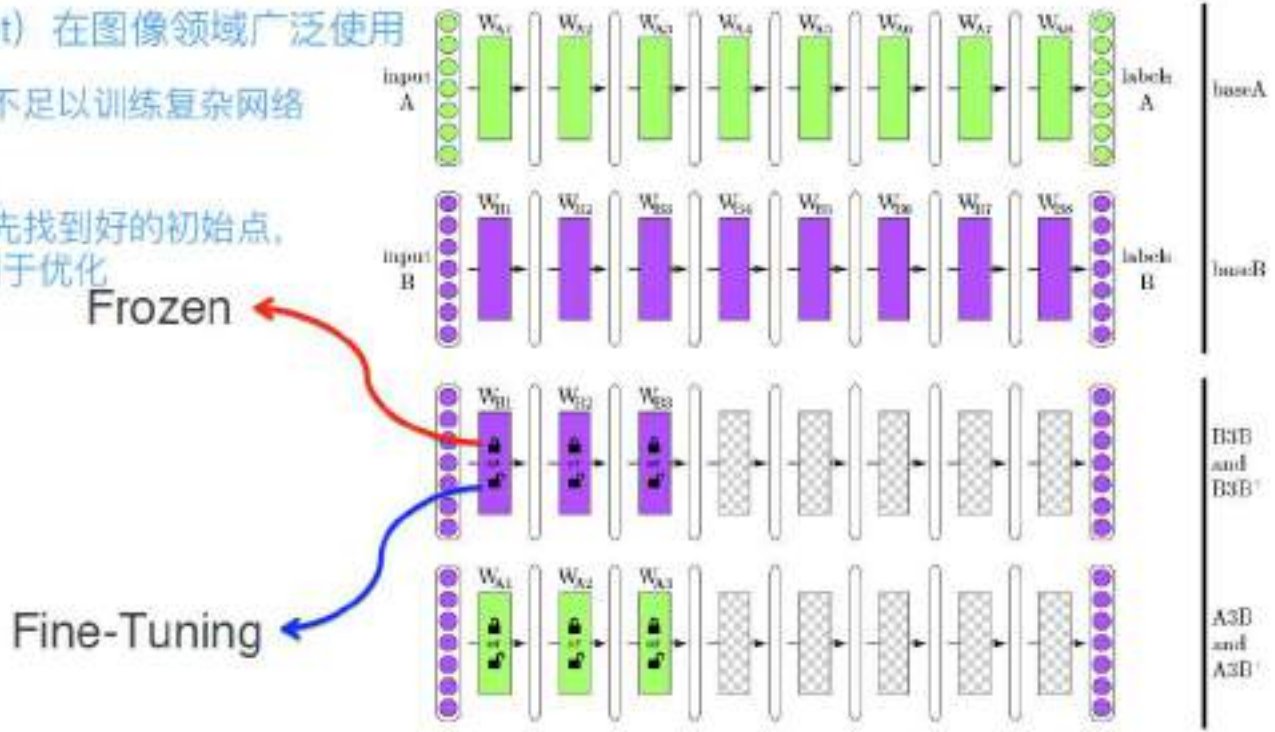
表示学习在图像领域的应用

预训练（say,imagenet）在图像领域广泛使用

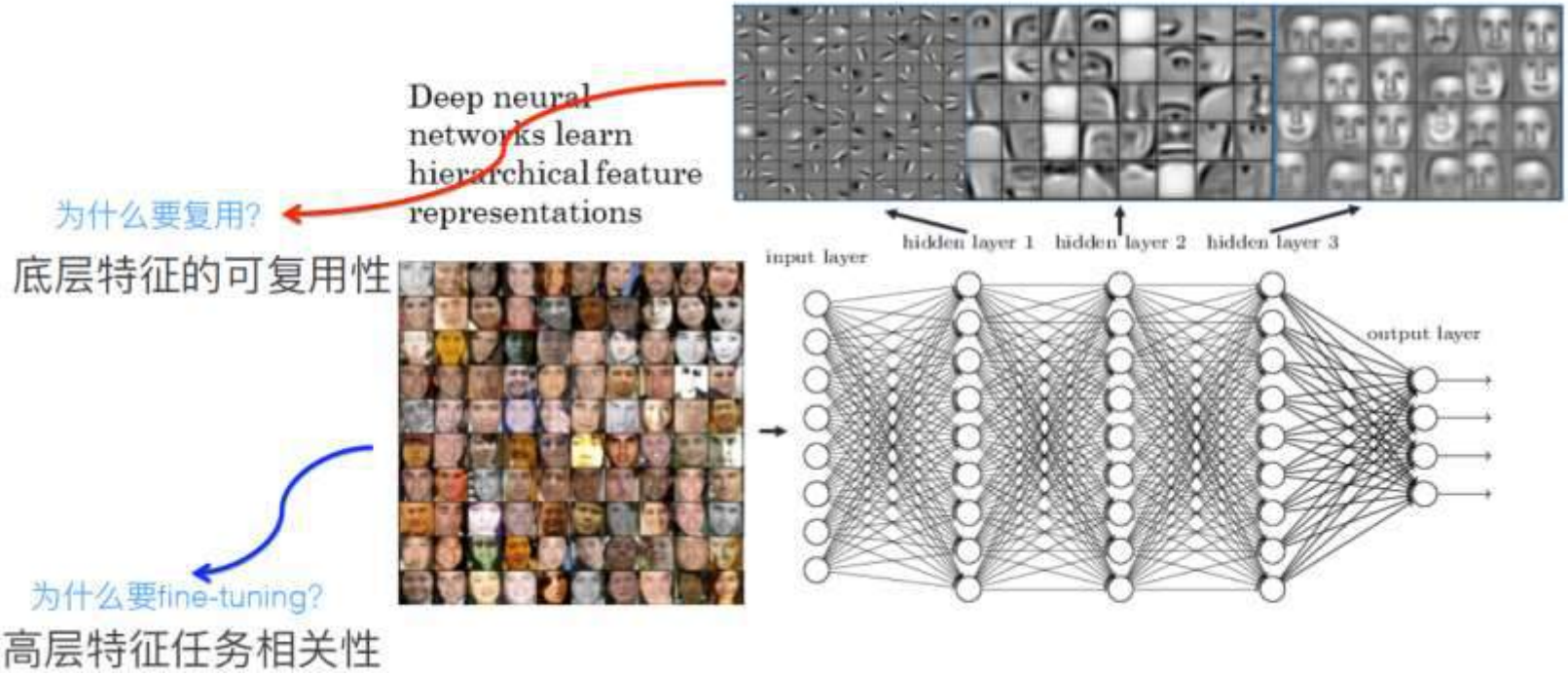
1.训练数据小，不足以训练复杂网络

2.加快训练速度

3.参数初始化，先找到好的初始点，有利于优化



表示学习在图像领域的应用



PS： Rethinking ImageNet Pre-training

表示学习在自然语言处理的应用

语言模型

Sentence 1:美联储主席本·伯南克昨天告诉媒体7000亿美元的救助资金

Sentence 2:美主席联储本·伯南克告诉昨天媒体7000亿美元的资金救

Sentence 3:美主车席联储本·克告诉昨天公司媒体7000伯南亿美行元

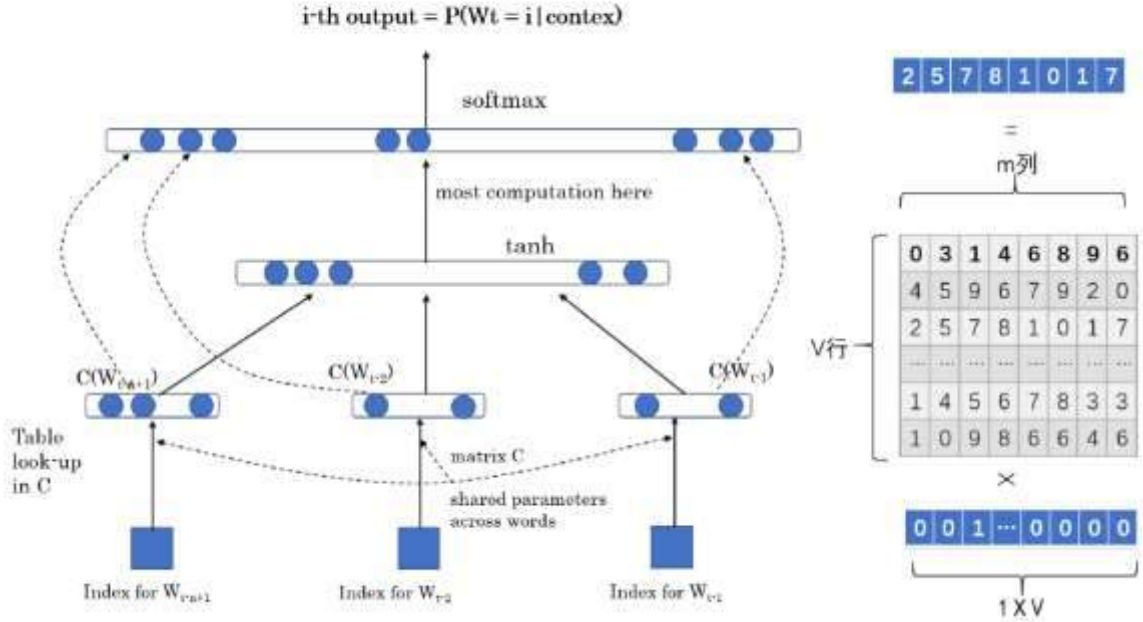
哪个句子更像一个合理的句子？如何量化评估?

$$P(S) = P(w_1, w_2, \ldots, w_n) \qquad P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\cdots P(w_n|w_1, w_2, \cdots w_{n-1})$$

语言模型： $L = \sum_{w \in C} logP(w|context(w))$

表示学习在自然语言处理的应用

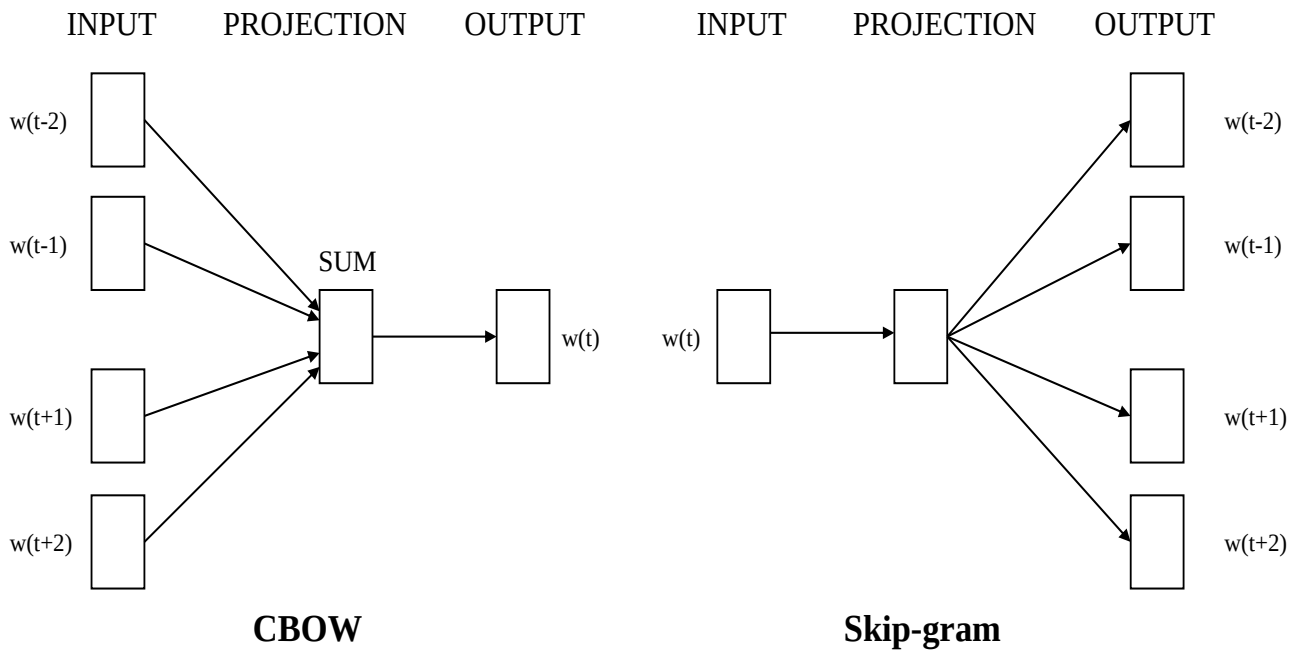
神经网络语言模型（NNLM）



论文：A Neural Probabilistic Language Model, JMLR,2003

表示学习在自然语言处理的应用

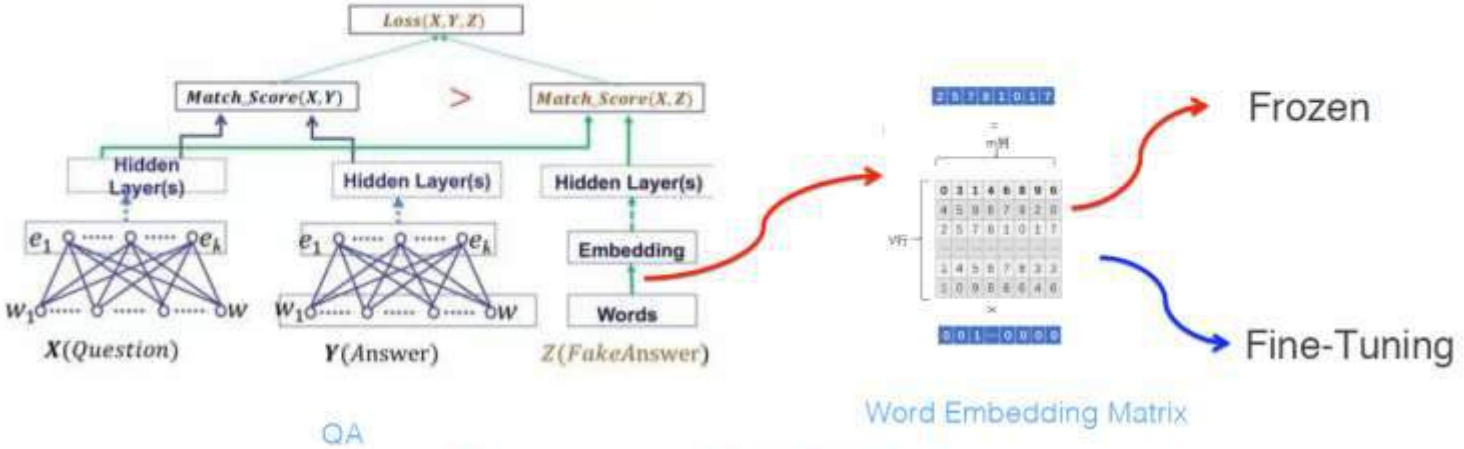
Word2Vec



论文：Efficient Estimation of Word Representations in Vector Space,ICLR,2013

表示学习在自然语言处理的应用

学会了单词的WE，怎么用?



这是18年之前NLP中典型的预训练模式!

表示学习在自然语言处理的应用

有什么问题值得改进?

..very useful to protect banks or slopes from being washed away by river or rain...

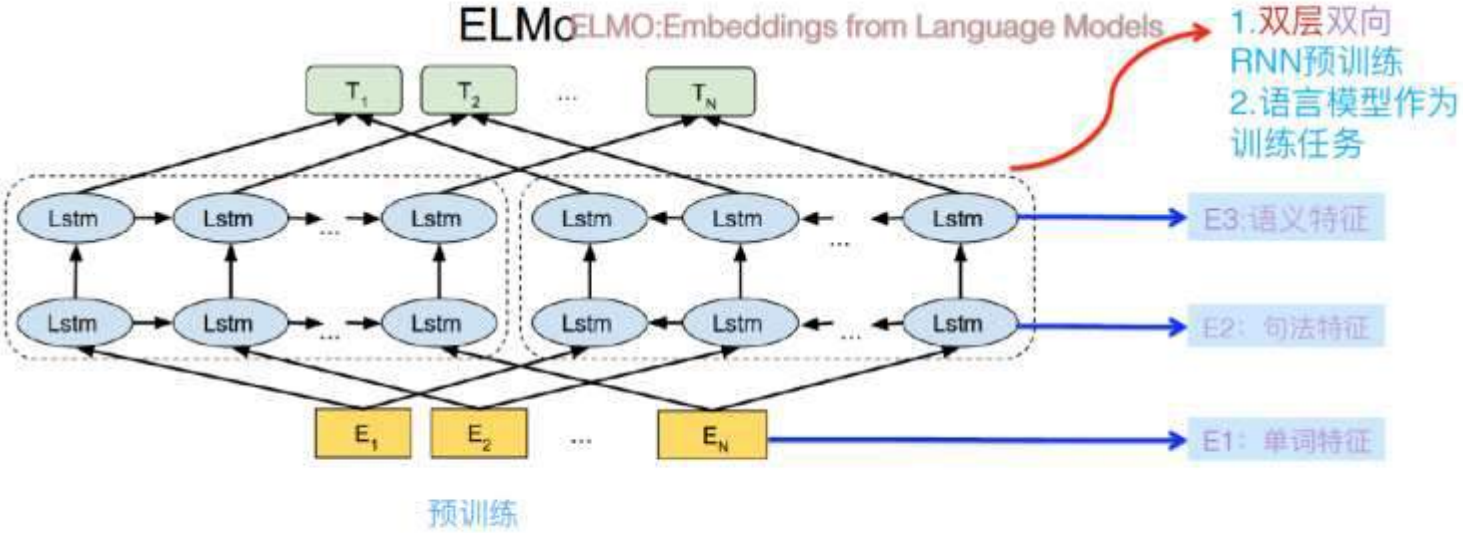
...the location because it was high, about 100 feet above the bank of river...

...The bank has plan to branch throughout the country...

...They throttled the watchman and robbed the bank...

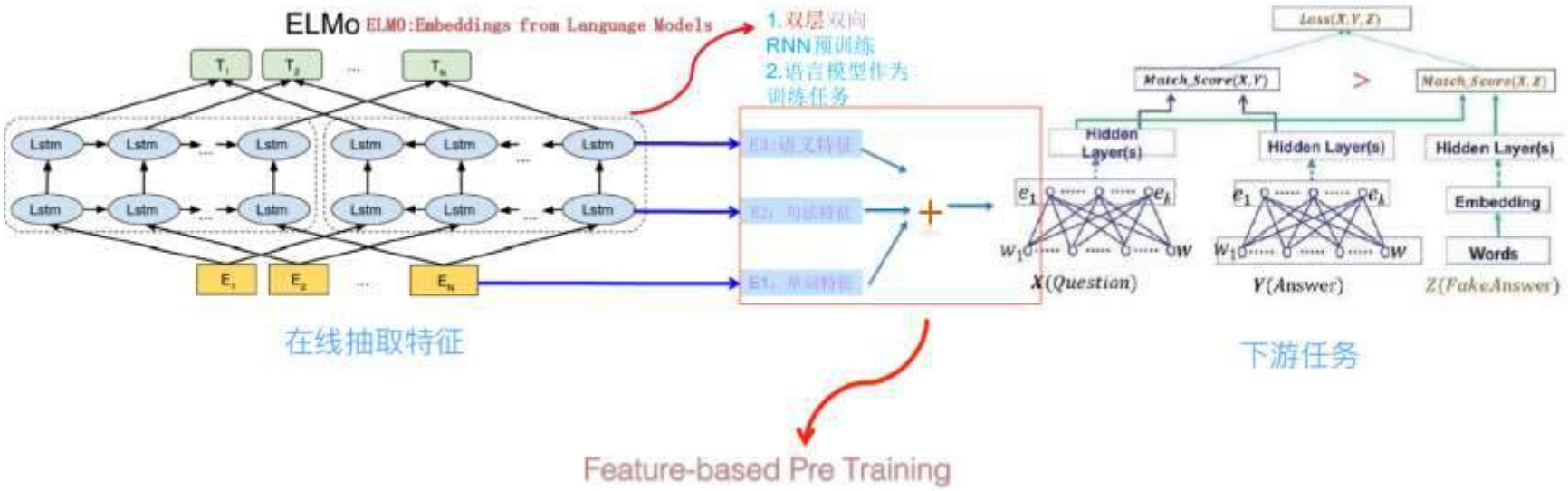
表示学习在自然语言处理的应用

从WE到ELMO：基于上下文的Embedding



NAACL 2018 最佳论文：Deep contextualized word representations

表示学习在自然语言处理的应用



表示学习在自然语言处理的应用

ELMO：多义词问题解决了吗?

(多义词)Play:
1.运动
2.音乐

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

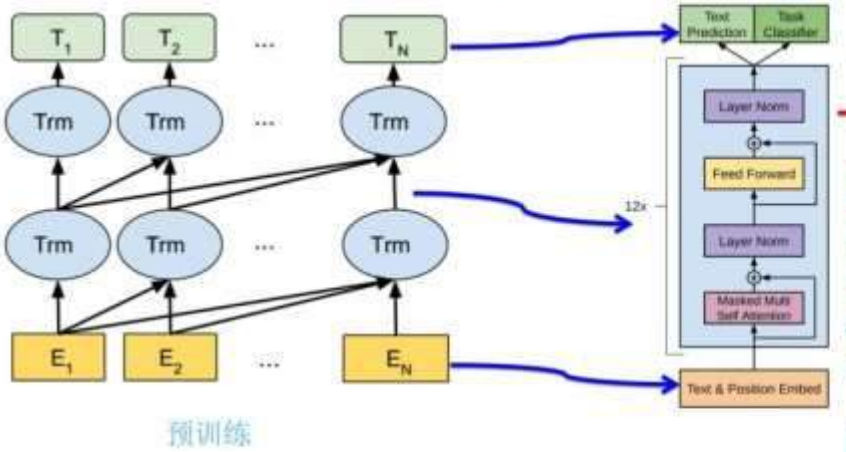
表示学习在自然语言处理的应用

ELMO：效果如何?

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

表示学习在自然语言处理的应用

从WE到GPT：Pretrain+Finetune两阶段过程

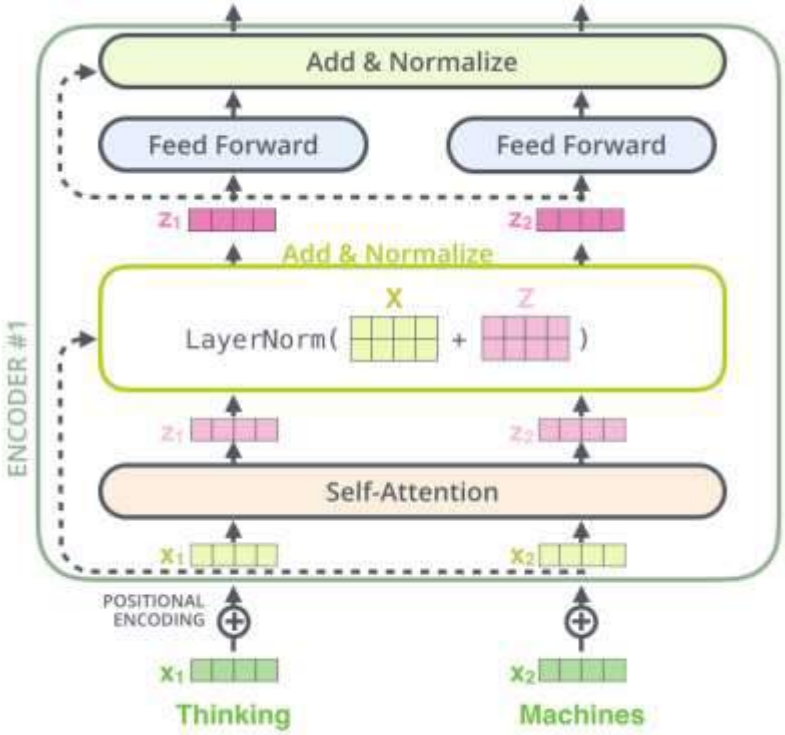


1.Transformer特征抽取器
2.语言模型作为训练任务（单向）

1.Transformer由Google2017年提出
2.本质上是个self attention叠加结构
3.目前是效果最好的特征抽取器，RNN在未来要被替代
4.优点：易于并行（RNN弱项）；捕获长距离特征能力强（Transformer>LSTM>CNN）;有深度有内涵

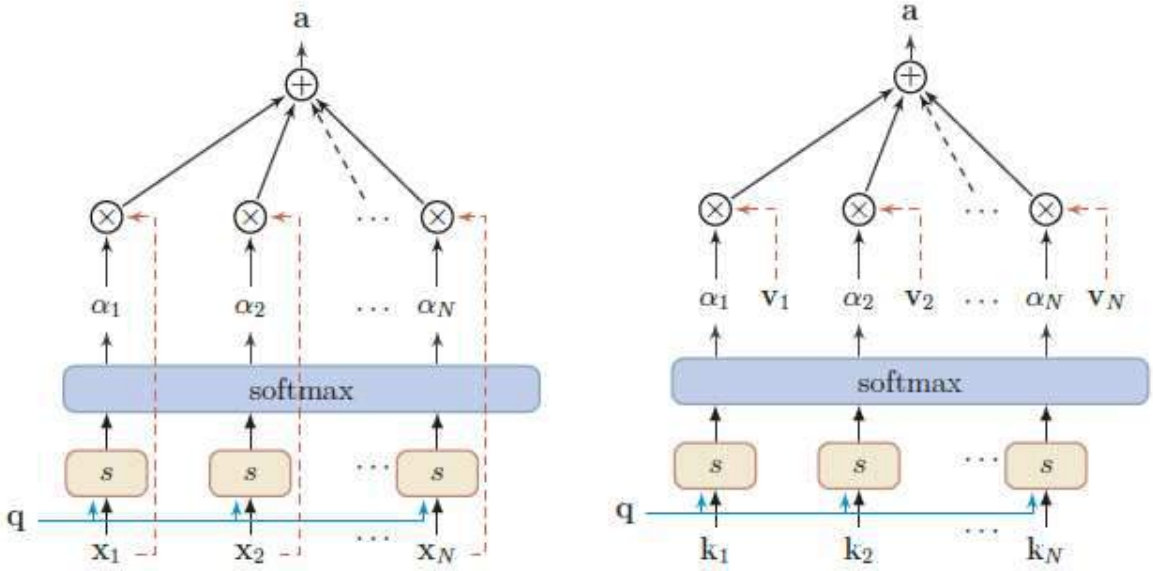
论文：Improving Language Understanding by Generative Pre-Training

表示学习在自然语言处理的应用



https://jalammar.github.io/illustrated-transformer/

表示学习在自然语言处理的应用



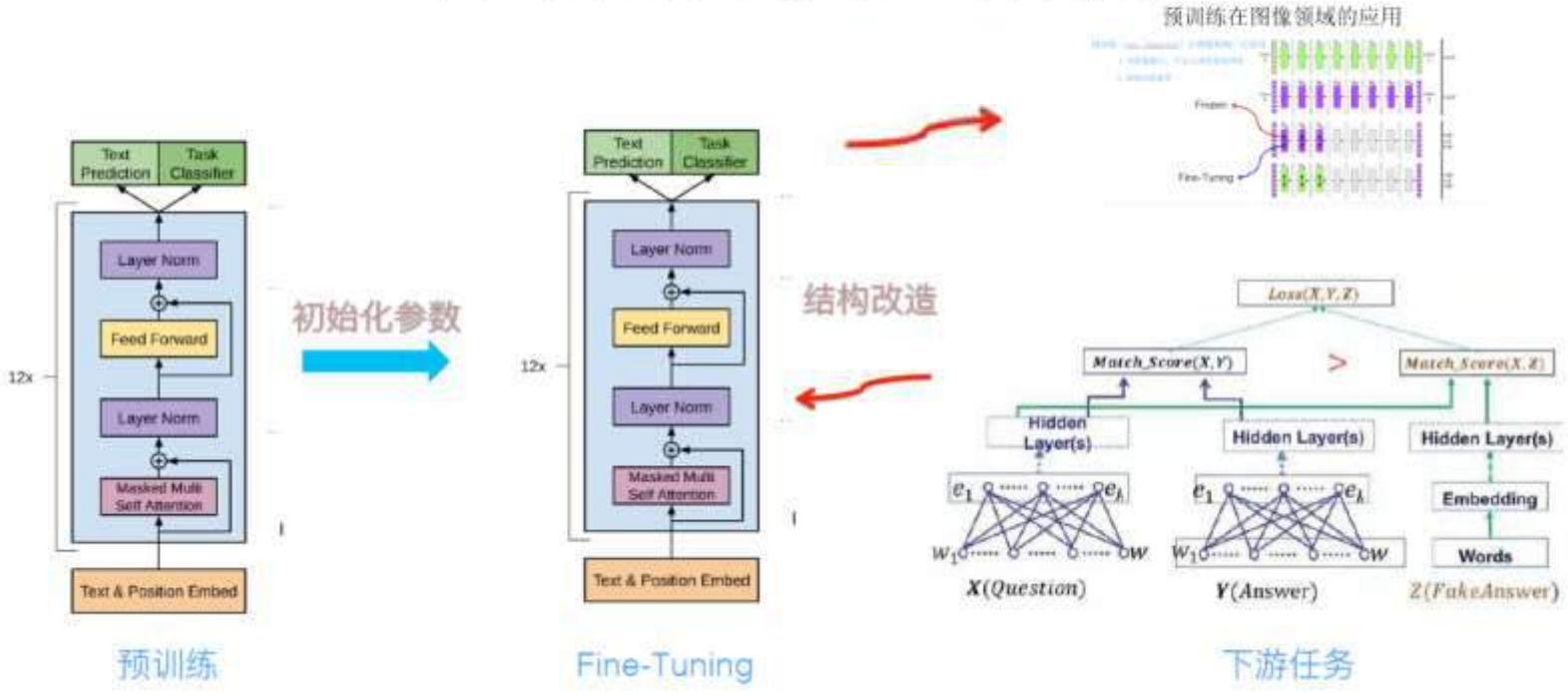
(a) 普通模式　　(b) 键值对模式

$$\mathbf{Attention(Query, Source)} = \sum_{i=1}^{L_x} \boldsymbol{Similarity(Query, Key_i) * Value_i}$$

论文:Attention Is All You Need,NIPS,2017

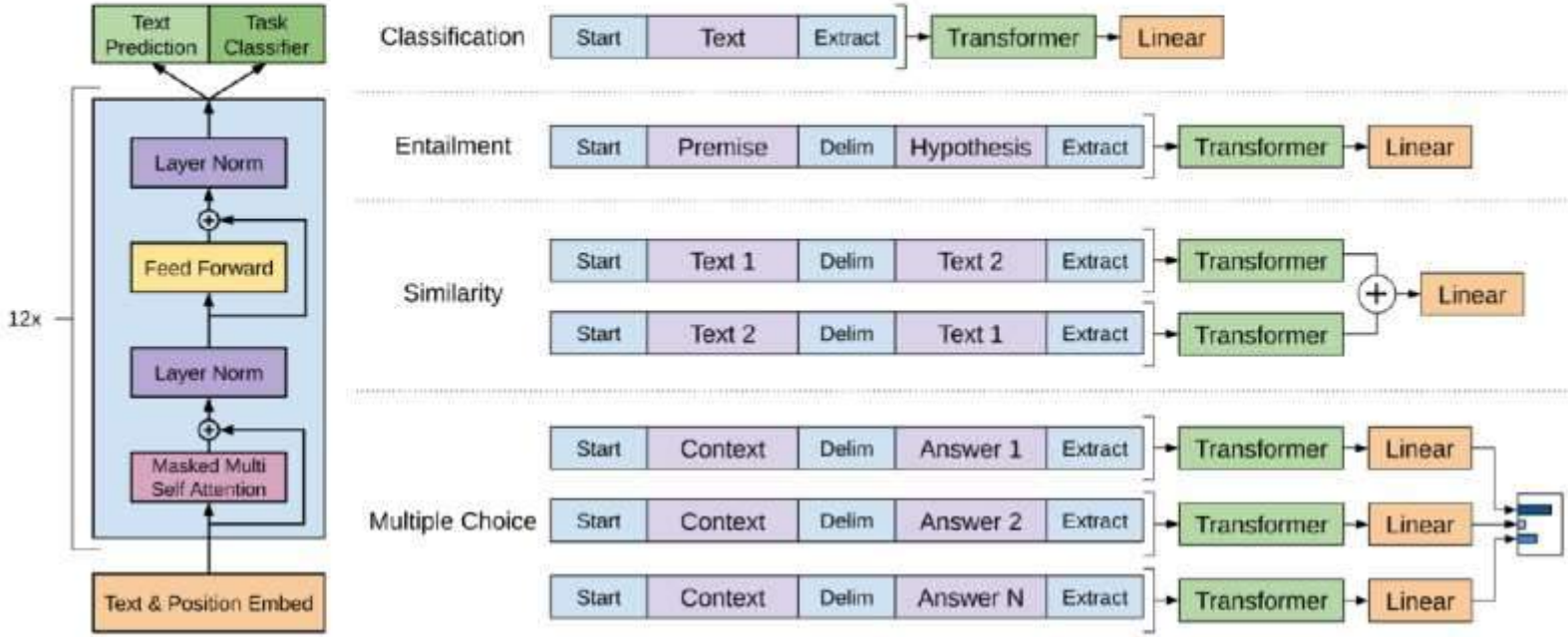
表示学习在自然语言处理的应用

GPT：训练好之后如何使用？



表示学习在自然语言处理的应用

GPT：如何改造下游任务？



表示学习在自然语言处理的应用

GPT：效果如何?

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

12个NLP任务：9个达到最好效果!

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Table 4: Semantic similarity and classification results, comparing our model with current state-of-the-art methods. All task evaluations in this table were done using the GLUE benchmark. (*mcc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

表示学习在自然语言处理的应用

GPT：有效因子分析

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

1.Transformer特征抽取能力远强于LSTM

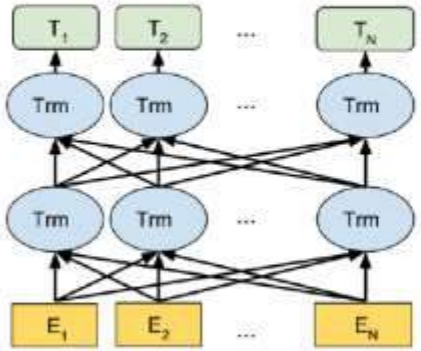
2.预训练至关重要

表示学习在自然语言处理的应用

从GPT和ELMO及Word2Vec到Bert：新星的诞生

1.Transformer特征抽取器
2.语言模型作为训练任务（双向）

BERT (Ours)

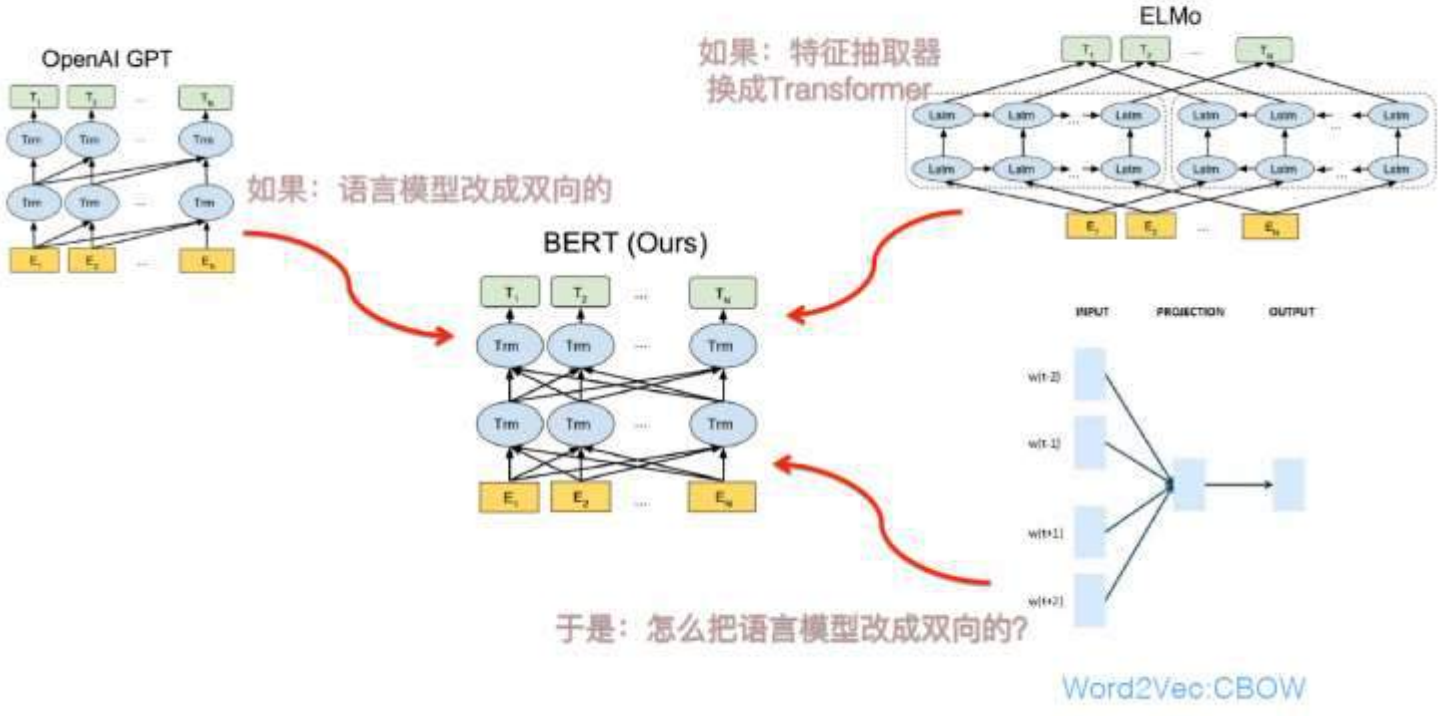


预训练

论文：BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

表示学习在自然语言处理的应用

从GPT和ELMO及Word2Vec到Bert：四者的关系



表示学习在自然语言处理的应用

Bert：效果如何?

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition results. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results were scored against the hidden labels by the SWAG authors. †Human performance is measure with 100 samples, as reported in the SWAG paper.

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

11个NLP任务：全面提升效果!

总结

1.BERT并未有重大创新，更像是最近几年NLP的集大成者

2.充分利用大量无监督数据作为预训练

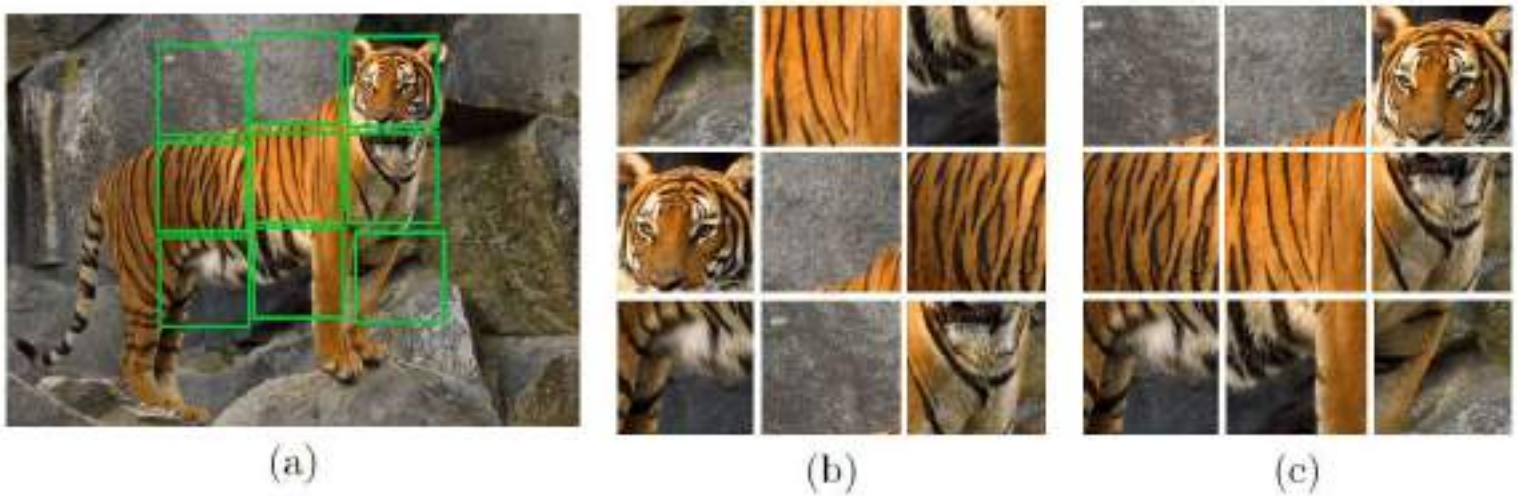
3.两阶段模型，超大规模预训练+具体任务fine tune将成为趋势，BERT彻底提高了预训练词向量的重要性。

4.关于特征提取，transform将逐步取代CNN和RNN

自监督学习

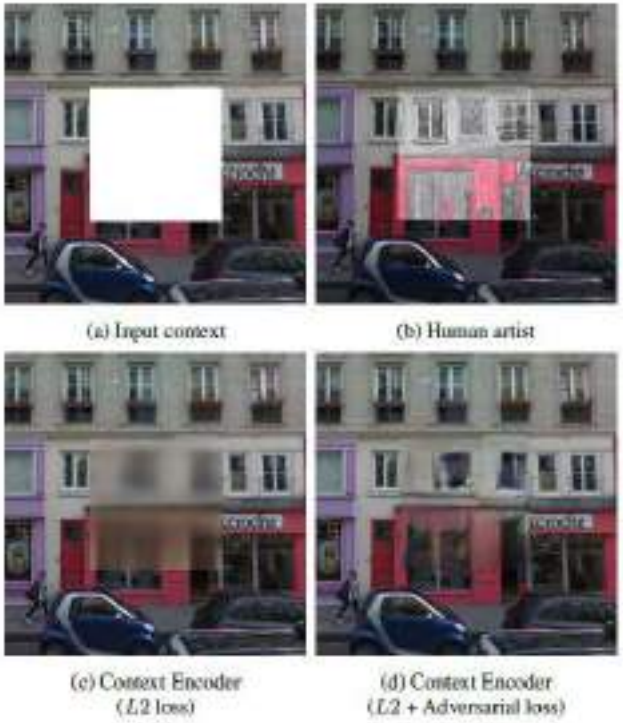
- 什么是自监督学习
  - 使用**自然存在的监督信号**用于训练
  - 基本没有人为干预
- 为什么需要自监督学习
  - 表示学习的重要性 Pre-training – Fine-tune模式
  - 自监督学习能够充分利用自身的标签进行表示学习

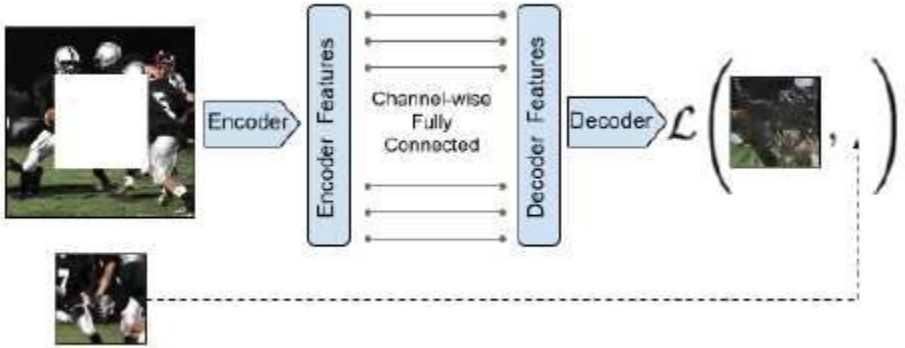
表示学习在图像领域的应用



论文：Unsupervised Learning of VisualRepresentations by Solving Jigsaw Puzzles, ECCV,2016

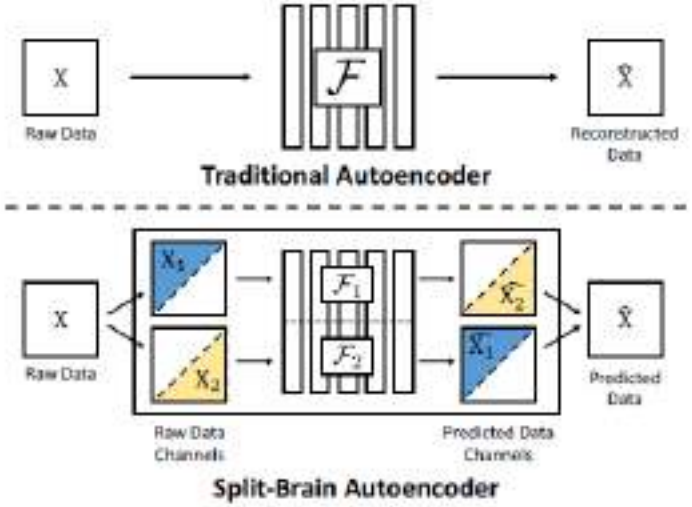
表示学习在图像领域的应用

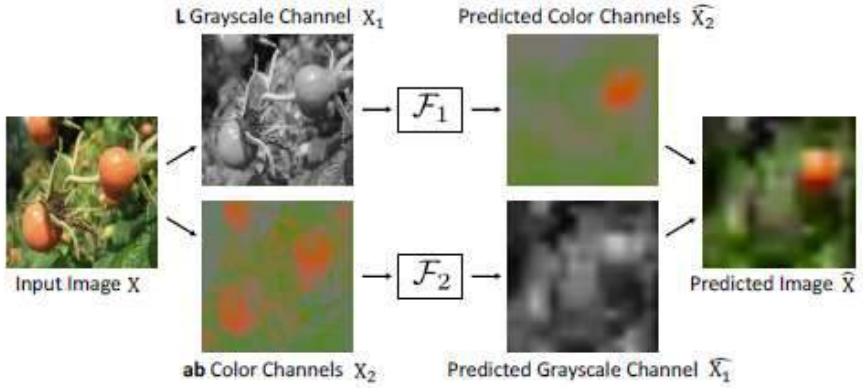




论文:Context Encoders: Feature Learning by Inpainting. In *CVPR 2016*.

表示学习在图像领域的应用





Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. *In CVPR 2017*

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick

基本概念

instance-level discrimination:每一个样本就是一个实例，自成一类

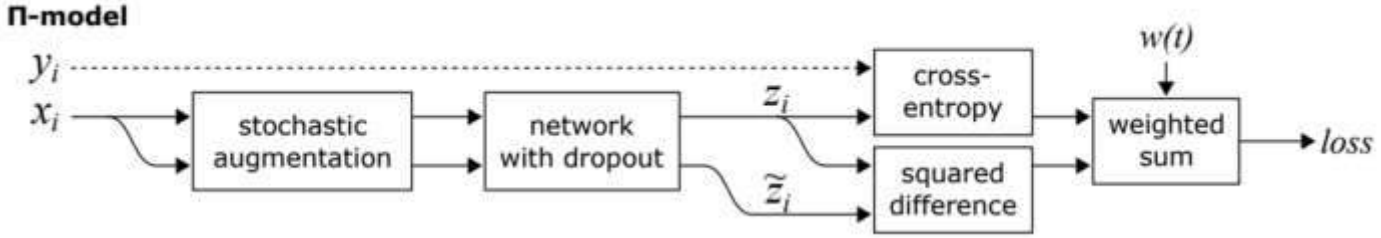
contrastive learning：即contrastive损失，这种损失可以处理孪生网络中的成对数据，这个loss真正work的关键点就是你要sample很多negative

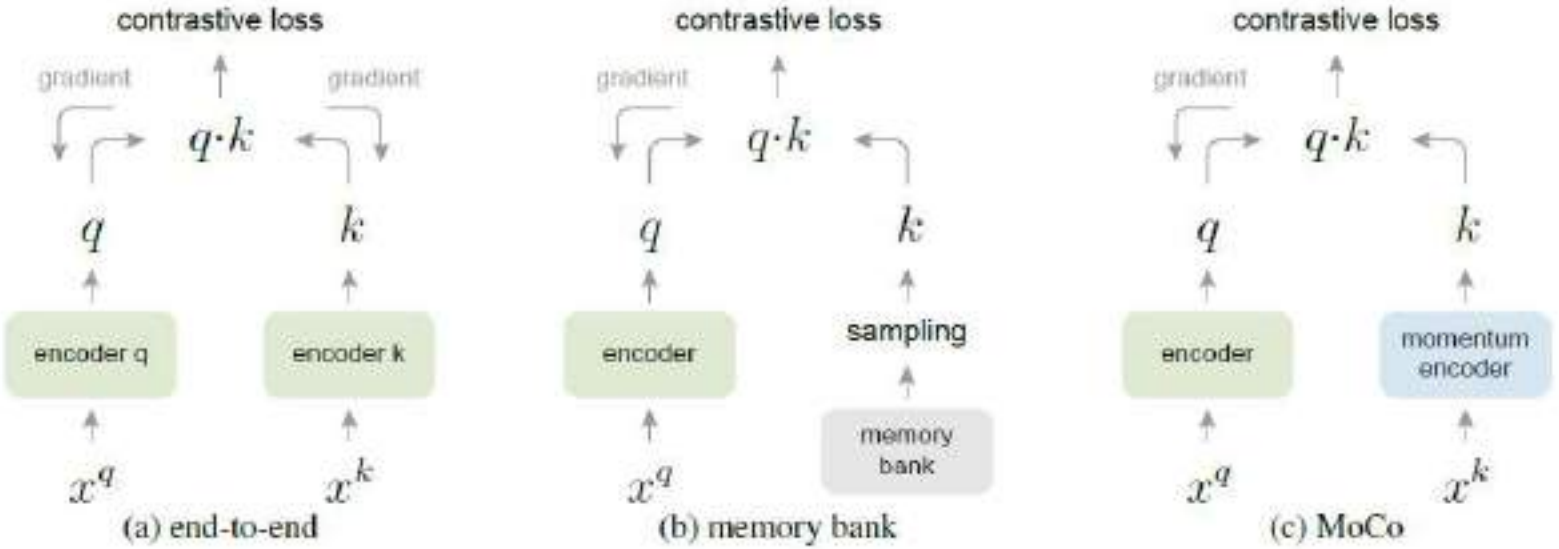
Noise-contrastive estimation：把多分类问题转化为二分类，不直接计算样本x对应每个类别的概题。

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{v}\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{w}_j^T \mathbf{v}\right)}.$$

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)},$$

Consistency Loss：Temporal Ensembling for Semi-Supervised Learning（ICLR17)




contrastive loss
gradient
gradient
$q \cdot k$
q
k
encoder q
encoder k
x^q
x^k
(a) end-to-end
contrastive loss
gradient
$q \cdot k$
q
k
encoder
sampling
memory bank
x^q
(b) memory bank
contrastive loss
gradient
$q \cdot k$
q
k
encoder
momentum encoder
x^q
x^k
(c) MoCo

三种不同的学习框架

Momentum Contrast

基本假设：好的特征可以通过一个涵盖大量负样本的词典(集合)来学习

1. 将词典作为队列：将词典保存为一个数据样本队列，可以重新利用来自之前mini-batch的编码，字典中的样本将被逐步替代。

2. 动量更新：因为编码器的快速变化会导致降低编码的一致性，直接采用反向传播会降低队列词典的作用，所以采取EMA的更新方式：

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

Algorithm 1 Pseudocode of MoCo in a PyTorch-like style.

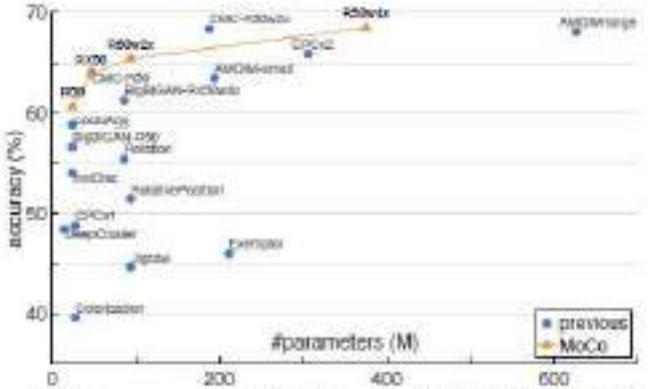
```
# f_q, f_k: encoder networks for query and key
# queue: dictionary as a queue of K keys (CxK)
# m: momentum
# t: temperature

f_k.params = f_q.params # initialize
for x in loader: # load a minibatch x with N samples
   x_q = aug(x) # a randomly augmented version
   x_k = aug(x) # another randomly augmented version

   q = f_q.forward(x_q) # queries: NxC
   k = f_k.forward(x_k) # keys: NxC
   k = k.detach() # no gradient to keys

   # positive logits: Nx1
   l_pos = bmm(q.view(N,1,C), k.view(N,C,1))

   # negative logits: NxK
   l_neg = mm(q.view(N,C), queue.view(C,K))

   # logits: Nx(1+K)
   logits = cat([l_pos, l_neg], dim=1)

   # contrastive loss, Eqn.(1)
   labels = zeros(N) # positives are the 0-th
   loss = CrossEntropyLoss(logits/t, labels)

   # SGD update: query network
   loss.backward()
   update(f_q.params)

   # momentum update: key network
   f_k.params = m*f_k.params+(1-m)*f_q.params

   # update dictionary
   enqueue(queue, k) # enqueue the current minibatch
   dequeue(queue) # dequeue the earliest minibatch
```

bmm: batch matrix multiplication; mm: matrix multiplication; cat: concatenation.

数据增广

动量更新

更新队列



method	architecture	#params (M)	accuracy (%)
Exemplar [15]	R50w3×	211	46.0 [36]
RelativePosition [11]	R50w2×	94	51.4 [36]
Jigsaw [43]	R50w2×	94	44.6 [36]
Rotation [17]	Rv50w4×	86	55.4 [36]
Colorization [62]	R101*	28	39.6 [12]
DeepCluster [3]	VGG [51]	15	48.4 [4]
BigBiGAN [14]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [59]	R50	24	54.0
LocalAgg [64]	R50	24	58.8
CPC v1 [44]	R101*	28	48.7
CPC v2 [33]	$R170^{*}_{wider}$	303	65.9
CMC [54]	$R50_{L+ab}$	47	64.1†
	$R50w2\times_{L+ab}$	188	68.4†
AMDIM [2]	$AMDIM_{small}$	194	63.5†
	$AMDIM_{large}$	626	68.1†
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

pre-train	AP_{50}	AP	AP_{75}
random init.	58.0	32.8	32.5
super. IN-1M	81.5	53.6	58.9
MoCo IN-1M	81.1 (−0.4)	53.8 (+0.2)	58.6 (−0.3)
MoCo IG-1B	81.6 (+0.1)	54.8 (+1.2)	60.3 (+1.4)

(a) Faster R-CNN, R50-**dilated**-C5

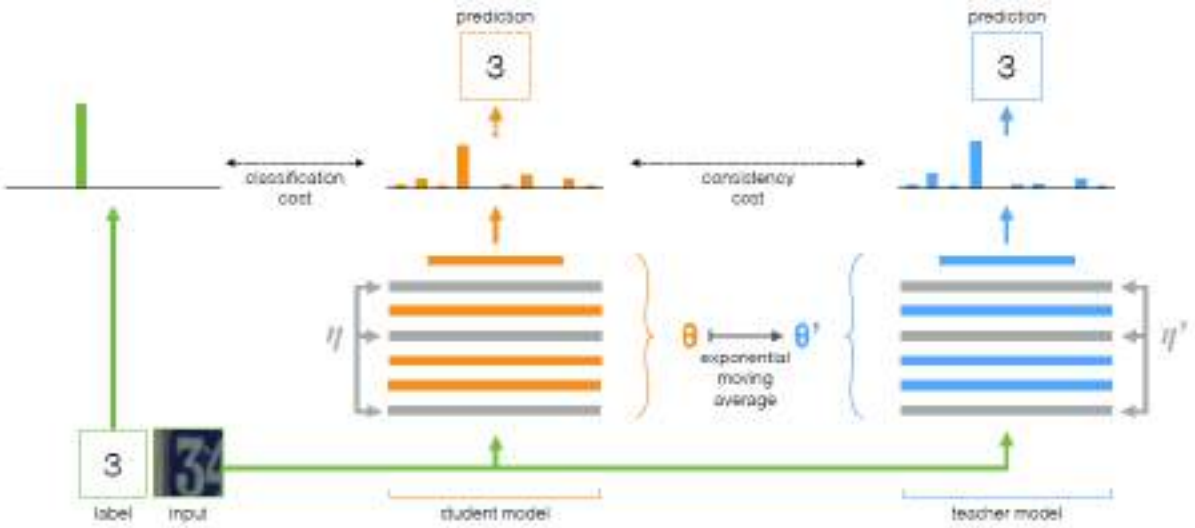
pre-train	AP_{50}	AP	AP_{75}
random init.	52.5	28.1	26.2
super. IN-1M	80.8	52.0	56.5
MoCo IN-1M	81.4 (+0.6)	55.2 (+3.2)	61.2 (+4.7)
MoCo IG-1B	82.1 (+1.3)	56.2 (+4.2)	62.3 (+5.8)

(b) Faster R-CNN, R50-C4

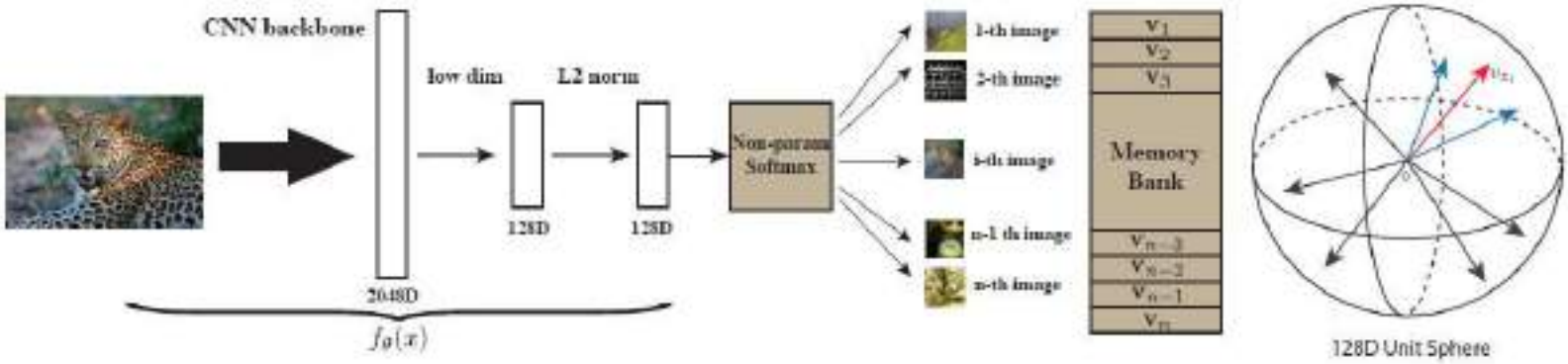
	R50-dilated-C5			R50-C4		
pre-train	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
end-to-end	77.8	50.1	53.8	79.7	53.0	57.9
memory bank	79.6	51.9	56.3	80.3	53.9	58.9
MoCo	**81.1**	**53.8**	**58.6**	**81.4**	**55.2**	**61.2**

CPCV1,CPCv2(ICLR悲剧)

CMC(ICCV悲剧)



Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results ,NIPS,2017



Unsupervised Feature Learning via Non-Parametric Instance Discrimination,CVPR,2018

总结

1.主体提取 key/query feature 并选择合适的损失函数训练，但是key 和 query 上用不同 feature extractor以及key feature extractor的更新十分重要。

2.选择key/query pair，本质上还是instance-level discrimination

3.关于BN，DeepMind之前也讨论过在无监督/半监督表示学习中的影响，并采用LN取代BN，本文作用采用了Shuffle BN来减少信息泄露

4. keypoints/densepose /detection提升显著，segmentation效果一般，可能与任务相关，但是采用更大的Ins数据集(10亿)就完全超过ImageNet pre-train

5.该论文很好总结了unsupervised learning的方法，并且在下游任务中做了详细的实验测试(有钱任性)。

https://arxiv.org/pdf/1905.09272.pdf

思考

1.NLP和CV在表征学习上有着密切的联系，由于文本之间天然的联系，可以很好构造Unsupervised learning任务，所以NLP的难点是在于如何进行Pre-train，BERT的出现基本解决了这个问题，成为类似ImageNet Pre-train的存在。

2.CV的难点在于如何进行Unsupervised learning，由于图像不如文本之间存在联系，图像的学习任务构造更加困难，目前主流的还是如self-supervised构造辅助任务，以及instance-level learning。

3.在视频领域存在比较强的信息相关，但是由于太多帧之间自带appearance相似性，从而导致视频的自监督陷入appearance相似性而根本不能学到高层次的语义相似性，所以有关视频的表示学习还是很不work的。

4. Multimodal representation learning也是比较make sense的方向，参考CMC

5.如何选取K个negative samples进行contrast learning

Local Aggregation of unsupervised learning of visual embeddings,ICCV,2019