AI 学习与算法集成平台技术报告

泛能源大数据与战略研究中心

目 录

第 1 章 开源机器学习平台概述 5

§1.1 平台介绍 5

§1.2 功能介绍 5

§1.2.1 开发平台 5

§1.2.2 开放能力 5

§1.2.3 知识学习 6

第 2 章 开发平台 7

§2.1 Tensorflow 7

§2.2 PyTorch 7

§2.3 Sklearn 7

§2.4 其他经典开源模型 7

第 3 章 开放能力 9

§3.1 回归预测 9

§3.1.1 海浪高度预测 9

§3.1.2 波士顿房价预测 11

§3.1.3 碳排放预测 14

§3.1.4 股票预测 14

§3.2 智能图像 14

§3.2.1 MNIST 手写体分类 14

§3.2.2 图片文字识别 15

§3.2.3 目标检测 16

§3.2.4 智能图像分割 18

§3.2.5 人体关键点检测 19

§3.2.6 图像生成 20

§3.3 自然语言处理 ····· 22
§3.3.1 文本情感分类 ····· 22
§3.3.2 机器翻译 ····· 23
§3.3.3 阅读理解 ····· 25
§3.3.4 语义匹配 ····· 25
§3.3.5 文本生成 ····· 25
§3.3.6 对话系统 ····· 26
§3.3.7 句法分析 ····· 27
§3.4 智能语音 ····· 29
§3.4.1 语音识别 ····· 29
§3.4.2 语音生成 ····· 30
§3.A 各模块模型服务接口文档 ····· 31
§3.A.1 回归预测 ····· 31

第 1 章

开源机器学习平台概述

§1.1 平台介绍

随着人工智能（artificial intelligent，AI）技术的不断发展，各种 AI 产品已经逐步进入了我们的生活。但是局限专业知识，非 AI 开发人士无法快速、简单地了解 AI 知识、使用 AI 产品。AI 平台以深度学习技术研究和业务应用为基础，集深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台于一体，是全面开源开放、技术领先、功能完备的产业级深度学习平台。AI 平台源于产业实践，始终致力于与产业深入融合。

§1.2 功能介绍

本平台主要包含**开发平台**、**开放能力**、**知识学习**三个模块。其中开发平台集成了主流深度学习工具的开发接口，开放能力集成了多领域的端到端的模型训练和预测服务，知识学习模块包含了当下主流的机器学习算法的介绍，例如：算法描述、优点、缺点、适用范围、实现方法、应用场景等。

§1.2.1 开发平台

集成了主流工具的 API 介绍，包括 TensorFlow、PyTorch、Sklearn 等，用户通过点击查阅即可了解相关 API 的知识。

§1.2.2 开放能力

平台开放的能力包含了当下主流的可供学习使用的机器学习任务，涉及智能图像处理、自然语言处理、智能语音等多领域多方面的服务，可以让用户根据自身需要，依据自己的数据获得端到端的 AI 模型服务。

用户可以依据当前的进行该任务所选择的机器学习算法进行自主调节参数，用户在选择完参数之后可以自主决定是否选择上传本地的训练数据文件，如果选择上传，则会依据上传的数据文件训练模型，如果不选择上传，则会使用平台默认的数据文件进行训练。在参数和文件全部输入完毕后，用户点击“训练开始”即可开始训练模型。在模型训练完毕后，则会给出模型相应的性能指标。各模块集成的模型如下：

回归预测模块

回归预测模块包含海浪高度预测、波士顿房价预测、碳排放预测和股票预测四个基础功能，并提供端到端的训练功能。

智能图像模块

智能图像模块包含 MNIST 手写体分类、汽车目标检测、智能图像分割、人体关键点检测、智能图像生成、图片文字识别等基础模型，用户可以在相应模块下上传图片，获得预测结果。

自然语言处理模块

自然语言处理模块 (Natural Language Process，NLP) 深度整合了顶级的 NLP 技术，包括文本情感分类、机器翻译、智能阅读理解、语义匹配、对话系统、文本生成、句法分析等模块，主要利用 Transformer 模型建模，提供端到端的预测服务。

智能语音模块

智能语音模块包含智能语音合成 (Text To Speech, TTS) 和智能语音识别 (Automatic Speech Recognition，ASR) 功能，智能语音合成将用户输入的文本合成成音频，智能语音识别则可以将用户上传的语音中的文字识别出来。

其他模块

本模块包括了鸢尾花识别等基础分类任务的训练和预测功能。

§1.2.3 知识学习

知识学习模块主要包含了对主流的监督学习、无监督学习、半监督学习、强化学习等不同机器学习、深度学习方法的讲解。用户可以在该模块了解传统机器学习和前沿模型的建模思路和特点，以便于在自己建模时选择合适的模型。

第 2 章

开发平台

§ 2.1 Tensorflow

§ 2.2 PyTorch

§ 2.3 Sklearn

§ 2.4 其他经典开源模型

第 3 章

开放能力

本章节将详细讲述每个模块模型实现方法以及用户自主训练、预测的方法。

§3.1 回归预测

回归是一种通过建模和分析变量之间关系的的方法，其目的是通过模型来计算得出一个具体的值。

回归预测在很多的方面得到应用，例如：天气预测、机械寿命预测、商品价格预测、股价预测等都有广泛的应用。本平台的回归预测模块选取了海浪高度预测、波士顿房价预测、碳排放预测和股票价格预测等四个典型预测问题，并提供多种不同类型的建模方法，方便使用者充分利用本平台构建自己的模型。

§3.1.1 海浪高度预测

模型构建

海浪高度预测属于简单的回归预测，其样本的结构如表3.1-1所示：因此，建模时采用简单的多

列名	WVHT_1	WDIR_1	WSPD_1	WDIR_2	WSPD_2	WDIR	WSPD
含义	t-1 时刻高度	t-1 时刻风向	t-1 时刻风速	t-2 时刻风向	t-2 时刻风速	当前风向	当前风速
数据类型	float	float	float	float	float	float	float

表 3.1-1 海浪高度预测模型的数据特征及说明

层感知机 (Multi-Layer Perceptron, MLP) 即可快速完成任务，模型构造如图3.1-1所示，输入数据通过 z-score(式3.1) 进行归一化，然后输入到感知机中，最后输出得到预测值，并与真实值进行比较，计算损失值进行迭代。

$$z(x) = \frac{x - \mu}{\sigma} \tag{3.1}$$

用户可以将自己的数据输入进模型，根据自定义的超参数进行训练，得到符合自己数据的专有模型；也可以直接输入特征数据进行预测。

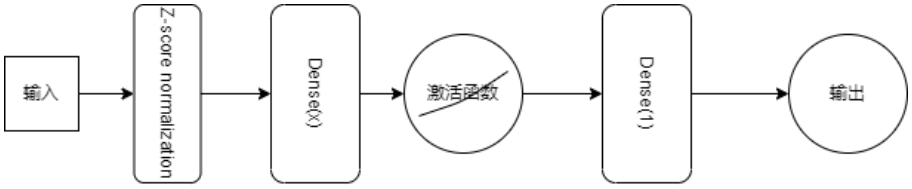


图 3.1-1 支持用户自定义的多层感知机

模型训练

训练界面如图3.1-2所示，该模型支持自定义神经元数量、学习率、训练轮数、激活函数、损失函数等超参数。用户通过输入上述超参数，可以选择上传本地数据进行训练，若不上传，则默认使

图 3.1-2 海浪高度模型训练界面

用线上数据。等待训练完成后，系统将返回该模型在验证集数据上的指标，如图3.1-3所示。此时，用户就得到了一个基于自设超参数和本地数据的模型。

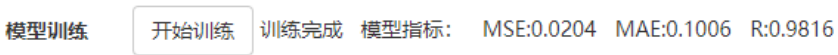


图 3.1-3 海浪高度模型训练指标

模型预测

平台支持使用用户自定义和模型进行预测和使用平台预先训练好的模型进行训练，以用户自训练模型预测 (图3.1-4) 为例 用户输入一条特征数据，点击**预测浪高**按钮，等待模型预测完成后，即可获得海浪高度的预测值。

图 3.1-4 使用自训练模型预测

§3.1.2 波士顿房价预测

波士顿房价预测属于经典的机器学习回归问题之一，该任务需要对 79 个与房价有关的特征进行处理和建模，最终得到能够预测房价的回归模型。该任务的评价指标是均方根误差 (Root Mean Square Error, RMSE)[1]

$$RMSE(y) = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

模型构建

由于特征维度较多，而且有很多的缺失值、部分特征存在异常值，因此首先进行特征分析和数据预处理。包括异常点检测、缺失值处理和数据填充、特征独热编码等。

1. **异常点检测**。异常点的分析以特征"GrLivArea" 为例，打印其与房屋售价"SalePrice" 的分布如图3.1-5所示，

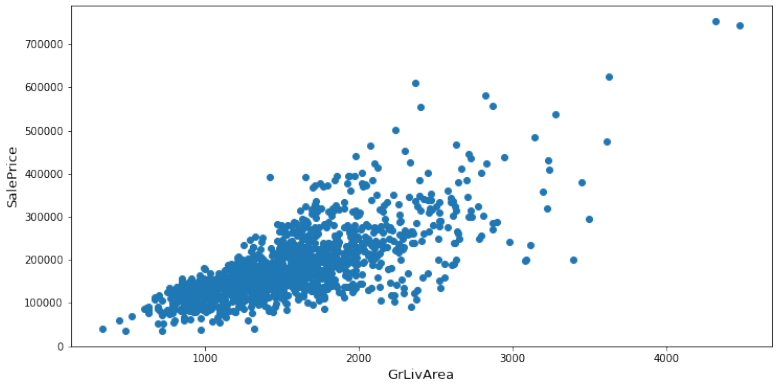


图 3.1-5 GrLivArea 与 SalePrice 分布图

清晰可见右上角有两个离群点，在训练时，可以考虑将这两条样本删除。同理，其他特征上包含异常值的样本，在数据预处理阶段也同时予以删除。

2. **缺失值处理**。对训练数据的缺失值进行分析，得到较高缺失比例的特征如图3.1-6所示，高缺失率的特征将影响模型的泛化能力，因此将缺失率 >30% 的特征 (PoolQC, MiscFeature, Alley, Fence, FireplaceQu) 删去，不参与后续训练。

	特征名称	缺失率
0	PoolQC	0.995885
1	MiscFeature	0.962963
2	Alley	0.937586
3	Fence	0.807270
4	FireplaceQu	0.473251
5	LotFrontage	0.177641

图 3.1-6 特征缺失比例

3. **缺失值填充**。对于剩余包含较少缺失值的特征列，分别采取填充 None 值、填充众数、填充平均数等方式，将缺失的部分填充上。

4. **正态性分析**。对于目标值 SalePrice，构建其直方图和正态概率分布图3.1-7，可见其呈明显正偏态分布，不利于回归建模。

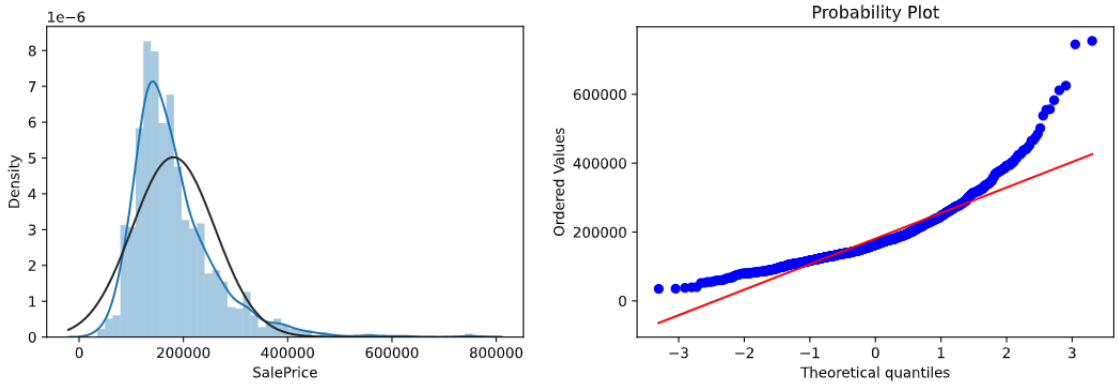


图 3.1-7 SalePrice 的直方图和正态概率分布图

因此，对价格列进行对数化处理，处理后的数据分布如图3.1-8，可见其基本符合正态分布。

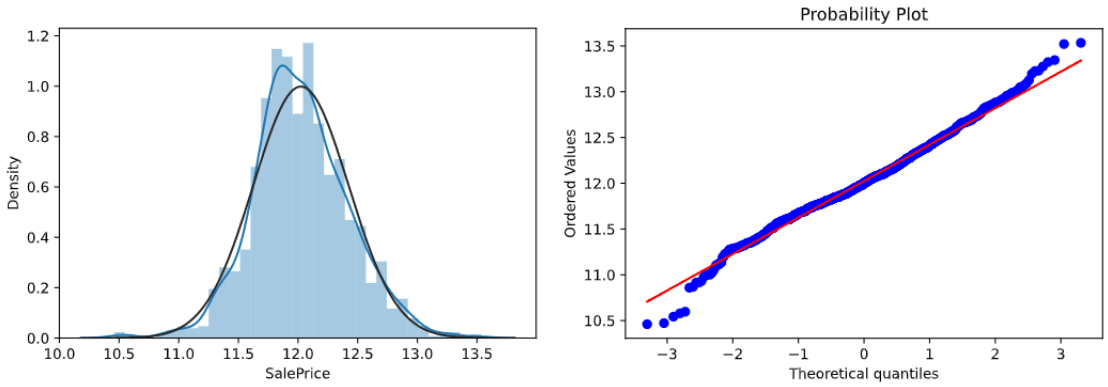


图 3.1-8 对数化后 SalePrice 的直方图和正态概率分布图

5. **特征独热编码**对于离散型的特征，需要将其进行独热编码 (one-hot)，才能转变为机器可以识别的特征。

经过数据预处理后，平台采用 xgboost 模型进行回归建模。xgboost(eXtreme Gradient Boosting) 是一种基于 GBDT 的算法，具有高效、灵活和轻便的特点，在数据挖掘、推荐系统等领域得到广泛的应用。通过预设的超参数集合，将训练样本输入 xgboost 模型中，经过数轮迭代后得到训练好的模型。

模型训练

平台提供了用户上传数据和自定义超参数训练模型的功能。用户可定义的超参数如表3.1-2所示

名称	含义	默认值
eta	学习率	0.05
max_depth	树的最大深度	10
subsample	随机采样比例	0.7
colsample_bytree	随机采样列数	0.8
num_boost_round	迭代轮数	1000
early_stopping_rounds	提前终止训练	200

表 3.1-2 房价预测模型的参数及说明

模型预测

用户可以选择使用自训练模型进行预测，也可以选择使用平台已经训练好的模型进行预测。预测时需要按照波士顿房价预测问题官方给出的数据集格式准备好自己的数据，并将数据在指定界面上传，平台将返回符合官方要求的数据格式的文件。

§3.1.3 碳排放预测

模型构建

模型训练

模型预测

§3.1.4 股票预测

模型构建

模型训练

模型预测

§3.2 智能图像

机器视觉（Machine Vision）是人工智能领域中发展迅速的一个重要分支，目前正处于不断突破、走向成熟的阶段。智能图像处理是指一类基于计算机的自适应于各种应用场合的图像处理和分析技术，本身是一个独立的理论和技术领域，但同时又是机器视觉中的一项十分重要的技术支撑。

智能图像处理的主要技术包括图像预处理、图像分割、目标识别和分类、目标检测和跟踪等主要研究方向。近年来，随着深度学习技术的不断发展，利用深度神经网络完成相关任务成为智能图像处理领域的主流方法。

在本模块中，平台基于深度神经网络技术，实现了综合性的图像智能处理功能。包括基础问题建模 (MNIST 手写体分类、图片文字识别)，智能目标检测、智能图像分割、人体关键点检测、智能图像生成 (通过文字描述生成图像) 等应用，可以快速根据需求生成所需的结果。

§3.2.1 MNIST 手写体分类

MNIST 手写体数据集来自美国国家标准与技术研究所 (National Institute of Standards and Technology, NIST). 训练集有 60000 个样本，测试集有 10000 个样本。样本经过归一化后，每张图片的大小是相同的。该数据集由 Yann LeCun 等人维护。图3.2-9展示了部分样本转换为图片后的样式。

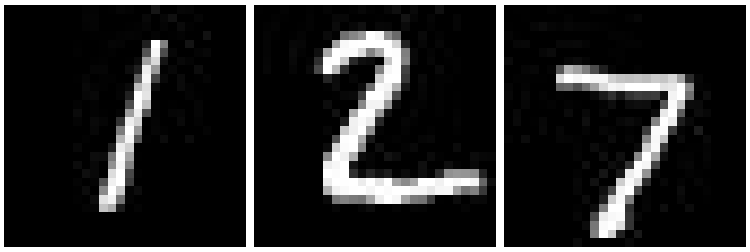


图 3.2-9 MNIST 数据集示例

模型构建

由于任务相对简单，我们以经典的 LeNet-5 来完成 MNIST 手写体识别任务。LeNet-5 网络结构如图3.2-10[2] 所示

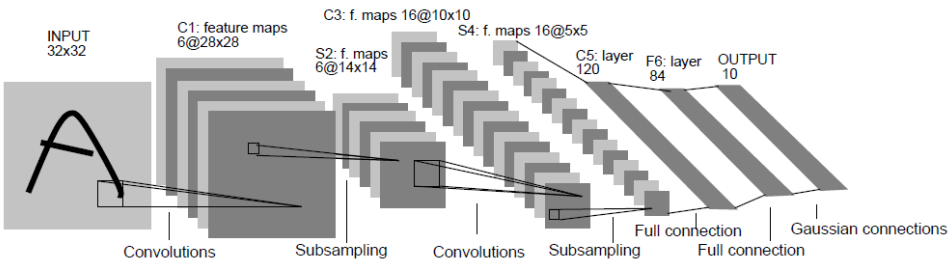


图 3.2-10 LeNet-5 网络结构 [2]

模型训练

用户可以上传数据并自设学习率和迭代轮数进行模型自训练。

模型预测

用户可以使用自训练的模型进行预测，也可以上传数据，使用平台预先训练好的模型进行预测。预测界面如图3.2-11所示

图 3.2-11 MNIST 预测界面

§3.2.2 图片文字识别

光学字符识别（Optical Character Recognition, OCR）是指对文本资料的图像文件进行分析识别处理，获取文字及版面信息的过程。亦即将图像中的文字进行识别，并以文本的形式返回。

模型构建

平台集成了开源离线 OCR 工具 tr，用户只需要上传图片，就可以得到图片中的文字。

模型训练

暂不支持自训练。

模型预测

用户上传带有文字的图片，平台经过计算后，返回图片中的文字，效果如图3.2-12所示

图 3.2-12 OCR 识别效果

§3.2.3 目标检测

YOLO(You Only Look Once) 系列算法采用单个神经网络直接预测物品边界和类别概率，实现端到端的物品检测。该方法的特点是将目标检测任务看作目标区域预测和类别预测的回归问题，实现快速检测的同时还达到较高的准确率。平台采用 YOLOv5 框架作为目标检测模型的基础框架。

模型构建

YOLOv5 的网络结构如图3.2-13所示，

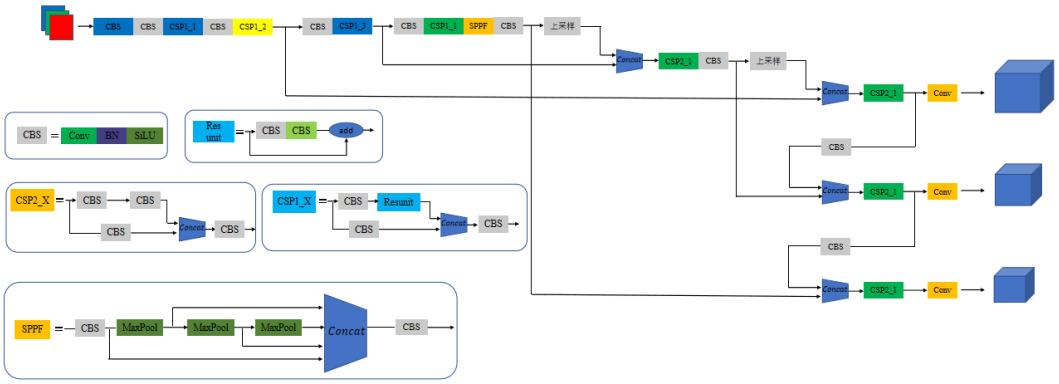


图 3.2-13 YOLOv5 网络结构图

由于模型代码已经开源 (https://github.com/ultralytics/yolov5)，平台直接利用了其开源代码搭建模型。

模型训练

通过 YOLOv5 的开源代码，我们可以利用预训练好的代码和权重进行训练。通过以下指令，就完成了基于 yolov5n 的网络结构和 COCO 数据集，BATCH 大小为 128 的模型的训练。

```
python train.py --data coco.yaml --cfg yolov5n.yaml --weights '' --batch-size 128
```

该模型训练耗时和算力消耗都比较大，因此平台不提供自主数据训练。

模型预测

平台支持使用不同参数量的 YOLOv5 模型进行预测，用户通过上传本地的图片，选择 YOLOv5 模型的版本，进行目标检测模型的预测。具体版本类型如表3.2-3所示:

Model	size(pixels)	Speed(ms)	params(M)	FLOPs@640 (B)
YOLOv5n	640	45	1.9	4.5
YOLOv5s	640	98	7.2	16.5
YOLOv5m	640	224	21.2	49.0
YOLOv5l	640	430	46.5	109.1
YOLOv5x	640	766	86.7	205.7

表 3.2-3 YOLOv5 可选版本及其部分参数

使用 YOLOv5x 对图片进行分割的结果如图3.2-14所示

图 3.2-14 YOLOv5x 检测结果

§3.2.4 智能图像分割

图像分割技术是目前计算机视觉领域最热门的技术之一，许多应用了计算机视觉的场景都需要对图像进行智能分割，以充分理解图像中的内容，以便于机器分析图像中各个部分之间的关系。图像分割技术主要分为语义分割和实例分割等。

光伏发电站，是指一种特殊材料电子元件 (诸如晶硅板、逆变器等) 组成的利用太阳能发电的体系，与电网相连并向电网输送电力。利用遥感图像和图像分割技术，可以对光伏发电站的空间分布进行精准确定。因此，本平台以光伏发电站的遥感图像为例，提供了光伏板智能图像分割服务。

模型构建

U-Net 是一种 U 型的类 Encoder-Decoder 卷积神经网络，最初被应用于医疗图像分割，后来实验证明其在其他领域的图像分割也有较好的效果，是实现语义分割或实例分割的较为基础的模型，其网络结构如图3.2-15[3] 所示:

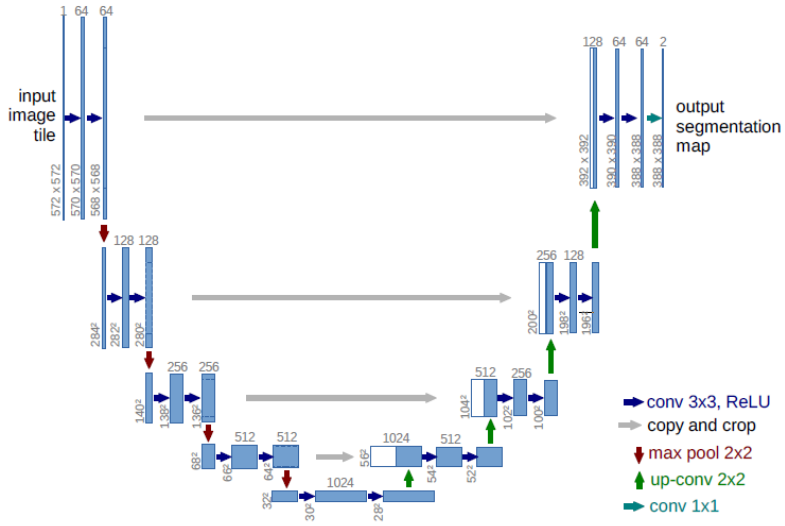


图 3.2-15 U-Net 的网络结构图

模型首先通过卷积操作进行特征融合，通过池化层进行降维，然后再通过卷积层进行上采样，最后通过 1×1 的卷积进行结果的输出。

模型训练

1. **数据准备**。U-Net 模型训练需要同时准备原图像和标注图像，平台使用了 zenodo 上的开源遥感光伏数据集 [4] 来建模，其示例如图3.2-16所示，标注图像需标注出光伏板的位置

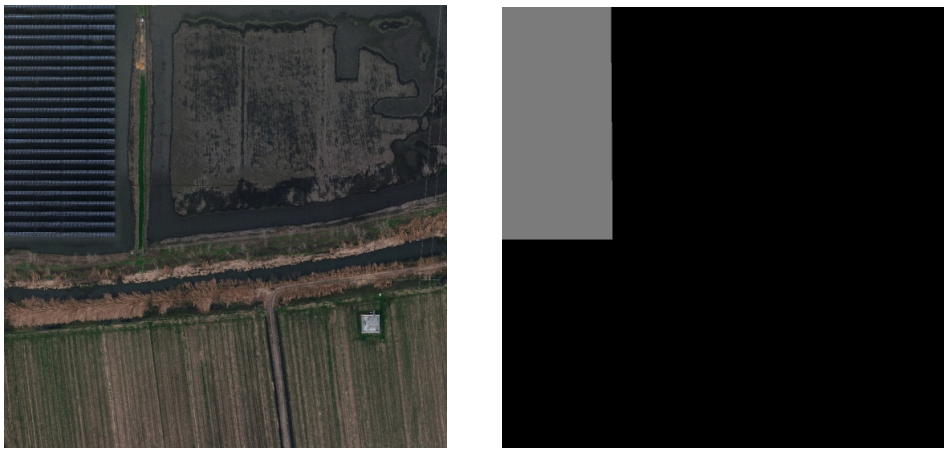


图 3.2-16 U-Net 数据集示例

2. **模型训练**。首先将原始图像进行数据增强，例如随机限制对比度、随机增加白噪声、随机锐化和模糊等，然后调整分表了使得图像像素长宽能被 32 整除，最后将图像规范化成为神经网络模型可接受的输入。训练时以 DiceLoss 作为损失函数，使用交并比作为训练时更新模型的指标。

模型预测

模型预测支持单张图片预测，也支持批量预测，用户通过平台上传本地图片，模型运行完成后会返回分割后的结果。最终分割的效果如图3.2-17所示

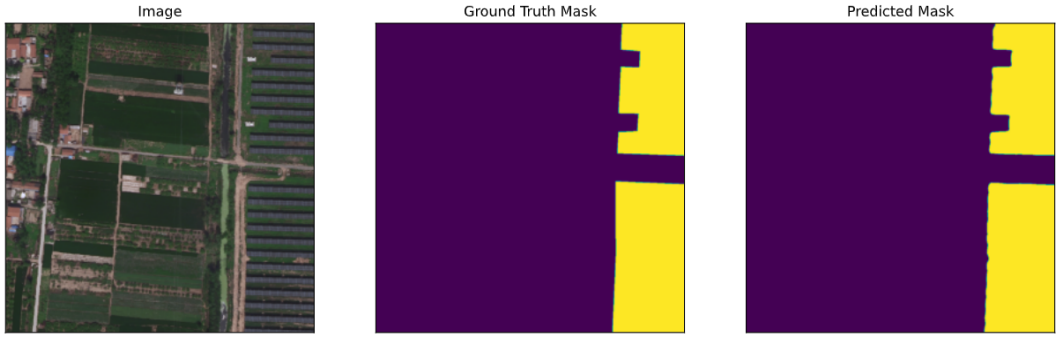


图 3.2-17 U-Net 分割光伏图像

§3.2.5 人体关键点检测

人体关键点检测，Human Pose Estimation，指通过识别一些人体关键点，如头部、肩膀、关节等，描述人体骨骼信息。人体关键点检测是计算机视觉的重要应用之一，以人体关键点检测为基

础，可以进行行为识别、人物跟踪、步态识别等相关研究。按照检测目标的维度，人体关键点检测一般可以分为二维关键点检测和三维关键点检测。

图 3.2-18 二维人体关键点检测 [5]

模型构建

MMPose 是一个基于 PyTorch 的开源姿态估计工具箱。基于 MMPose 的开源框架，我们可以简单的搭建一个人体关键点检测的建模、训练和预测服务。

模型训练

平台选择了 MMPose 自带的预训练模型搭建服务，MMPose 框架也支持使用者基于自有数据进行模型微调或者重新训练，指令如下

```
python tools/train.py ${CONFIG_FILE}$ [optional arguments]
```

模型预测

用户可以直接上传本地图片进行预测，模型预测的效果如图3.2-19所示

图 3.2-19 预测结果示例

§3.2.6 图像生成

根据文本生成对应的图像是多模态任务之一。解决这一问题的主流方法有 VAE(Variational Auto-Encoder)，DRAW（Deep Recurrent Attention Writer）以及生成对抗网络 (Generative Adversarial Networks, GAN) 等。

模型构建

FuseDream 是一个结合 CLIP(Connecting Text and Images) 和 GAN 的开源文本生成图像模型

1. **生成对抗网络模块**。GAN 模块的结构如下图3.2-20所示

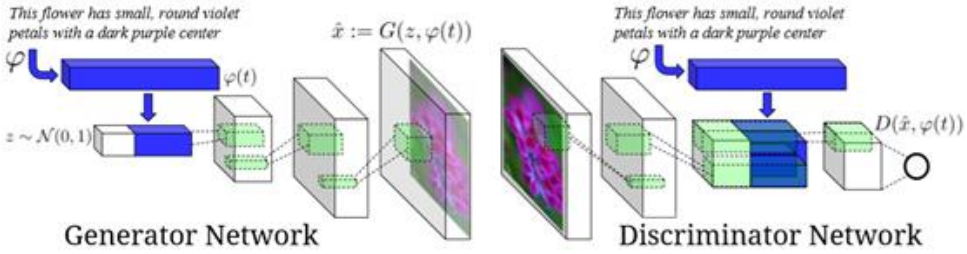


图 3.2-20 生成对抗网络结构图

生成对抗网络模型由两部分组成，一个是生成模型 (Generator)，一个是判别模型 (Discriminator Network)。生成模型的任务是尽力模仿输入的数据生成类似的数据。判别模型的任务是判断给定的实例是否是由生成器伪造的。

2. **CLIP 模块**。CLIP 的英文全称是 Contrastive Language-Image Pre-training，即一种基于对比学习的文本-图像对的预训练模型。如下图3.2-21所示，CLIP 包括两个模型：Text Encoder 和 Image Encoder，其中 Text Encoder 用来提取文本的特征，可以采用 transformer 模型；而 Image Encoder 用来提取图像的特征，可采用 CNN 模型。

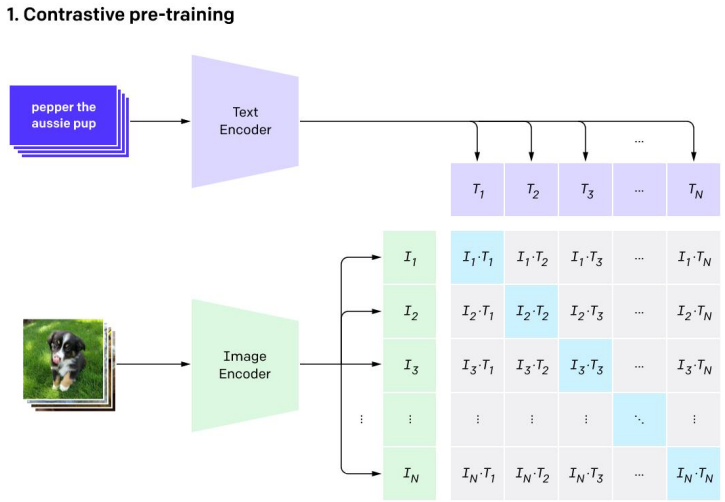


图 3.2-21 CLIP 结构图 [6]

模型训练

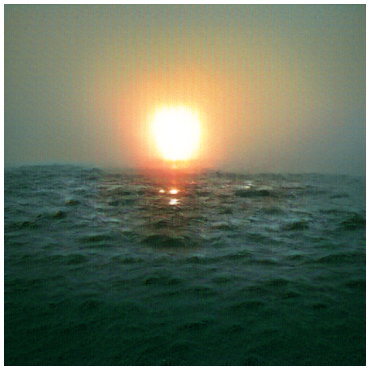
训练时按照如下步骤进行

1. 通过对数据集随机扰动，进行数据增强。
2. 使用 CLIP 模型计算文本与图像的相似度，依靠增强的数据，增加 CLIP 的鲁棒性。
3. 将 CLIP 模型计算得分较高的一部分数据送入 GAN 模型，进行训练。该模型利用了预训练的 BIG-GAN 权重，在使用时无需重复训练，根据用户输入的文本进行微调即可。

模型预测

用户输入文字，模型经过生成后，输出对应的图片。示例如下

(a) A photo of a blue dog.



(b) Sun rises from sea.

(c) A small flower blooms on the cliff.

图 3.2-22 生成图像示例

§3.3 自然语言处理

自然语言处理 (Natural Language Processing, NLP) 是计算机科学与人工智能领域中的一个重要研究方向。它是关于如何使计算机理解人所使用的自然语言的研究，是融合计算机科学、语言学、数学等学科于一体的科学。自然语言处理任务通常包括句法分析、文本分类、机器翻译、阅读理解、人机对话等多种任务，在舆情分析、智能机器人等研究上扮演了重要的角色。平台以将这些任务为例，结合最新预训练模型技术进行建模。

§3.3.1 文本情感分类

文本分类任务是 NLP 任务中较为基础的应用。文本情感分类一般是指按照文本内容，将文本情感分为两类 (积极、消极) 或三类 (积极、中性、消极)。

传统文本情感分类任务一般是基于中文分词、词向量结合 TextCNN 等神经网络模型实现的，近年来，随着各种 Transformer 预训练模型的出现，基于预训练模型进行微调 (Fine-Tune) 的文本分类模型逐渐成为主流。本平台的文本分类模型即采用了这种方式。

模型构建

使用 Bert 等 Transformer 模型进行微调预测的网络结构如下图3.3-23所示，中文文本一般不再需要分词，而是经过在词表中的位置映射，转换为机器能够理解的编码。对于文本分类任务，BERT 模型在文本前插入一个 [CLS] 符号，经过训练和，该符号可以学习到整段文本的语义表示，并将该符号对应的向量输出得到最后的分类结果。

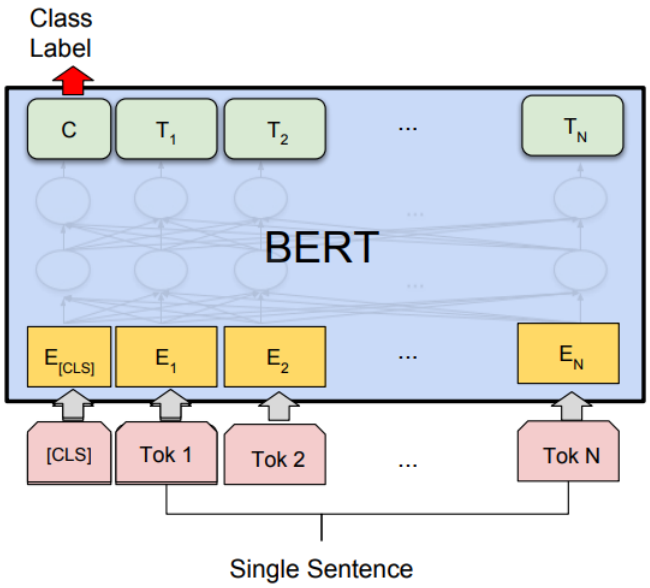


图 3.3-23 Bert 文本分类 [7]

构建模型时，采用 RoBERTa 作为基础的预训练模型，相比于 Bert，RoBERTa 基于更大的训练数据和更长的预训练时间进行预训练。

模型训练

平台支持文本分类任务的自主训练，用户同时上传训练数据集和测试数据集，平台建模完成后，会提供预测的结果文件，可供用户下载。

模型预测

用户也可以只提交测试文件，平台基于线上模型进行预测，同时返回预测结果。

§3.3.2 机器翻译

机器翻译 (Machine Translation)，是利用计算机把一种自然语言翻译为另一种自然语言的过程。早期的机器翻译一般是利用规则或者统计模型，基于规则或统计学模型进行翻译。近年来，随着深度神经网络和编码器-解码器 (Encoder-Decoder) 模型的发展，使用这类模型完成机器翻译任务逐渐成为主流方式。

模型构建

平台采用了基于编码器-解码器和注意力机制的 Transformer 模型来训练英译中模型。结构如图3.3-24所示：

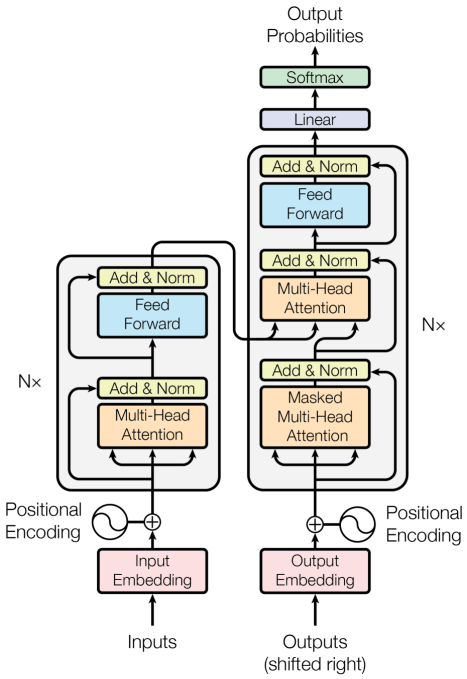


图 3.3-24 注意力机制的 Endcoer-Decoder 模型 [8]

各部分的构成如下：

1. 编码器 Encoder。首先通过字向量与位置编码将原文映射，然后多次重复 x 个 encoder 块，每个 encoder 块包括注意力机制 (Attention) 层、层归一化 (Layer Normalization, LN) 和前馈神经网络 (FFN) 层。

2. 解码器 Decoder。解码器将译文作为输入，同时接收编码器的输出，经过 Attention 层、FFN 层和层归一化后，得到对应位置每个词的输出概率。

模型训练

平台支持用户自训练模型，用户可以自己准备训练语料，将分词后的原文、译文按照“原文 + Tab + 译文”的格式上传 (原文、译文可以选择任意语种)。

模型预测

用户可以使用自主训练的模型，输入文本进行翻译，也可以基于平台的预训练中译英模型，输入中文来获得对应的英文。

§3.3.3 阅读理解

模型构建

模型训练

模型预测

§3.3.4 语义匹配

语义匹配是 NLP 领域的基础任务之一，直接目标就是判断两句话是否表达了相同或相似意思。语义匹配可应用于推荐系统、智能问答、搜索引擎、文本查重等。传统的文本匹配技术有词袋模型 (Bag of Words)、TF-IDF、Jaccard 矩阵、SimHash 等算法，主要基于词的层级进行匹配，近年来，结合预训练模型的深度神经网络模型逐渐成为这一领域的主流模型。

模型构建

我们利用 Bert 的 [SEP] 字符的特性，可以将两句话拼接在一起，然后基于 [CLS] 字符来表示两句话是否具有相近语义，构造一个简单的二分类任务，网络框架如下图3.3-25所示

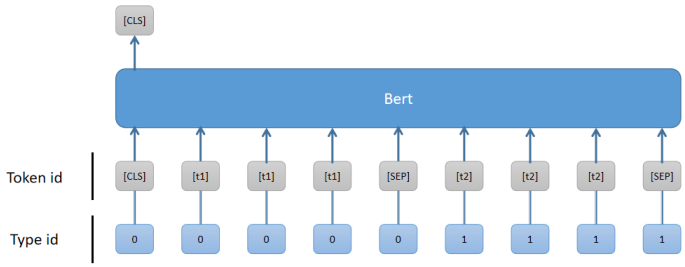


图 3.3-25 注意力机制的 Endcoer-Decoder 模型

模型训练

训练用数据集每行由两列文本和一列标记组成，标记用 1/0 来表示相似/不相似。经过 tokenizer 后变为上图中 bert 的输入。损失函数选择 cross_entropy，优化器选择 Adam 进行训练。

模型预测

用户在页面输入待预测的两段文本，模型最后返回其预测结果。

§3.3.5 文本生成

文本生成是基于给出的信息，生成近似于自然语言的文本序列。根据应用场景，自然语言处理下的文本生成任务又可以用于机器翻译、对话系统、故事、诗歌生成、摘要生成等。近年来，随着多模态应用的兴起，根据图片、视频生成文本方面也出现了很多优秀的模型。本平台提供的文本生成服务是基于语言模型 (Language Model) 和随机采样的文章续写。

模型构建

基于单向语言模型，以自回归的方式进行解码的方式进行建模。单向语言模型结构如下图3.3-26所示，即在预测 t 时刻的字符时，只能依赖 t 时刻以前已知的字符进行推理。而自回归解码器将

解码器自己当前步的输出加入下一步的输入，通过这种方式，可以逐步推导接下来最可能生成的文字，直至推导出结束符<EOS>或达到最大长度。

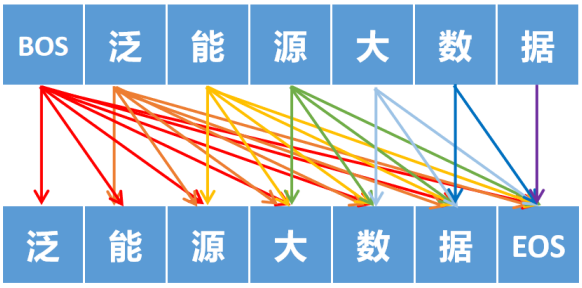


图 3.3-26 单向语言模型

模型训练

按照建模思路，平台将使用 NEZHA??的预训练权重作为语言模型。由于文章续写只依赖于文字本身，因此不需要使用 segment_id 来区分更多细节。

模型预测

模型解码时，按照自回归的方式逐步解码，并依据概率，在解码空间内随机采样，逐步生成后续的文本。

§3.3.6 对话系统

智能对话系统按照需求类型一般分为任务导向型 (task-oriented) 对话系统和非任务导向型 (non-task-oriented) 对话系统。任务型对话系统是基于一定的任务目标，通过预设流程或者插槽填充 (slot filling) 的方式引导对话完成；非任务导向性对话系统一般是以闲聊为目的，利用 Encoder-Decoder 模型，基于大量聊天对话语料进行训练，基于当前对话以及上文内容，自动生成回复内容。

模型构建

我们选择非任务导向性对话系统作为平台的智能对话系统，相对于一般的文本生成任务，对话任务有两个难点，一是需要对说话人的身份予以区分，二是生成的文本需要符合上文的语境和连贯性。利用基于预训练模型的 Transformer 结构进行微调，其网络结构如下图3.3-27所示

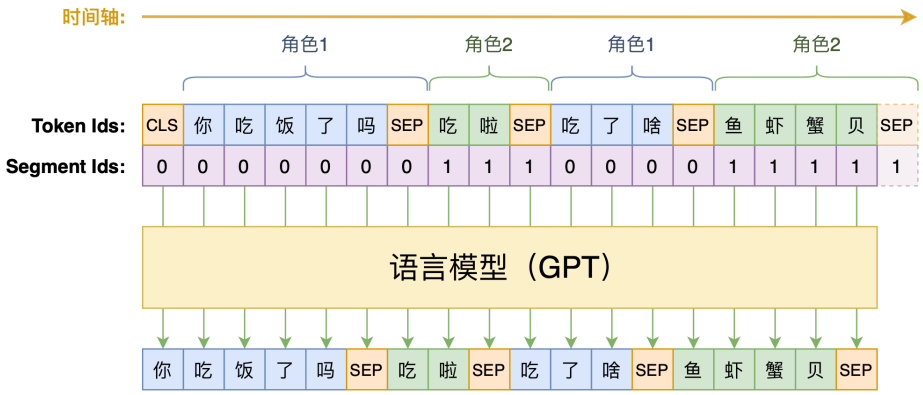


图 3.3-27 闲聊机器人模型架构 [9]

模型的输入文本是历史的对话记录，将每句话用 [SEP] 符号分隔开，同时以不同的 Segment Id 作为说话人角色的区分，同时使用相对位置编码的 NEZHA 作为基础的预训练模型。与基于 Bert 的分类任务不同，基于 Bert 的生成任务，不再以 [CLS] 的输出作为结果，而是利用 decoder 进行解码，再通过集束搜索 (Beam Search) 或贪婪搜索 (Greedy Search) 等方式将输出结果映射到解码空间。

模型训练

平台基于 LCCC 数据集 [10] 和 GPT 预训练模型进行训练。数据集给出一段段对话记录，每一段对话记录都可以通过拆解获得多条训练样本，每条样本包含一句或多句历史对话和一句目标文本，历史对话用于输入模型，目标文本用于验证模型的输出、计算损失并优化模型。

模型预测

模型支持两人之间的对话，用户通过将历史对话输入平台，平台运行模型得到最后的输出。

§3.3.7 句法分析

句法分析的主要任务是识别出句子所包含的句法成分以及这些成分之间的关系，主要分为以下两个任务：

1. 成分句法分析 (Constituency Parsing): 分析句子的成分，给出一棵树由终结符和非终结符构成的句法树。

2. 依存句法分析 (Dependency Parsing): 分析句子中词与词之间的依存关系，给一棵由词语依存关系构成的依存句法树。

成分句法分析（constituent parsing）是自然语言处理中的一个基础任务，它的任务是给定一个句子，分析出句子的短语结构句法树。例如给定句子“The little boy likes red tomatoes .”，它的成分句法树如下图3.3-28所示：

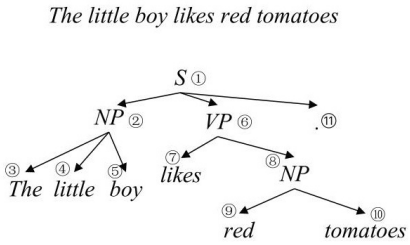


图 3.3-28 Syntactic tree[11]

依存句法分析则是为了分析句子中各个部分，即主、谓、宾、定、状、补，以及他们之间的依存关系，如主谓关系、动宾关系等。我们可以将依存分析抽象为一棵树??

模型构建

我们基于 HanLP 框架 [12] 构建了句法分析的服务，HanLP 是面向生产环境的多语种自然语言处理工具包，支持 PyTorch 和 Tensorflow 框架。

模型训练

框架嵌套了基于预训练模型的算法，因此不需要外部数据的训练。

模型预测

用户输入待预测的句子，模型分析完成后返回词的成分以及词之间的依存关系，结果如图3.3-29所示

```
中心旨在紧密围绕山东省和国家需求，构建互联互通的泛能源大数据数据体系。
```

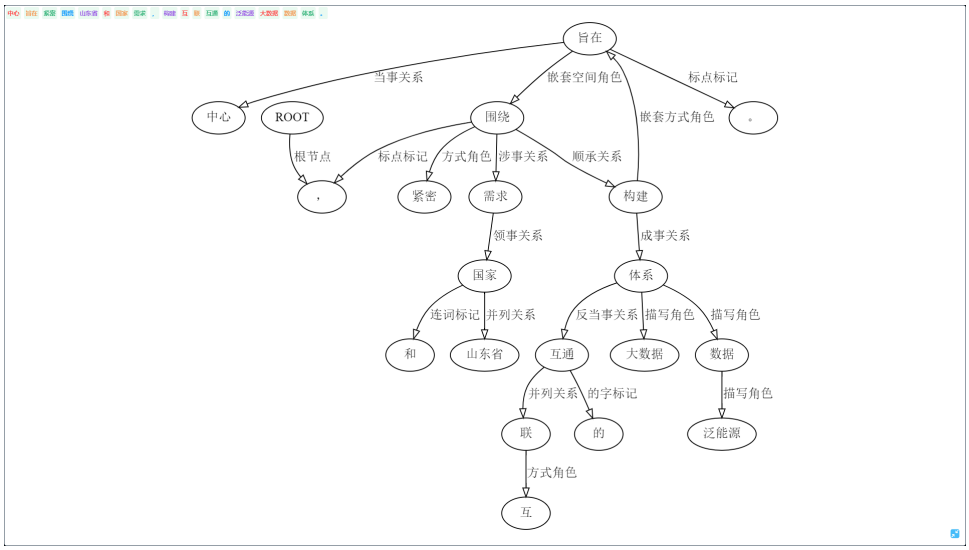


图 3.3-29 依存句法树

§3.4 智能语音

早期的语音技术主要包括利用信号处理的相关知识，进行语音的去噪、波形分析等，随着人工智能技术的发展和计算机算力的提升，基于深度学习模型的智能语音技术蓬勃发展。广义的智能语音技术主要包括**矢量信号处理 (Vector Signal Processing, VSP)**、**自动语音识别 (Automatic Speech Recognition, ASR)**、**自然语言处理 (Natural Language Processing, NLP)**，以及**语音合成 (Text to Speech, TTS)**。这几个模块构成了一个完整的**语音交互 (Voice User Interface)**系统，其结构如下图3.4-30所示

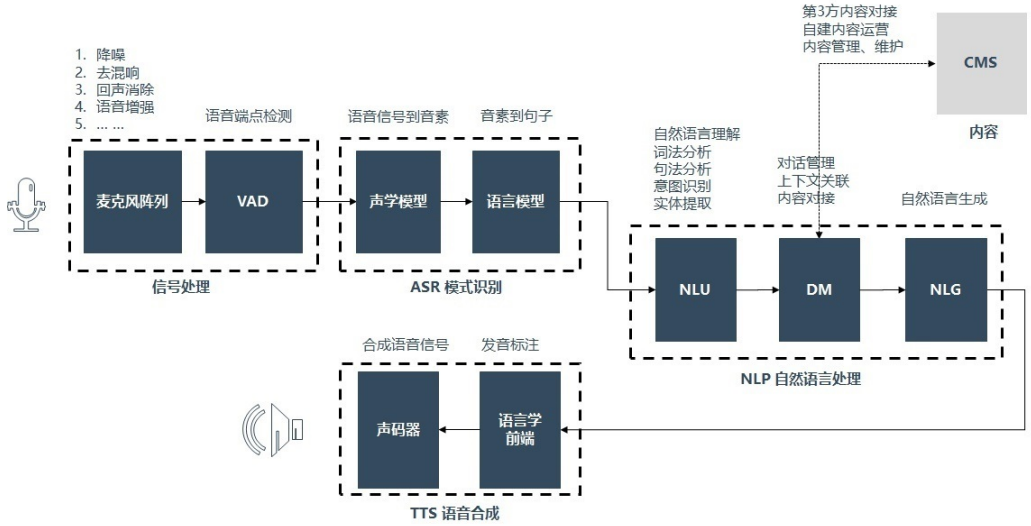


图 3.4-30 语音交互系统 [13]

其中，语音识别和语音合成属于智能语音技术的核心部分。

§3.4.1 语音识别

语音识别是将音频文件转为自然语言文字的技术，主要由特征提取、声学模型、语言模型构成。其衍生技术包括声纹识别、机器唤醒、情绪识别等。

语音识别最终是统计优化问题，给定输入序列 $O = \{O_1, \ldots, O_n\}$，寻找最可能的词序列 $W = \{W_1, ..., W_m\}$，即寻找使得概率 $P(W|O)$ 最大的词序列。最开始的研究都是分别求取声学和语言模型。随着深度学习和大数据的端到端（End-to-End）方法发展起来，能将声学和语言模型融为一体，直接计算 $P(W|O)$。

模型构建

构建语音识别模型主要步骤包以下步骤:

1. 特征提取。特征提取是将语音的波形信号转换为机器能识别的特征，主要方法包括梅尔频谱特征、FBank(Filter Bank) 特征提取和梅尔倒谱系数 (Mel-scaleFrequency Cepstral Coefficients, MFCC) 特征提取

2. 建立数据集。语音选择的基本单位是帧（Frame），一帧数据是由一小段语音经过 ASR 前端的

声学特征提取模块产生的，整段语音就可以整理为以帧为单位的向量组。每帧的维度固定不变，但跨度可调，以适应不同的文本单位，比如音素、字、词、句子。

3. 模型搭建。传统方式是音学模型和语言模型分别建模，即音学模型使用隐马尔科夫模型和高斯混合模型学习信号的动态和静态特征；语言模型基于音学模型的特征求解 $P(W|O)$，常用的方法基于 n 元语法（n-gram Grammar）或 RNN。端到端的方法可以使用连接时序分类，(Connectionist temporal classification, CTC) 或者 Attention 的方法。

模型训练

平台使用 WeNet??模型作为语音识别模型的训练和预测工具，其使用 conformer 网络结构和 CTC/attention loss 联合优化方法，具备统一的流式/非流式语音识别方案。训练过程如下：

1. **数据准备**。首先要按照如下形式准备 wav.scp 和 text 两个文件。

列名	**wav_id**	**wav_path**
示例	id001	/path/to/id001.wav

表 3.4-4 wav.scp

列名	wav_id	text_label
示例	id001	text of id001.wav

表 3.4-5 text

2. **提取 cmvn 特征 (可选)**。提取倒谱均值方差归一化 (cepstral mean and variance normalization, CMVN) 系数，用于对声学特征归一化。

3. **生成文本和 token 的对应词典**。由于计算机无法直接理解汉字，所以需要将音频中出现的汉字以及特殊字符如静默字符<blank>，未知字符<unk>，音频开始/结束字符<sos/eos>映射到数值上，一般用自增的整数来表示每个字符，这些整数称为 token。

4. **构造 WeNet 可识别的数据格式**。wenet 模型需要以 json 格式表示的表征音频和文本的统一数据格式，由 key, value 和 txt 三个键值对组成，示例如下：
```
{"key":"id001", "wav":"/path/to/001.wav", "txt":"本段音频的文本内容"}
```

5. **模型训练**。使用run.sh中的脚本进行训练。

模型预测

用户上传 wav 音频文件后，平台通过识别最后得到该段文本的文字，并将文本展示到界面上。

§3.4.2 语音生成

智能语音合成 (TTS) 是把文字智能地转化为自然语音流的技术，即输入文本，输出机器可以解析的波形。近年来，个性化 TTS、带有情绪的 TTS 等逐渐成为研究的热点。

模型构建

我们使用了基于 Tacotron2 的模型训练了 TTS 模型。Tacotron 是 Google Brain 推出的端到端的开源 TTS 深度神经网络模型，Tacotron 是一个带有 Attention 机制的序列到序列 (Sequence-To-Sequence，Seq2Seq) 生成模型。使用 < 文本序列，语音声谱 > 配对的方式进行训练。Tacotron2 是 Tacotron 的改进版，使用了 LSTM 和卷积层代替了原来的 CBHG 模块，并且使用 WaveNet 的声码器代替了 Griffin-Lim 算法。

模型训练

Tacotron2 原模型是针对英文语音合成的模型，要训练针对中文的模型，首先要对数据进行**拼音化处理**，即：

1. 对于输入文本从左向右遍历，优先从词拼音库中以词的方式匹配读音，若没有匹配到词语，从字拼音库中获取该汉字的拼音；
2. 对于数字、ip 地址等，首先根据规则转化成文字，再转化为拼音；
3. 由于输入是文字转化而来的拼音序列，所以在合成阶段，允许部分或全部的拼音输入。

数据处理完成后，即可使用原有脚本进行本地训练。

模型预测

用户可以输入中文句子，也可以输入拼音进行合成。合成后的音频结果可以在平台直接播放，也可以点击下载进行下载保存。

§3.A 各模块模型服务接口文档

§3.A.1 回归预测

海浪高度预测

参考文献

[1] House prices - advanced regression techniques. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview.

[2] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[3] Z. Zhou, Mmr Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020.

[4] Jiang Hou, Yao Ling, and Liu Yujun. Multi-resolution dataset for photovoltaic panel segmentation from satellite and aerial imagery, August 2021. Data document can refer to the preprint https://es sd.copernicus.org/preprints/essd-2021-270/.

[5] 计算机视觉方向简介 | 人体骨骼关键点检测综述. https://developer.aliyun.com/article/639017.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *arXiv*, 2017.

[9] 苏剑林. 动手做个 dialogpt：基于 lm 的生成式多轮对话模型. https://spaces.ac.cn/archives/7718, Sep 2020.

[10] Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer, 2020.

[11] Jiangming Liu and Yue Zhang. In-order transition-based constituent parsing. *Transactions of the Association for Computational Linguistics*, 5:413–424, 2017.

[12] Han He and Jinho D. Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] 智能语音专题（一）：智能语音交互的概念. https://zhuanlan.zhihu.com/p/109885562.