大规模预训练模型概述

王浩帆

2021年6月9日

https://haofanwang.github.io/

概要

1. **预训练概述**
2. 语言预训练模型
3. 视觉预训练模型
4. 多模态预训练模型

什么是预训练？

通俗的例子是：武侠小说中，一个人若想成为武林高手，需要有扎实的内功基础，内功修炼好之后，再去学各种招式，就能够非常轻易的上手并发挥其最大效用。比如说金庸小说《倚天屠龙记》的主角张无忌，在偶然习得内功《九阳真经》之后，再学“乾坤大挪移”、“太极拳”、“太极剑”等招式就如鱼得水，进步神速。**小说中的“修炼内功”就可以理解为“预训练”的过程**。

https://www.zhihu.com/question/327642286/answer/1642346894

预训练诞生的背景

标注资源稀缺而无标注资源丰富: 某种特殊的任务只存在非常少量的相关训练数据，以至于模型不能从中学习总结到有用的规律

为什么要做预训练？

1. 基于大规模的**无标注**数据，模型可以隐式地学习到了**通用的表征**。
2. 从开放领域学到的知识**迁移到下游任务**，以改善低资源任务。
3. 预训练模型在几乎所有 NLP 任务中都取得了**目前最佳**的成果。
4. 预训练模型+微调机制具备很好的**可扩展性**，在支持一个新任务时，只需要利用该任务的标注数据进行微调即可。

预训练的关键技术

先在**大规模通用数据**(文本/图像)上训练，学习到**通用的表征**，然后再针对性的在**特定任务上进行迁移**训练。

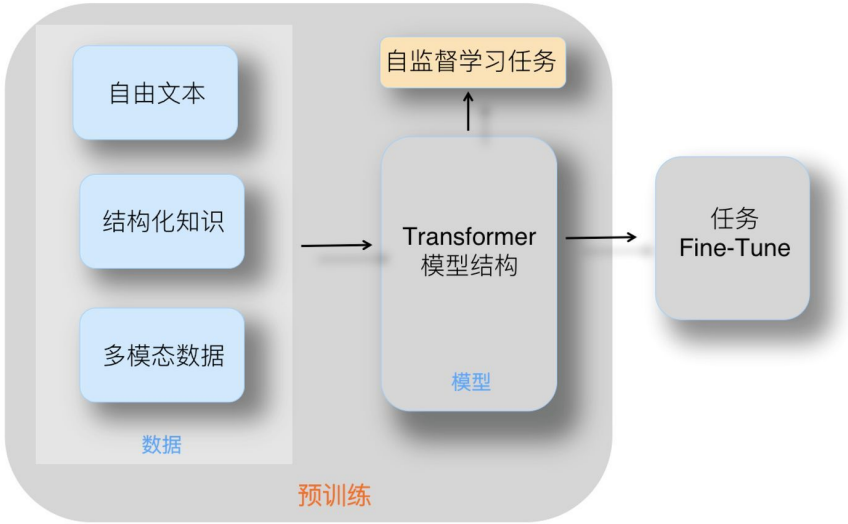
预训练的关键技术

巧妙的训练方式：自监督

高效的模型结构：变形器

先在**大规模通用数据**（文本/图像）上训练，学习到**通用的表征**，然后再针对性的在**特定任务上进行迁移**训练。

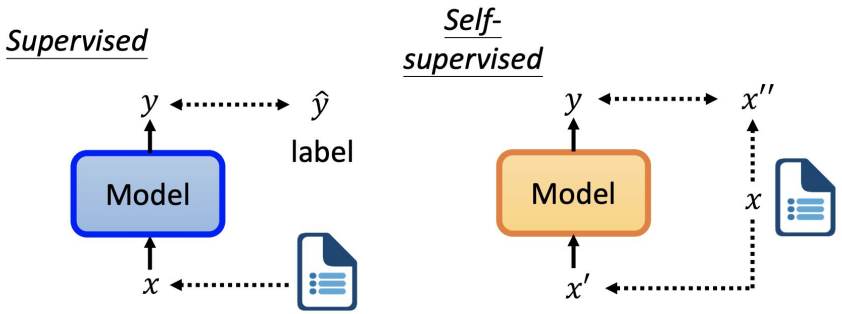
快速的知识迁移：微调

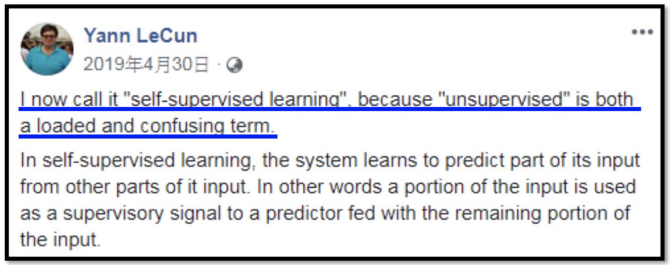


https://zhuanlan.zhihu.com/p/254821426

预训练的关键技术

1. 变形器（Transformer）
   a. Self-attention (contextual information)
2. 自监督（Self-supervised Learning, Unsupervised）
   a. Contrastive Learning（MoCo）-- CV
   b. Prediction (MLM, NSP) -- NLP
3. 微调（Fine-tuning）

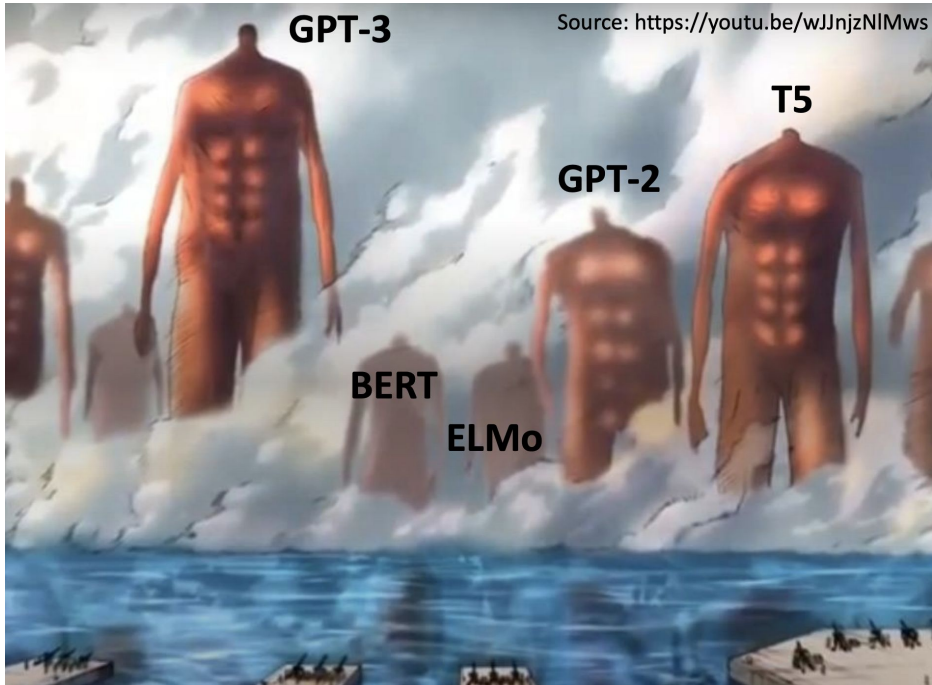




https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/bert_v8.pdf

预训练的趋势

1. 更大的模型
   a. BERT（3亿参数），GPT-3（1700亿参数），悟道2.0（1.75万亿参数）
2. 更丰富的训练策略
   a. 丰富的自监督任务
3. 单语言，多语言，多模态
   a. BERT，ViT，CLIP



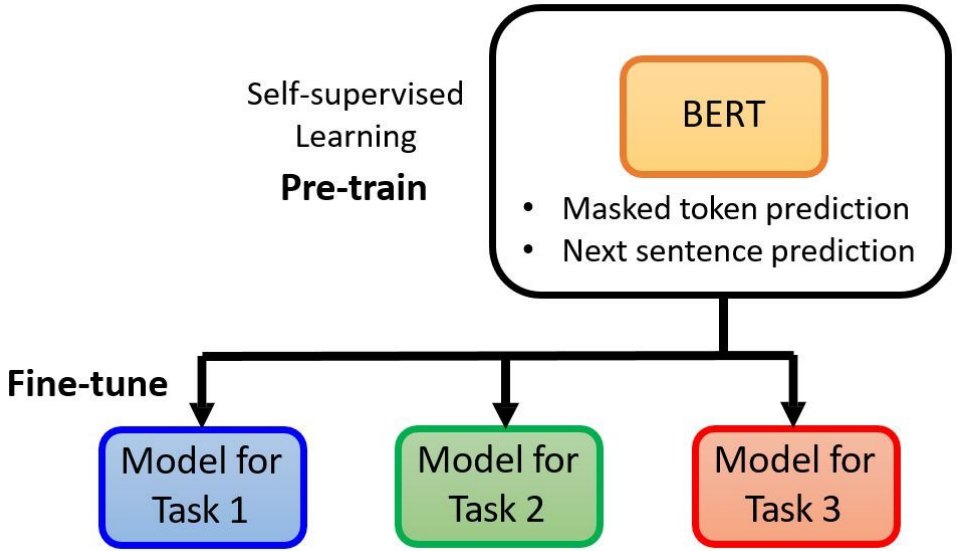
https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/bert_v8.pdf

概要

1. 预训练概述
2. **语言预训练模型**
3. 视觉预训练模型
4. 多模态预训练模型

大规模语言预训练模型

1. BERT (2018)
   a. Bidirectional Encoder Representations from Transformers
   b. 预训练任务：单词级的Mask语言模型**MLM**、句子级的下一句预测任务**NSP**



Credit to Dr. Hung-yi Lee

大规模语言预训练模型

1. BERT (2018)
   a. Bidirectional Encoder Representations from Transformers
   b. 预训练任务:单词级的Mask语言模型MLM、句子级的下一句预测任务NSP
   c. Multi lingual BERT（让模型对于不同语言做预训练任务）
      i. **也许不同语言的词汇的Embedding是很接近的**

• English: SQuAD, Chinese: DRCD

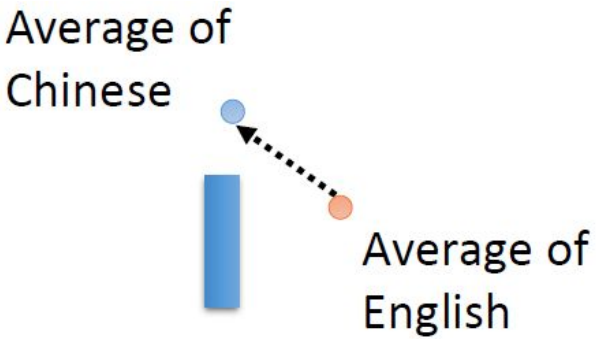
Model	Pre-train	Fine-tune	Test	EM	F1
QANet	none	Chinese	Chinese	66.1	78.1
BERT	Chinese	Chinese		82.0	89.1
	104 languages	Chinese		81.2	88.7
		English		63.3	78.8
		Chinese + English		82.6	90.1

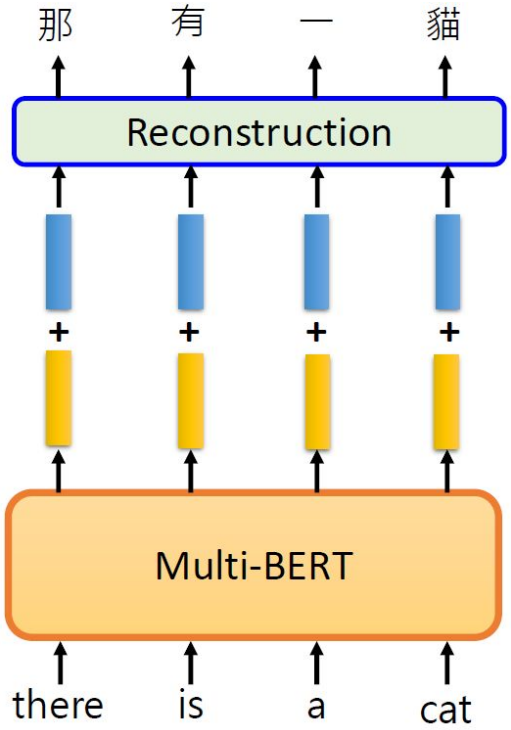
F1 score of Human performance is 93.30%

Credit to Dr. Hung-yi Lee

大规模语言预训练模型

1. BERT (2018)
   a. Bidirectional Encoder Representations from Transformers
   b. 预训练任务:MLM、NSP
   c. Multi lingual BERT（让模型对于不同语言做预训练任务）
      i. **也许不同语言的词汇的Embedding是很接近的**





Credit to Dr. Hung-yi Lee

大规模语言预训练模型

1. BERT (2018)
   a. Bidirectional Encoder Representations from Transformers
   b. 预训练任务：单词级的Mask语言模型MLM、句子级的下一句预测任务NSP
2. RoBERTa (2019)
   a. 充分训练的BERT模型
   b. 证明了NSP对于模型效果没什么影响

RoBERTa

充分训练：大BatchSize 增加训练步数

增加训练数据

Baseline

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

大规模语言预训练模型

1. BERT (2018)
   a. Bidirectional Encoder Representations from Transformers
   b. 预训练任务：单词级的Mask语言模型MLM、句子级的下一句预测任务NSP
2. RoBERTa (2019)
   a. 充分训练的BERT模型
   b. 证明了NSP对于模型效果没什么影响
3. T5 (2020)

大规模语言预训练模型

1. 更高质量、更大规模的数据
   a. 说明目前Transformer的capacity是足够的。（大数据+大模型的暴力美学）

数据量及数据质量的影响

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

数据量最大，
噪音多，有负面影响

数据质量高前提下，数据规模没那么大效果也很好

From:Google T5

大规模语言预训练模型

1. 更高质量、更大规模的数据
2. 增加模型的容量和复杂度

模型复杂度的影响

Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

模型参数放大四倍，相对Baseline效果大幅提高

From:Google T5

大规模语言预训练模型

1. 更高质量、更大规模的数据
2. 增加模型的容量和复杂度
3. 更充分的训练模型（RoBERTa）
4. 有难度的预训练（自监督）任务

各种语言模型预训练任务

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

From:Google T5

概要

1. 预训练概述
2. 语言预训练模型
3. **视觉预训练模型**
4. 多模态预训练模型

大规模视觉预训练模型

1. BiT (Big Transfer)
   a. 调整了预训练+微调中的部分操作
   b. 上游预训练阶段：
      i. 模型尺寸与数据规模的关系。单纯增加数据量或者模型容量可能会损害性能，需要同时增加二者

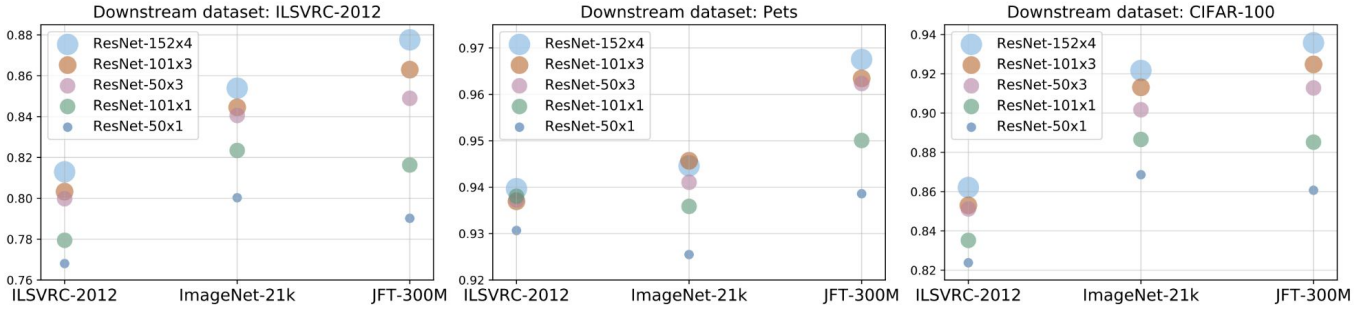


Fig. 5: Effect of upstream data (shown on the x-axis) and model size on downstream performance. Note that exclusively using more data or larger models may hurt performance; instead, both need to be increased in tandem.

Credit to Dr. Xiaohua Zhai

大规模视觉预训练模型

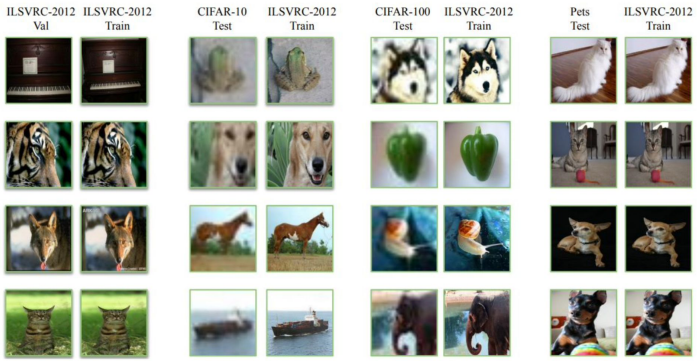
1. BiT (Big Transfer)
   a. 调整了预训练+微调中的部分操作
   b. 上游预训练阶段：
      i. 模型尺寸与数据规模的关系。单纯增加数据量或者模型容量可能会损害性能，需要同时增加二者
      ii. Batch Norm不适合迁移学习，数据域变化导致统计量发生变化。而且模型尺寸比较大的时候，在每个device上batch size无法太大。因而需要设备间通信，这样性能很差。GN会更好。
   c. 下游微调阶段：
      i. Mix-up正则化（预训练阶段由于数据充足，Mix-up效果不明显）
      ii. 分辨率影响（训练阶段使用更高分辨率）。

Credit to Dr. Xiaohua Zhai

大规模视觉预训练模型

1. BiT (Big Transfer)
   a. 调整了预训练+微调中的部分操作
   b. 上游预训练阶段：
   c. 下游微调阶段
   d. 数据泄漏
      i. 大规模预训练模型性能好会不会是因为训练阶段见到过相同图片？

Detailed analysis: Deduplication

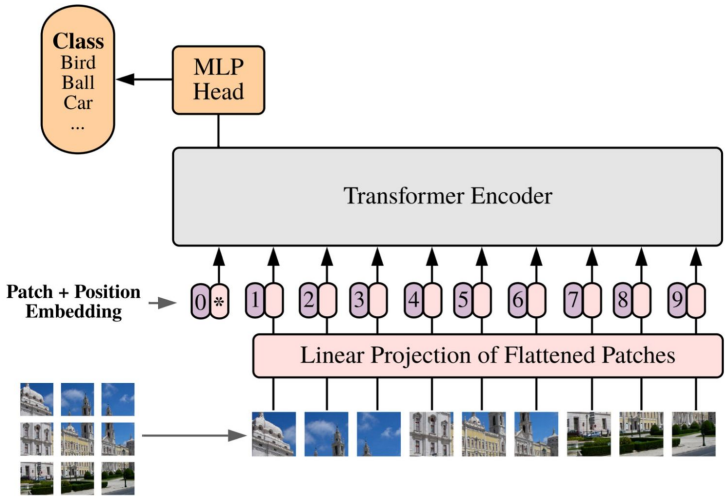


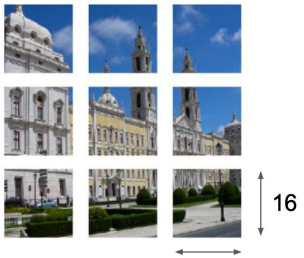
Credit to Dr. Xiaohua Zhai

大规模视觉预训练模型

1. BiT (Big Transfer, ResNet)
2. ViT (Vision Transformer)

Vision Transformer (ViT)



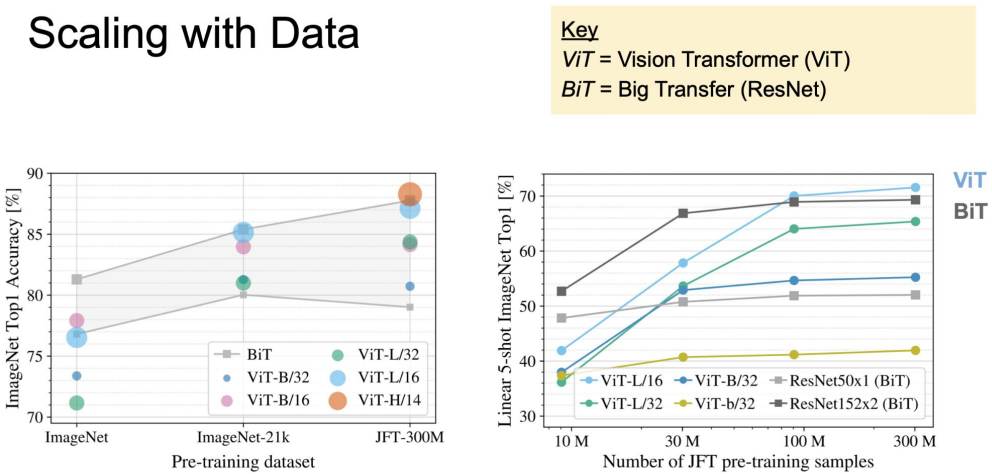


notation e.g. ViT-L/16

Credit to Dr. Xiaohua Zhai

大规模视觉预训练模型

1. BiT (Big Transfer, ResNet)
2. ViT (Vision Transformer)
   a. Transformer结构在大规模数据训练下，可以实现比ResNet更好的效果。



Conclusion: despite heavy regularization efforts ViT overfits on ImageNet, but is much better on larger datasets.

Credit to Dr. Xiaohua Zhai

概要

1. 预训练概述
2. 语言预训练模型
3. 视觉预训练模型
4. **多模态预训练模型**

大规模多模态预训练

1. 多模态预训练
   a. 本质上，多模态预训练要学习的知识是两种模态之间，或者多种模态之间的知识单元映射关系。比如对于文字-图片这两种多模态信息来说，我们可以把图片想像成一种特殊类型的语言，多模态预训练希望让模型学会这两种**不同模态之间的语义映射关系**，比如能够将单词“苹果”和图片中出现的苹果区域建立起联系。或者说，希望通过将不同模态的信息映射到相同的语义空间，来学会两者之间的语义映射关系。

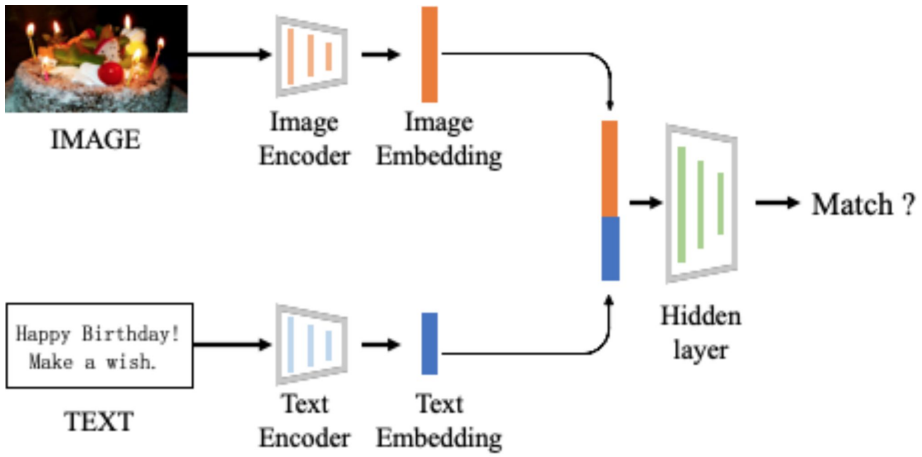
Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型

a. 在图文跨模态预训练模型中，早期的架构基本都采用基于BERT的单塔结构。

b. 视觉信息与语言**信息进行融合拼接**后，作为整体进行特征提取，如VisualBERT、UniCoder等。



Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型

a. 在图文跨模态预训练模型中，早期的架构基本都采用基于BERT的单塔结构。
b. 视觉信息与语言信息进行融合拼接后，作为整体进行特征提取，如VisualBERT、UniCoder等。
c. 单塔模型结构相似，预训练任务也大同小异。

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Tan & Bansal, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018) + BooksCorpus (Zhu et al., 2015) + English Wikipedia	1) masked language modeling 2) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

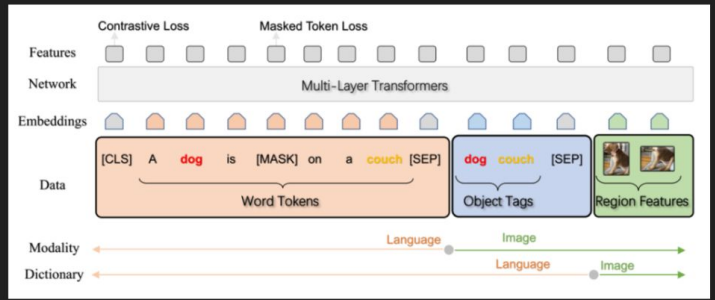
‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型

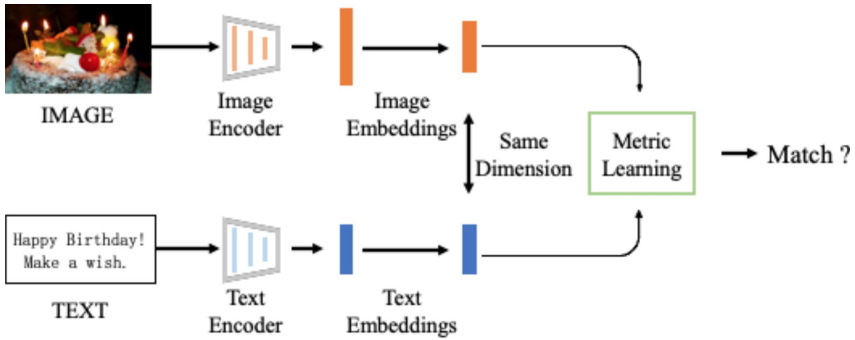
- UNITER的预训练任务：
  - 掩码语言建模（即完形填空，Masked Language Modeling）
  - 掩码区域建模（针对图片，Masked Region Modeling）
  - 图文匹配（Image-Text Matching）Triplet Loss
  - 单词和图片区域对齐（Word-Region Alignment）
- OSCAR的预训练任务：
  - Masked Language Modeling
  - Masked Region Modeling
  - Image-Text Matching Triplet Loss
  - 加入图像中检测出来区域的类别信息



Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型
2. **双塔模型**
   a. 单塔结构在模态交互上有天然优势，但是在跨模态任务上**效率过低**，不适合大规模任务，在实际落地场景中有较大局限性。
   b. 2021年来典型工作：CLIP，Align，文澜。
      i. 视觉、语言信息通过**独立分支进行编码**。
      ii. 对图文对的对齐要求降低，极大降低了数据采集的要求，可使用互 联网数据训练。
      iii. 使用对比学习的方式，缓解双塔结构不做细粒度对齐的影响。

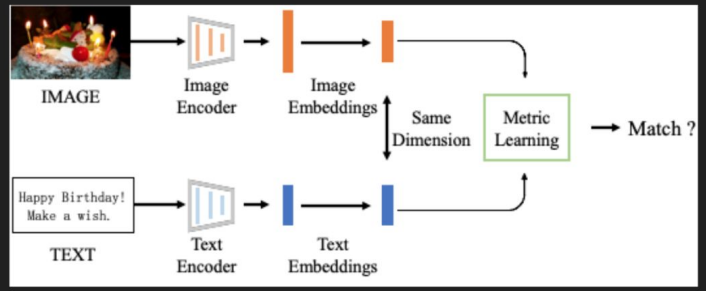


Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型
2. **双塔模型**

- 悟道-文澜：
  - 预训练数据：无清洗（仅敏感信息过滤）的中文图文对
  - 网络结构：image/text encoder + image feature map with grid pooling + image/text self-attention
  - 对比学习：MoCo + DeepSpeed加速
- OpenAI CLIP：
  - 预训练数据：经过清洗的英文图文对
  - 网络结构：image/text encoder
  - 对比学习：SimCLR
- Google ALIGN：
  - 预训练数据：无清洗的英文图文对
  - 网络结构：image/text encoder
  - 对比学习： SimCLR



Credit to Dr. Zhiwu Lu

大规模多模态预训练

1. 单塔模型
2. **双塔模型**

多模态预训练：单塔vs双塔

- 单塔模型优点：
  - 能够学到图文数据细粒度上的特征关联，对某些任务更有优势
- 单塔模型缺点：
  - 准备数据的代价大，需要每个图文对之间有很强的关联
  - 应用开销大，以图文检索为例，对每张图片/每条文本，都要与所有候选文本/图片构建图文对输入模型以获得该图文对的匹配程度
- 双塔模型优点：
  - 数据无需清洗，不需要图文对有很强的关联
  - 应用时效率高，可以提前获得候选数据的特征表示，每当有query时只需提取query特征与候选集的所有特征计算相似性
- 双塔模型缺点：
  - 只关注图文整体的匹配，在图片区域/单词等细粒度上表现可能不如单塔模型

Credit to Dr. Zhiwu Lu

资源

1. 乘风破浪的PTM：两年来预训练模型的技术进展
2. Self-Supervised Learning
3. ADL116期大规模预训练模型报告Slide合集
4. 前沿热点的论文集锦