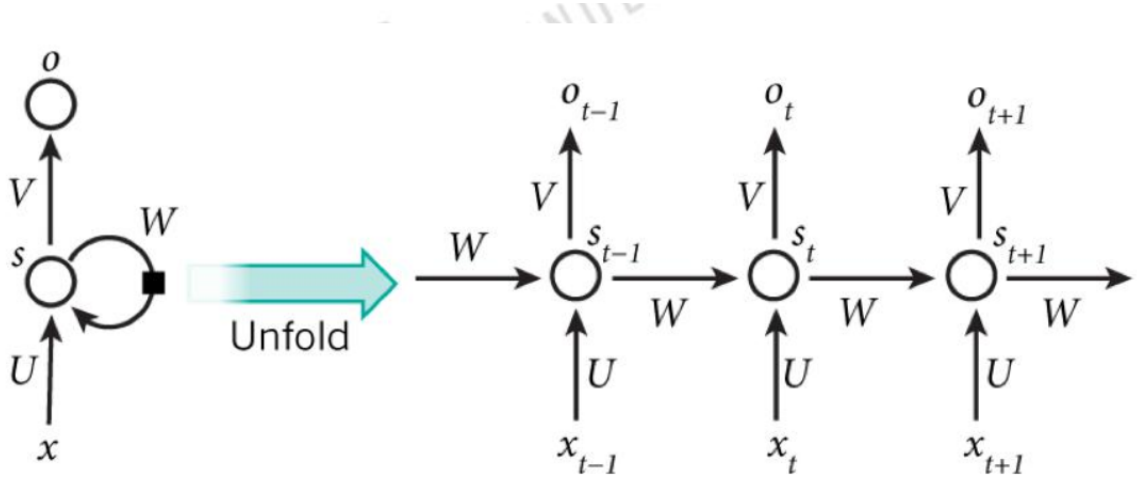
1. CNN、RNN、Transformer谁做NLP特征抽取器?

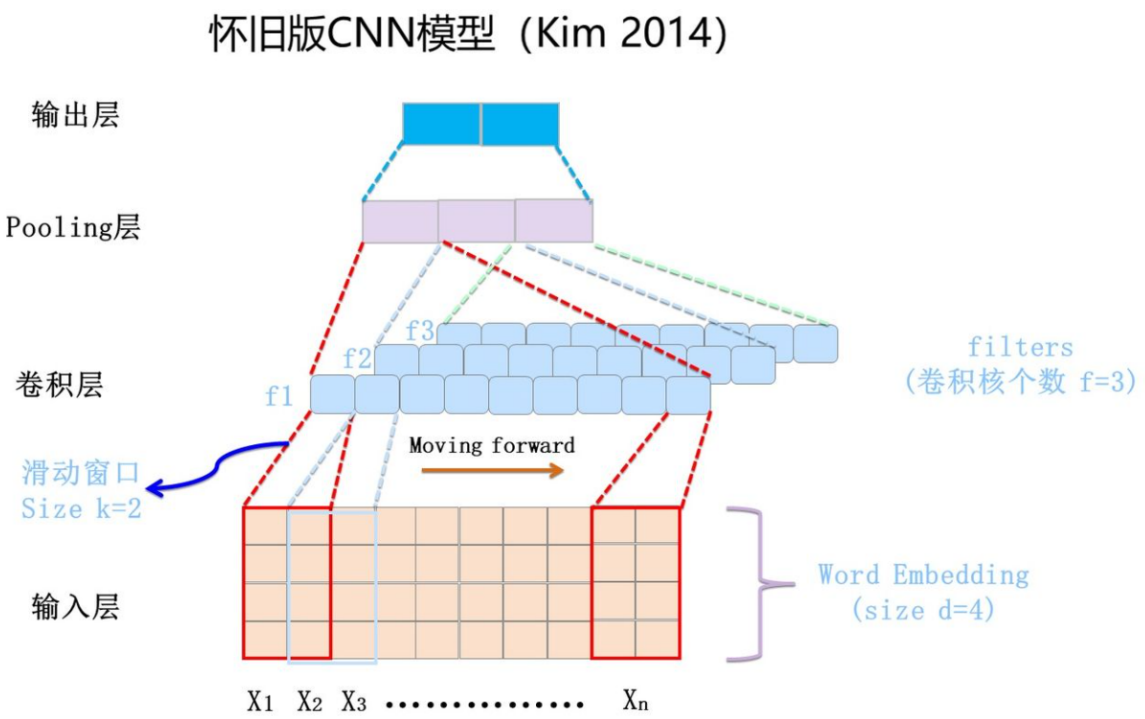
- 语义特征提取能力： transformer
- 长距离特征捕获能力：RNN \ Transformer > CNN
- 任务综合特征抽取能力： 机器翻译中Transformer
- 并行计算能力及运行效率：transformer

1.1 RNN(-2018)



> 不定长序列输入
>
> 梯度消失、梯度爆炸--> LSTM....
>
> 并行计算能力差（序列依赖问题）--> SRU、SRNN
>
> 无法获取全局信息

1.2 CNN(2014-)

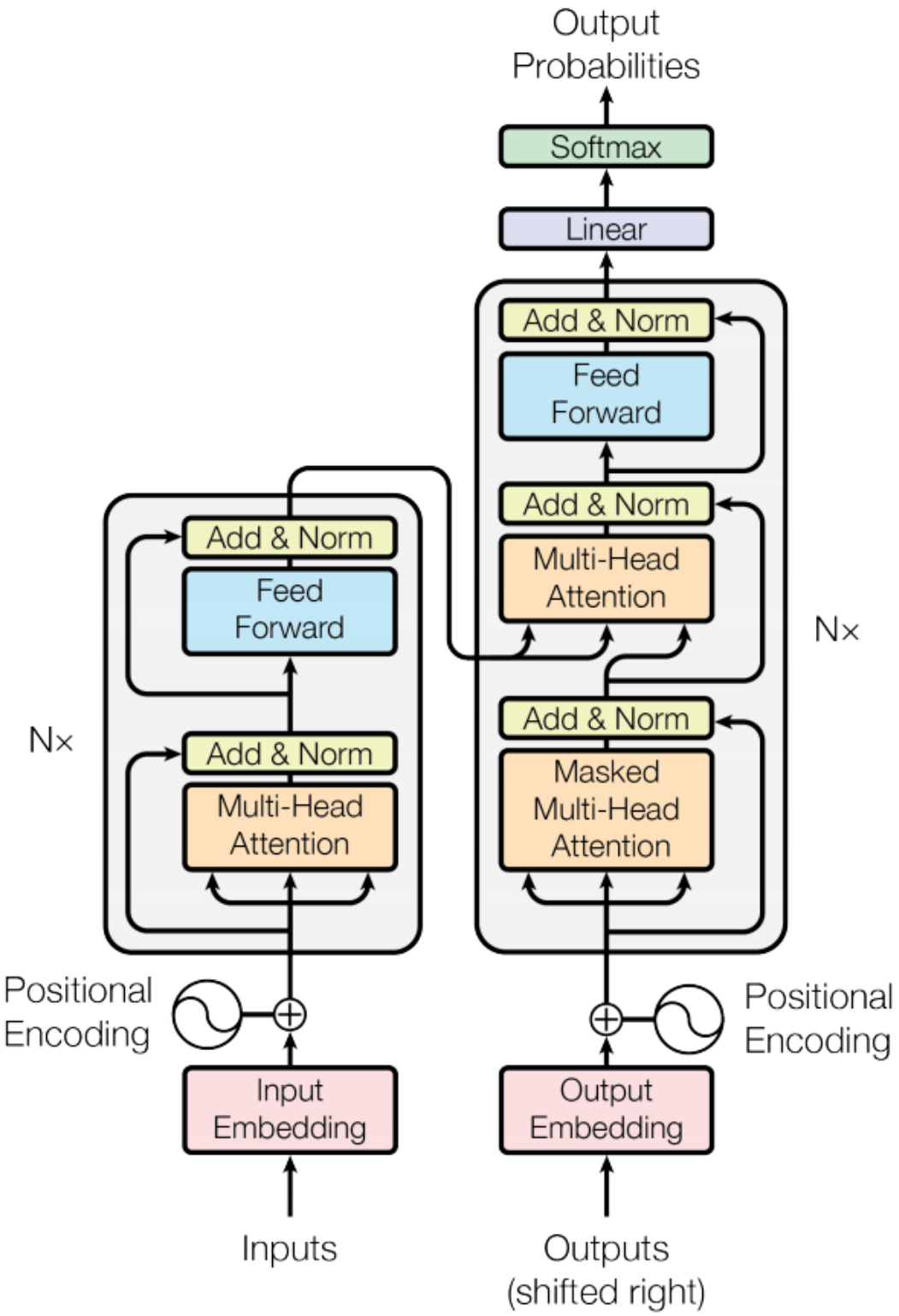


> 一维卷积 --> 位置信息的丢失 --> 抛弃pooling，使用位置编码

> 远距离特征，单层CNN无法捕获--> Dilated CNN、加深CNN网络

> 详情可参考：www.cnblogs.com/sandwichnlp/p/11876987.html

1.3 transformer(2017-)



> transformer encoder作为特征抽取器

> multi-head self-attention、skip connection、layernorm

> 不定长问题：padding填充，过长的例如文本摘要，时间复杂度过高-->transformer-xl

> 位置编码问题：相对位置编码

1.4 结合（塞）

Pay Less Attention With LightweightI and Dynamic Convolutions

> 参考：
>
> - https://zhuanlan.zhihu.com/p/54743941
> - Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures
> - how much attention do you need a granular analysis of neural machine translation architectures
> - efficient transformers：A Survey

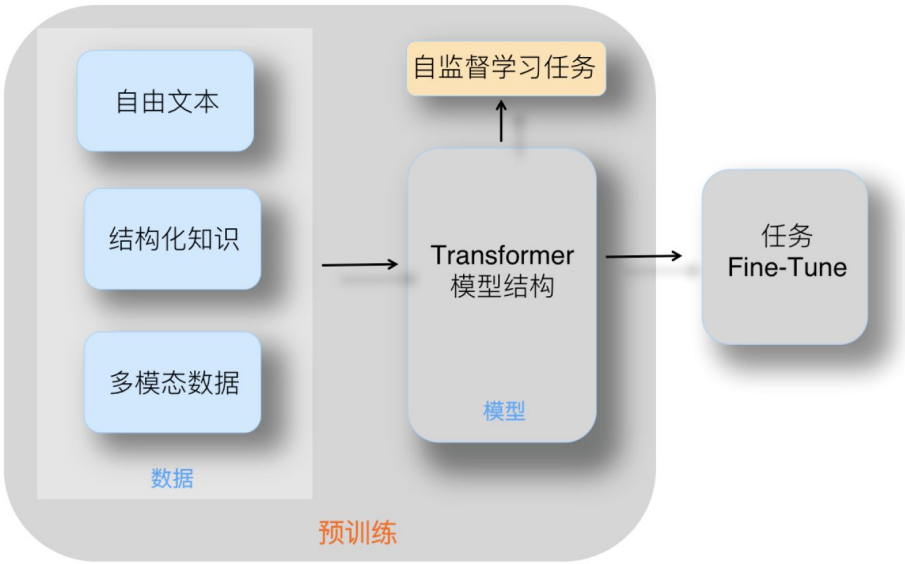
2. transformer in NLP

2.1 预训练模型回顾

2.1.1 预训练 in CV

- 训练数据小、不足以训练复杂网络
- 加快训练速度
- 参数初始化，先找到好的初始点，有利于优化
- **底层特征的可复用性+高层特征任务**相关性fine-tuning

2.1.2 预训练 in NLP



- word2vec(静态) --> 存在多义词问题
- ELMo(动态)
  - 预训练模型
  - 抽取特征：ELMo为下游提供每个单词的特征
  - 下游任务：特征集成
- bert家族

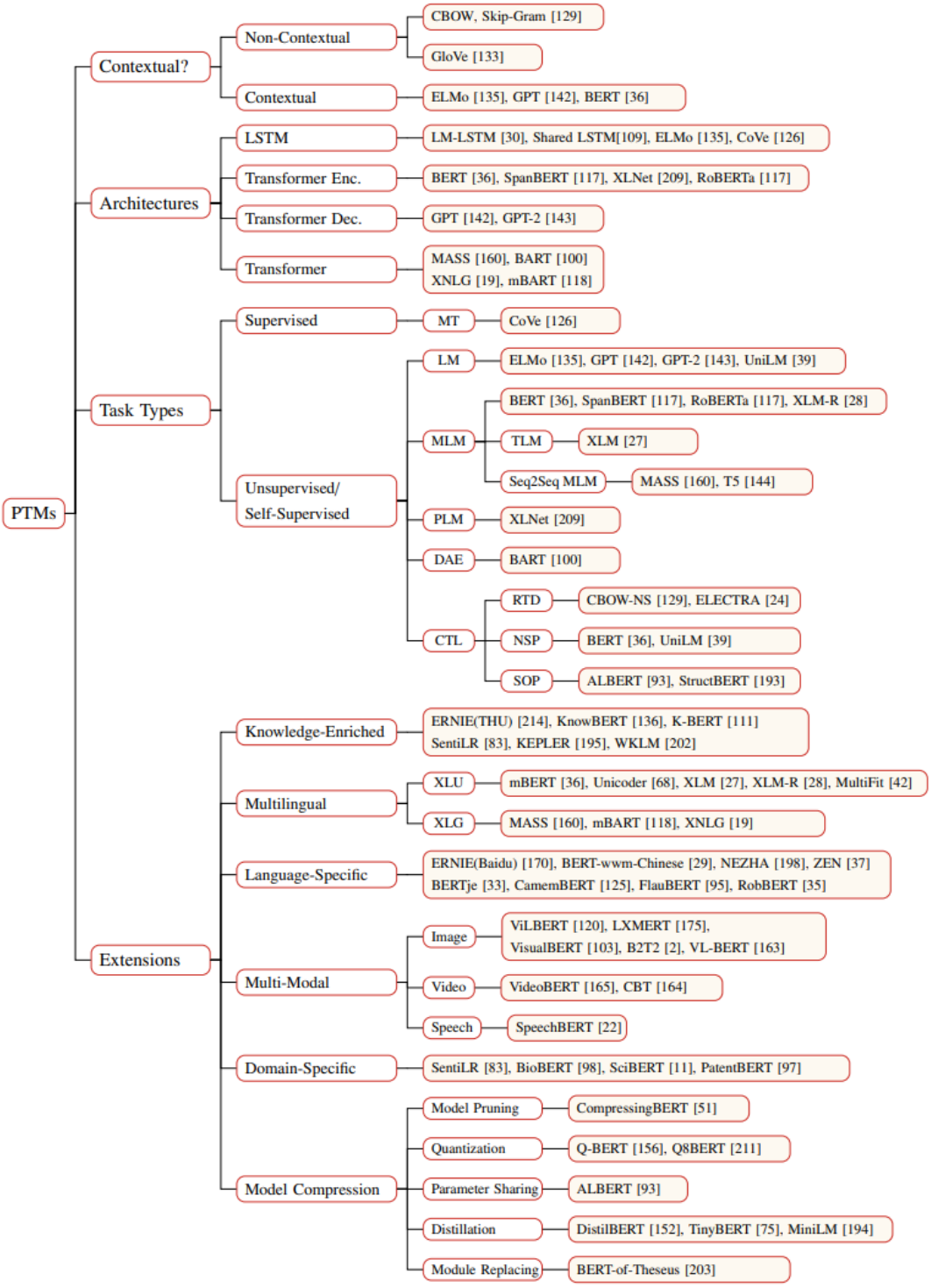
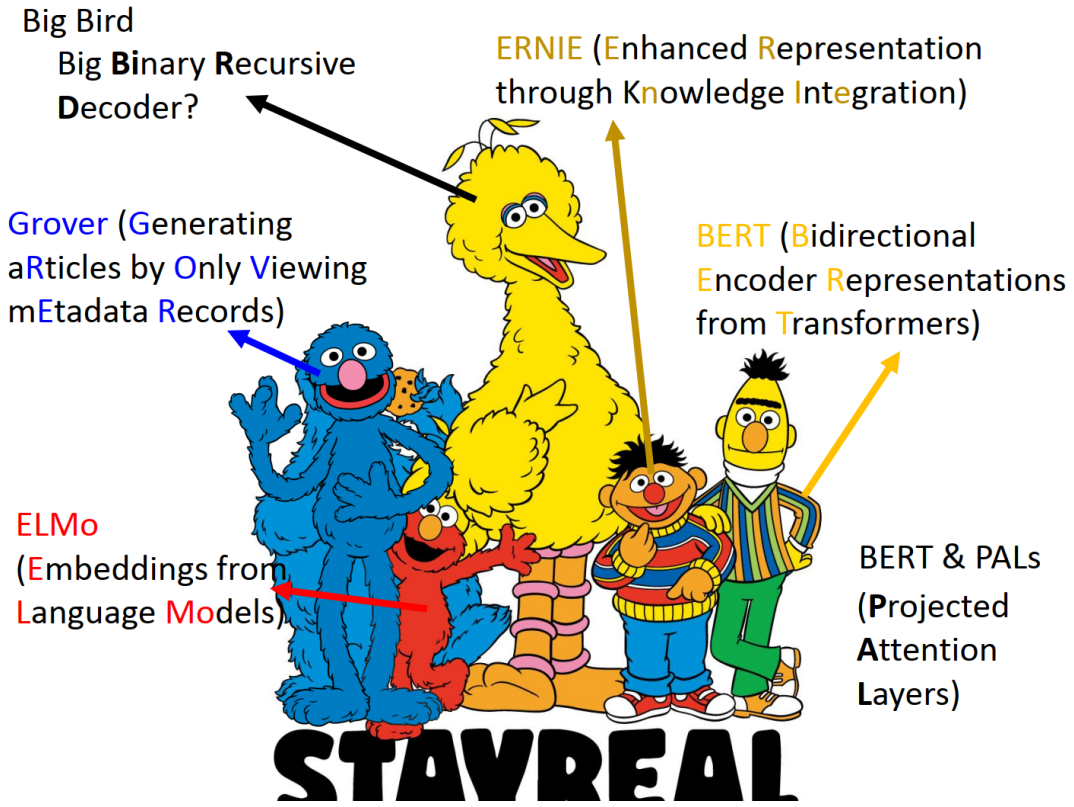


Figure 3: Taxonomy of PTMs with Representative Examples

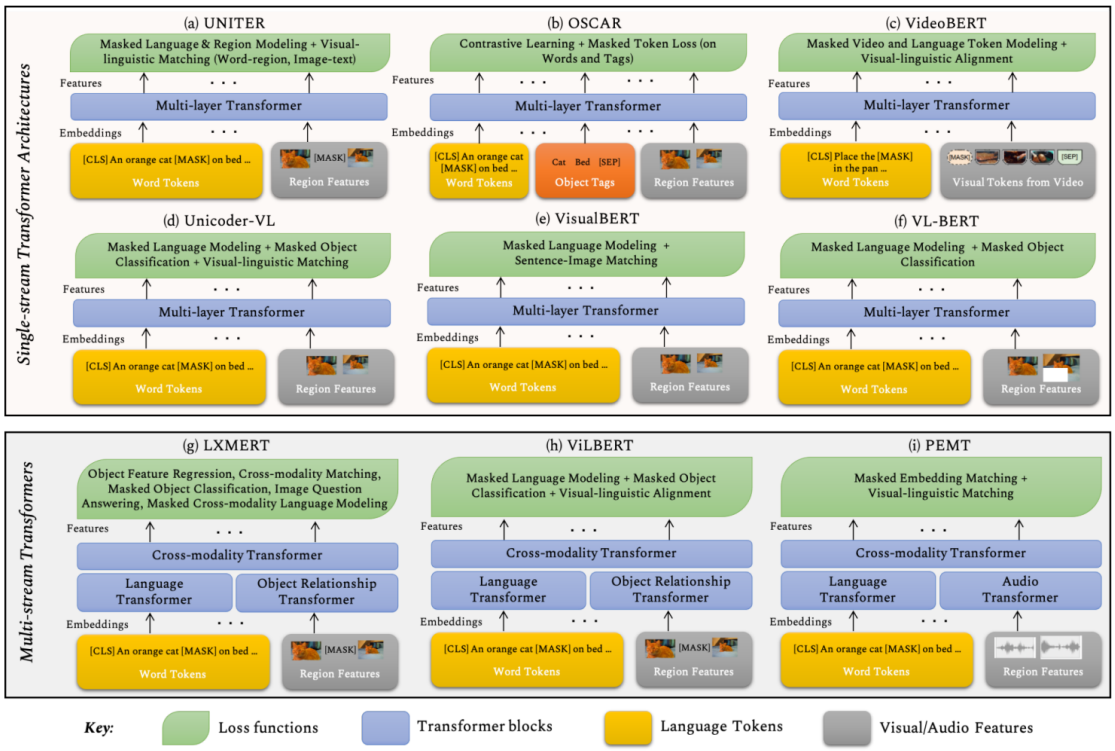


> LSTM特征抽取能力弱于 transformer
>
> 拼接方式双向融合特征，融合能力较差

> 参考：
>
> - deep contextualized word representions
> - don't stop pretraining:Adapt Language Models to Domain and tasks
> - 综述：Pre-trained Models for Natural Language Processing: A Survey

2.1.3 预训练 in 多模态

14



这里讨论text-img，基本上为VQA任务，其他详情可参考 Pre-trained Models for Natural Language Processing: A Survey

- MS-COCO
- Visual Gnome
- 双流交互：LXMERT：learning Cross-Modality Encoder Representations from transformer
  - img用faster-RCNN模型识别出图像中包含的多个物体及其对应的矩形位置信息，将高置信度的物体及其对应的位置信息作为图像侧tranformer的输入
  - transformer作为本文抽取器
  - **ViLbert**
- 单流：Unicoder-VL：A Universal Encoder for vision and language by cross-modal pre-training
  - VisualBERT
  - VL-VERT
  - UNITER
- B2T2、VLBERT、Unicoder-VL、UNITER
- 语音-文本：speechBERT

> - Pre-trained Models for Natural Language Processing: A Survey
> - M6-v0:Vision-and-Language Interaction for Mutli-modal Pretraining
> - github.com/yuewang-cuhk/awesome-vision-language-pretraining-papers

2.2 GPT

- transformer 作为特征抽取器，利用语言模型作为训练任务
- 通过fine-tuning的模型解决下游任务

> 单向+只利用了上文的信息，而没有融合下文的信息进行预测，在阅读理解等任务中不好（AR适合生成类）

> - Imporving Language Understanding by generative Pre-Training

2.3 Bert

从word2vec的CBOW到BERT：把语言模型改成双向的，学习CBOW的抠词的方法

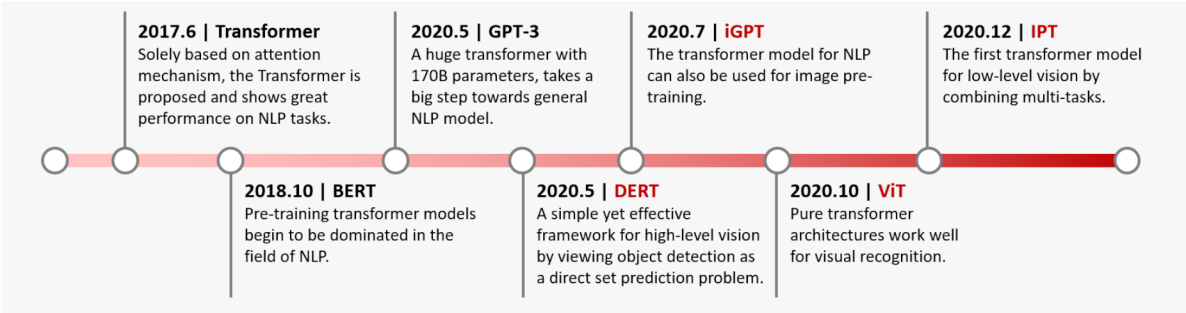
从ELMo到BERT：特征抽取器换成了transformer

从GPT到BERT：语言模型换成双向的

- 语言模型预训练
- Fine-tuning

> **RoBerta**：bert是半成品，RoBerta是完成品
>
> XLNet：基于transformer-xl（Encoder-AE+Decoder-AR,Encoder-AE适合语言理解类任务，Decoder-AR适合语言生成类任务，将二者结合会怎样?）
>
> RoBERTa,
>
> 语言理解类任务：Google T5，ALBERT
>
> 语言生成类：GPT3
>
> ELECTRA，XLNet，GPT3，BART，UNILM v2, StructBert，MacBert

3. transformer in CV



将transformer迁移到CV领域：

- 将self-attention机制与常见的CNN架构结合： transformer in transformer： **huawei TNT**
  - 论文链接：https://arxiv.org/abs/2103.00112
  - 代码链接：https://github.com/huawei-noah/noah-research/tree/master/TNT
- 用self-attention机制完全替代CNN：GAN --> **transGAN**：over GAN based in CNN

> 感兴趣可以参考：A Survey on Visual Transformer

4. transformer in speech

4.1 ASR with transformer

- 传统GMM-HMM --> DNN-HMM --> BiLSTM+attention
- **Speech-transformer** -->
- 使用预训练模型（结合transformer，**AALBert**）
- Transducer模型（Transformer融入RNN-T） -->
  - 2019 facebook **transformer-transducer**
  - 2020 google **transformer-transformer**
  - 2020 huawei **Conv-transformer**
- Joint Attention/CTC 模型
  - 时延下降，准确率也没下降多少
- **Conformer**：卷积增强，transformer善于捕获长序列依赖+conv善于对局部特征建模

> 用transformer的自注意力机制需要对全序列上下文信息进行建模，计算复杂度会随着语音时长的增加而增加 + 多层累积 --> 在流式ASR中延迟严重
>
> - Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset

> 参考：
>
> - Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss
> - 语音应用中transformer和RNN的比较：https://zhuanlan.zhihu.com/p/309390439
> - 实践：wenet：https://github.com/mobvoi/wenet/

4.2 TTS with transformer

> - Neural Speech Synthesis with transformer Network

> blog参考： https://zhihu.com/people/zhang-jun-lin-76/posts
>
> https://github.com/dk-liang/Awesome-Visual-Transformer
> https://github.com/DirtyHarryLYL/Transformer-in-Vision