模型简述

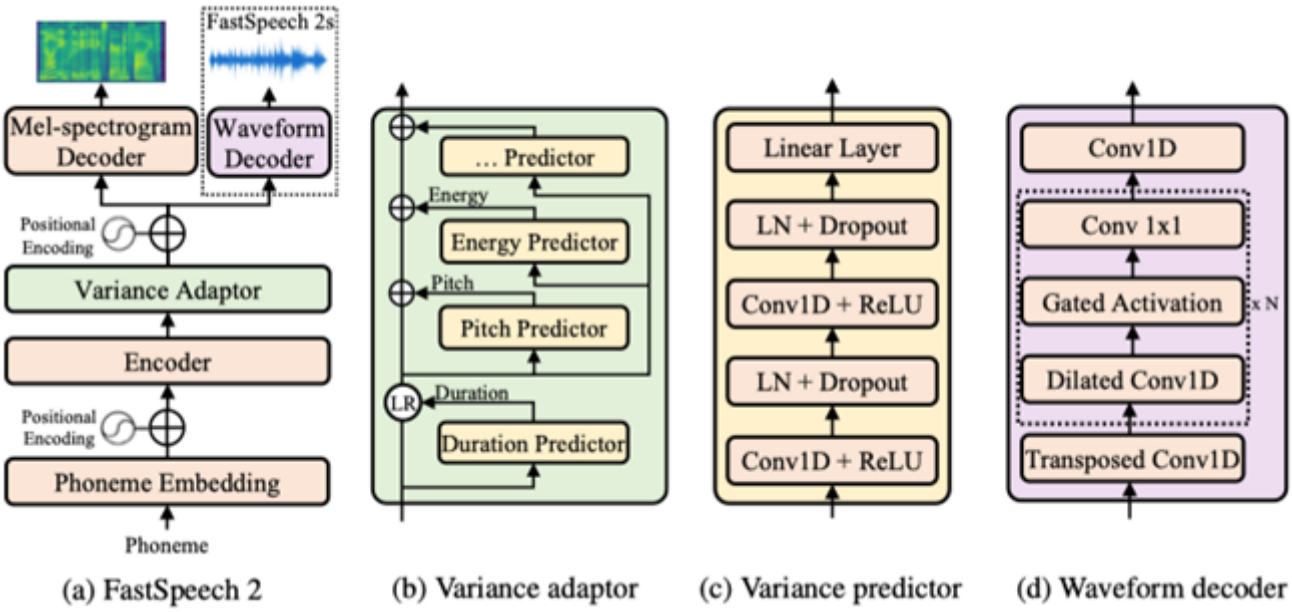
Text

FastSpeech2Conformer

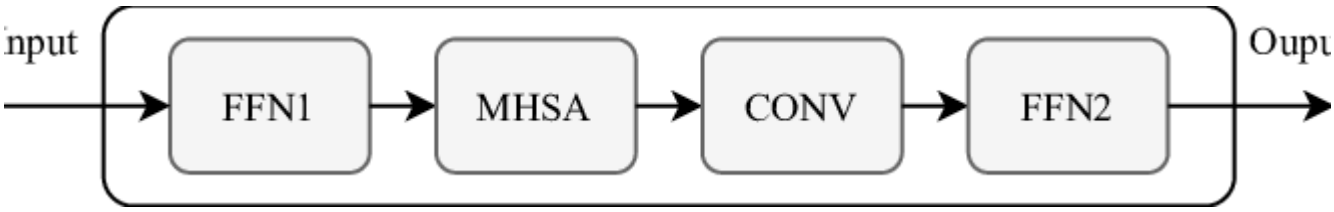
https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/fastspeech2_conformer

语音合成模型。Conformer是Convolution-augmented Transformer

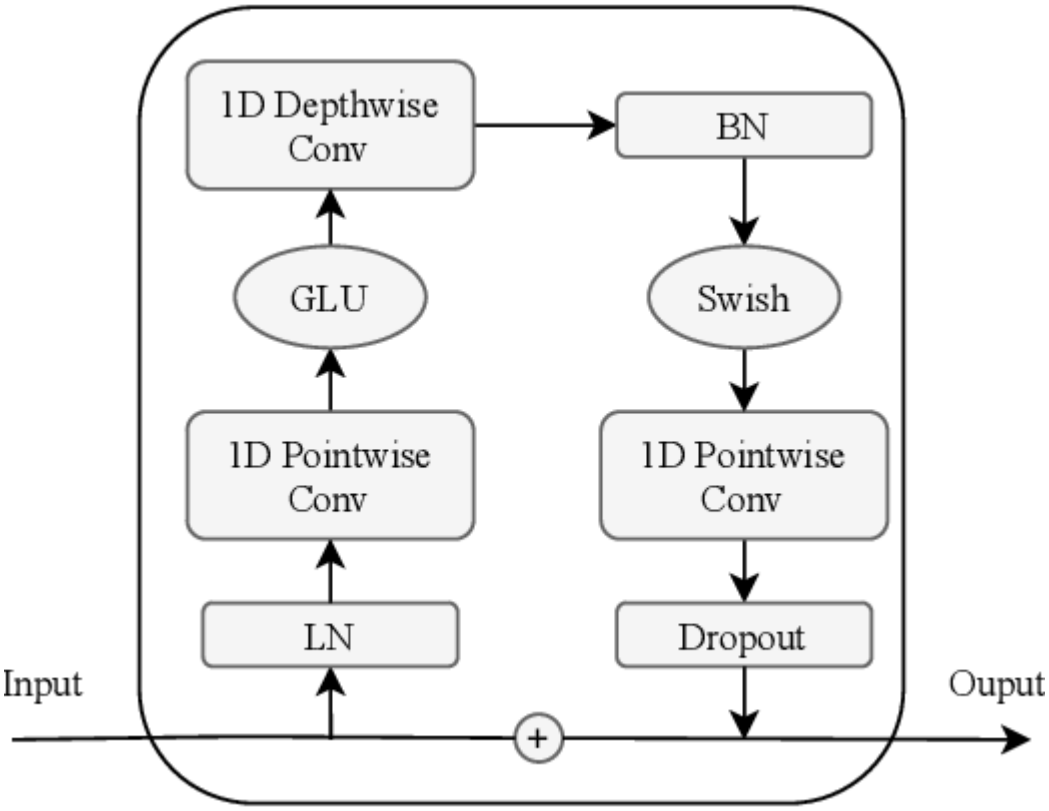
FastSpeech2 Model Architecture



Conformer Blocks



Convolution Module



Funnel Transformer

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/funnel
是一个双向Transformer，类似于BERT，但在每个层块之后有一个池化操作，这有点类似于传统CNN。
解决的问题：以低成本利用大量未标记数据

LED

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/led
LED结合了局部窗口注意力和由任务驱动的全局注意力，使处理数千个令牌或更长的文档变得容易
解决的问题：Transformer的模型由于其自注意力操作的计算复杂度与序列长度成二次方比例增长，无法处理长序列

Marian

没看懂 是一个高效且自成一体的神经机器翻译框架，配备了基于动态计算图的集成自动微分引擎。Marian完全用C++编写。

SwitchTransformer

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/switch_transformers
Switch Transformer 模型使用了稀疏的 T5 编码器-解码器架构，其中**MLP 被专家混合体**

（**MoE**） 所替换。一个路由机制（在本例中为 top-1 experts）将每个令牌与一个专家关联，每个专家都是一个密集的 MLP。

I-BERT

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/ibert
是RoBERTa的量化版本，该方案将整个推理过程**仅使用整数运算**进行量化。基于轻量级的仅整数近似方法，用于非线性操作如GELU、Softmax和层归一化，I-BERT实现了端到端的仅整数BERT推理，无需任何浮点计算。
解决的问题：减少了模型对浮点计算的依赖，还减少了模型的内存占用和功耗

QDQBert

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/qdqbert
整数量化： QDQBERT 采用整数量化技术，将模型参数从浮点数转换为整数，这样做可以显著减少模型的存储需求并提高计算效率。
高通量整数指令的利用： 提出了一个能够在所有研究的网络上维持与浮点基线相比 1% 内精度的 8 位量化工作流。

CANINE

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/canine
首批不使用明确的分词步骤（如字节对编码（BPE）、WordPiece 或 SentencePiece）来训练 Transformer 的论文之一。该模型直接在 Unicode 字符级别进行训练。在字符级别训练不可避免地会带来更长的序列长度，CANINE 通过一种高效的降采样策略解决了这一问题，然后应用深层 Transformer 编码器。

GPT-J

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/gptj
它是一个类似于 GPT-2 的因果语言模型，训练于 Pile 数据集
用于：**文本生成，语言理解**

XLM-ROBERTa-XL

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/xlm-roberta-xl
该模型旨在通过更大规模的多语言掩码语言模型来提升多语言理解的效果。
展示了两个更大的多语言掩码语言模型的结果，分别拥有 3.5B 和 10.7B 参数
用于：**跨语言理解 ， 多语言分类、问答和其他 NLP 任务，自然语言处理的基础研究**

Vision Model：

Mask2Former

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/mask2former
Mask2Former是一个统一的框架，用于全景、实例和语义分割。
其关键组件包括Masked-attention，该技术通过限制预测的掩蔽区域内的交叉注意力来提取局部特征。

SegFormer

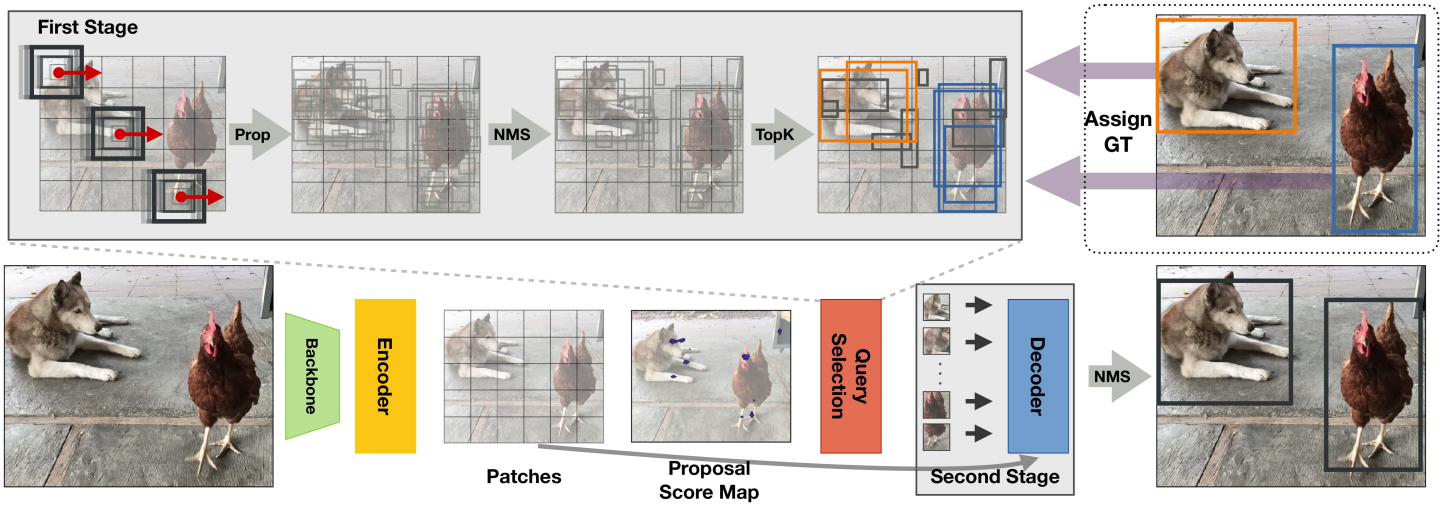
https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/segformer
该模型包括一个分层的 Transformer 编码器和一个轻量级全 MLP 解码头，能在图像分割基准测试如 ADE20K 和 Cityscapes 上取得出色的结果。
分层结构的 Transformer 编码器，输出多尺度特征。它不需要位置编码，从而避免了位置码的插值，这在测试分辨率与训练时不同时会导致性能下降。
MLP 解码器聚合了不同层的信息，因此结合了局部注意力和全局注意力以呈现强大的表示。

DETR

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/detr
DETR 包括一个卷积背景网络，后接一个可以端到端训练的编码器-解码器变换器，用于对象检测。此外，通过在解码器输出上简单添加一个掩码头，DETR 还可以自然扩展到执行全景分割任务。
用于：**对象检测 全景分割**

DETA

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/deta



我们的检测器训练 Deformable-DETR，采用传统的基于 IoU 的标签分配，在 12 个周期内（1x 计划）达到了 50.2 COCO mAP，使用 ResNet50 作为主干网络，在此设置下超越了所有现有的传统或基于变换器的检测器。

1. **从一对一匹配到一对多分配**：DETA 模型通过将 DETR 模型中的一对一匹配机制替换为传统检测器中的一对多标签分配，这种方法更加灵活，能够更好地处理多标签场景。
2. **结合非极大抑制（NMS）**：在一对多分配中加入 NMS，帮助模型在保持高检测精度的同时减少重复检测，提高检测效率。
   用于：**对象检测**

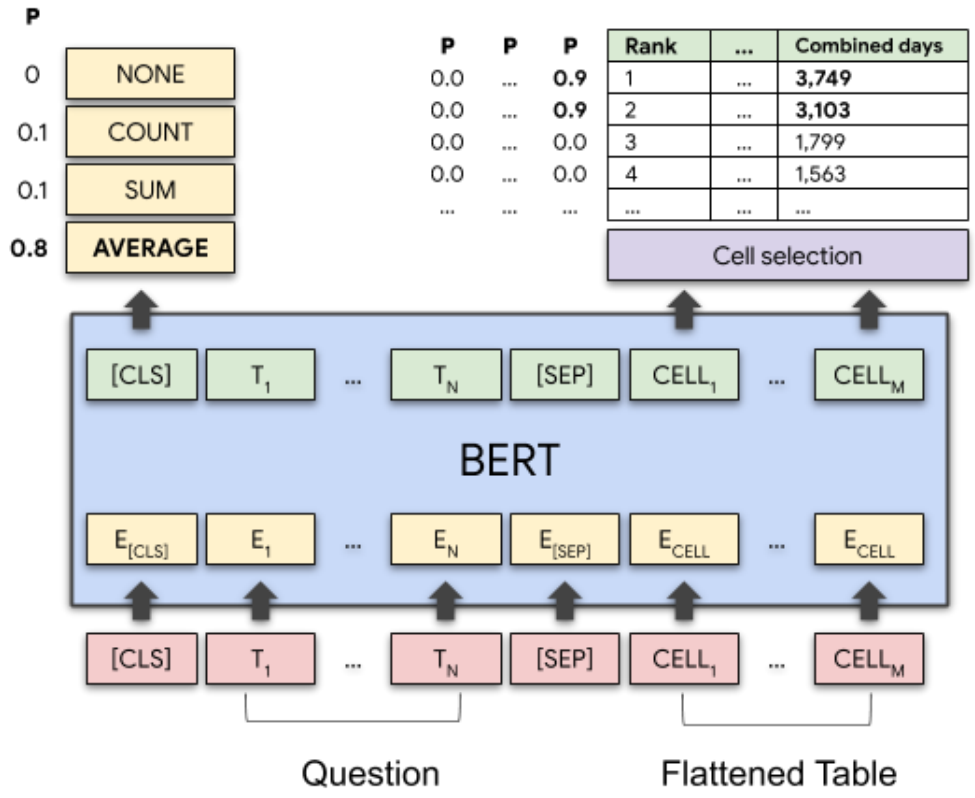
NAT Neighborhood Attention Transformer

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/nat
它是基于邻域注意力的分层视觉变换器，邻域注意力是一种滑动窗口自注意力模式。
提出邻域注意力（NA），这是第一个高效且可扩展的视觉领域中的滑动窗口注意力机制。NA是一种像素级操作，将自注意力（SA）局限于最近邻像素，因此与SA的二次复杂度相比，具有线性的时间和空间复杂度。滑动窗口模式允许NA的感受野在不需要额外像素移动的情况下增长，并保持平移等变性，这与Swin Transformer的窗口自注意力（WSA）不同。
我们开发了NATTEN（邻域注意力扩展），一个包含高效C++和CUDA内核的Python包，使NA的运行速度比Swin的WSA快40%，同时内存使用减少了25%。
我们进一步提出了基于NA的新型分层变换器设计——邻域注意力变换器（NAT），该设计提升了图像分类和下游视觉性能。
用于：**图像分类**，**下游视觉任务**

TAPAS

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/tapas
这是一个基于BERT的模型，专为回答关于表格数据的问题而设计（并进行预训练）。
与 BERT 相比，TAPAS 使用**相对位置嵌入**，并具有编码表格结构的7种令牌类型。
TAPAS 在一个包含来自英语维基百科的数百万表格及相应文本的大数据集上进行了掩码语言模型（MLM）目标的预训练。

对于问答任务，TAPAS 顶部有两个头：一个单元选择头和一个聚合头，用于（可选地）在选定单元之间执行聚合（如计数或求和）。采用**弱监督学习方法**，通过预测表格中的单元和可能的聚合操作来解析和回答问题，而不是依赖于完整的逻辑形式。



用于：

- **表格问答**：TAPAS主要用于解决自然语言问题与表格数据间的交互，适用于多种表格问答任务，包括序列问答、表格事实验证等。
- **表格蕴含**：在表格数据的上下文中，验证一条声明是否由表格支持，属于二分类任务。

VideoMAE

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/videomae
VideoMAE 将掩码自编码器（MAE）技术扩展到视频领域，提出了定制的视频管掩蔽和重建。这些简单的设计结果有效地克服了由视频重建过程中的时间相关性导致的信息泄露问题
用于：

- **视频分类**：通过学习视频中的视觉内容和动态信息，VideoMAE 能够对视频进行分类，识别视频中的场景或活动。
- **自监督学习**：作为一种自监督学习方法，VideoMAE 可以在没有标注数据的情况下进行预训练，从而学习视频的有效表示。这对于其他下游视频处理任务也是有益的。

Audio Model

Wav2Vec2-BERT

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/wav2vec2-bert
此模型在超过 143 种语言的 450 万小时未标记音频数据上预训练，并需要微调才能用于下游任务，如自动语音识别（ASR）或音频分类。
用于：

- **自动语音识别（ASR）**：转录语音为文字。
- **音频分类**：对音频进行分类或标记。
- **实时多语种翻译**：同时翻译多种源语言到目标语言，适用于实时通信。

WavLM

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/wavlm
WavLM 基于 HuBERT 框架构建，强调口头内容建模和说话人身份保留。
门控相对位置偏差：为 Transformer 结构引入了门控相对位置偏差，这有助于改善模型在语音识别任务中的表现，特别是在处理长距离依赖关系时。
话语混合训练策略：通过创建重叠话语并在训练中使用它们，这种策略增加了模型处理复杂语音场景（如多人对话）的能力，提高了说话人区分能力。
用于：全栈下游语音任务

SpeechT5

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/speecht5
该框架探索了用于自监督语音/文本表示学习的编解码器预训练。SpeechT5框架包括一个共享的编解码网络和六个模态特定（语音/文本）的预处理/后处理网络。通过预处理网络对输入的语音/文本进行预处理后，共享的编解码网络对序列到序列的转换进行建模，然后后处理网络根据解码器的输出在语音/文本模态中生成输出。利用大规模未标记的语音和文本数据，我们对SpeechT5进行预训练，学习统一模态的表示，希望提高对语音和文本的建模能力。为了将文本和语音信息对齐到这一统一的语义空间，我们提出了一种跨模态向量量化方法，该方法使用潜在单元作为编码器和解码器之间的接口，随机混合语音/文本状态。广泛的评估显示，所提出的SpeechT5框架在各种口语处理任务上表现出色，包括**自动语音识别、语音合成、语音翻译、声音转换、语音增强和说话人识别**。

MultiModel：

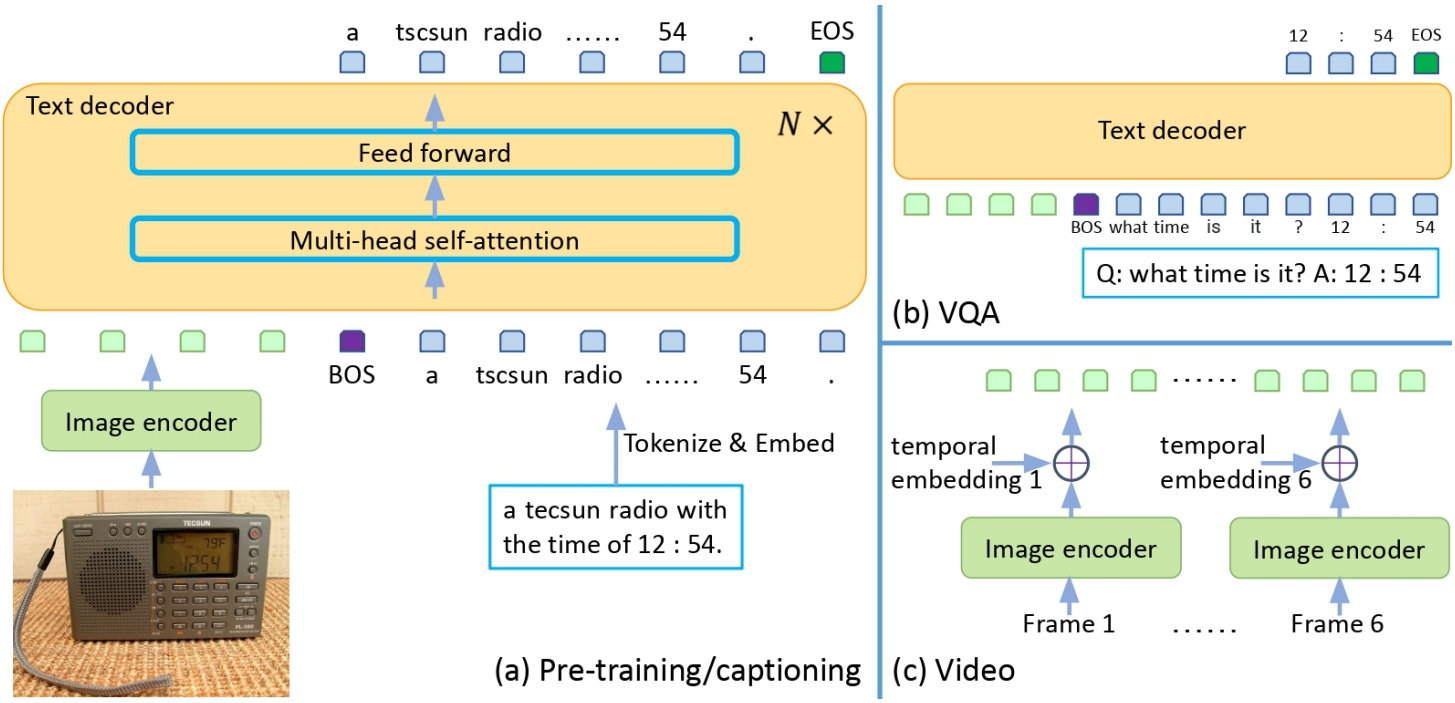
FLAVA

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/flava
一种基础的语言和视觉对齐模型。该论文旨在创建一个单一统一的基础模型，该模型可以跨视觉、语言以及视觉-语言多模态任务工作。

GIT

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/git
GIT 是一个仅包含解码器的变换器模型，它利用 CLIP 的视觉编码器来处理视觉输入以及文本，以统一视觉-语言任务，如图像/视频描述和问答。

在 GIT 中，我们将架构简化为一个图像编码器和一个文本解码器，都在单一的语言建模任务下。我们还扩大了预训练数据和模型大小以提升模型性能。不借助任何特殊技巧，我们的 GIT 在 12 个具有挑战性的基准测试中建立了新的最高水平，并且优势巨大。
用于：图像描述和视觉问答



Speech EncoderDecoder

Vision EncoderDecoder

VisionTextDualEncoder

OneFormer

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/oneformer
一个**通用的图像分割**框架，能够在单一的全景数据集上训练，以执行**语义、实例和全景分割**任务。
提出了一种任务条件联合训练策略，使得在单一多任务训练过程中可以对每个领域（语义、实例和全景分割）的真实情况进行训练；OneFormer 使用任务令牌来调节模型关注的任务，使得架构在训练时是task导向的，在推理时是task动态的；训练期间提出使用查询文本对比损失，以建立更好的任务间和类别间区分。
解决的问题：只训练一次，并在所有三个图像分割任务上实现SOTA（最先进）性能

Time Series

Time SeriesTransformer

https://huggingface.co/docs/transformers/v4.37.0/en/model_doc/time_series_transformer
a vanilla encoder-decoder Transformer for time series forecasting

Others

Wav2Vec2-Conformer
not found