平治 AI 数字人的多模态大模型技术白皮书

二零二四年五月

目录

1 概述

1.1 人工智能技术战略意义

当前，新一代人工智能已成为世界各国的竞争焦点，抢占未来技术战略制高点意义重大。由于持续开放的动态环境、各行业领域不断攀升的系统复杂度以及快速扩大的数据规模总量，智能技术应用需求不断增长，智能形态和认知水平持续深入发展。从互联网到移动互联网再到物联网、星联网时代，计算硬件体积不断压缩、功耗与成本持续降低，新一代人工智能已经成为共性支撑技术，推动经济、社会、民生、国家安全、制造等领域进行数字化和智能化转型。另一方面，伴随互联网/行业大数据、并行计算能力、机器学习算法的突破和人类智能本质认知探索的深入，新一代人工智能发展还在继续加速。随着生成式预训练(GenerativePre-Train，GPT)、基于 Transformer 的双向编码器表达(BidirectionalEncoderRepresentation from Transformers, BERT)、GPT-3、DALL-E、SwitchTransformer、华为盘古、悟道、ERINE、M6 等大规模预训练模型快速涌现，人工智能研究领域正在经历一场有监督学习向无监督学习条件下“大数据+大模型”的大规模预训练范式转变，即基于海量广域数据训练并且经过微调学习自动适应应用于广泛下游任务的模型。大规模预训练模型起源于自监督的语言模型，自监督的深度语言神经网络模型最初只在自然语言处理领域展开研究，直到 2018 年 BERT 模型在 11 项 NLP 任务基准上都打破了纪录，取得了巨大成功，性能远超第二名。2019 年以后，基于自监督学习

的语言模型已成为基础性方法，这与 2012 年基于卷积神经网络 AlexNet 在 ImageNet2012 上的突破很相似，标志着一个大模型时代的开始。当下，自然语言处理(NaturalLanguage Processing，NLP)领域几乎所有的目前最先进的模型(State-0f-The-Art model,SOTA)都是基于 Transformer 的大模型架构进化而来的，而这种趋势也正在向图像、视频、语音等不同模态、不同领域扩散蔓延。

人工智能从单模态有监督迈向多模态自监督学习时代。目前网络数据中 90%以上是图像与视频，更多知识蕴含其中。人类的信息获取环境感知、知识学习与表达，都是采用跨模态的输入输出方式。如何设计计算机模型并使其具有强大的无监督学习与通用知识迁移能力，使不同领域任务在统一框架下实现基于低标注代价的性能提升?一种可行的路径是通过跨模态语义关联,提升多模态融合理解以及跨模态转换与生成性能。

当前，单模态预训练模型在数据规模和模型性能方面已经遇到瓶颈，而且单模型只涵盖了互联网数据中的单一模态信息，更丰富的包含文本、语音、图像、视频等多种模态数据的信息并未被充分利用与学习。此外，人类的信息获取、环境感知、知识学习与表达，都是通过多模态信息方式进行执行。因此，为实现更加通用的人工智能模型预训练模型必然由单模态往多模态方向发展,需将文本、语音、图像视频等多模态内容联合起来进行学习，并专注多模态内容之间的关联特性与跨模态转换问题。这样一方面可以引入多维度的信息，另一方面可以利用互联网上大量的多模态数据，使得模型能够学习更通用

化的特征表示，以此增强模型的通用性和泛化能力。

1.2 平治 AI 数字人的多模态大模型简介

平治 AI 数字人的多模态预训练大模型架构与 GPT 和 BERT 类似，也是基于自注意力机制 Transformer 深度学习模型,其最大特点是模型的输入由单一模态的文本拓展到文本、语音、图像、视频等多个模态数据同时作为输入，强调音频的流式输入、输出以达成数字人交互的顺畅性，强调理解图像、视频的人体动作、表情、情绪并在数字人驱动参数上体现出对应的反馈。平治 AI 数字人的多模态大模型主要指输入包括两种及以上模态的、参数量大于亿级的深度学习网络模型。单模态大模型主要是指模型输入只包括一种模态(如只包括语音、图像或文本)的、大规模参数量的深度神经网络模型。

一个关键的科学问题是如何设计神经网络模型并使其具有强大的无监督学习与通用知识迁移能力，使不同领域任务在统一框架下实现基于低标注代价的性能提升。一种可行的路径是通过跨模态语义关联，提升多模态融合理解以及跨模态转换与生成性能。

多模态预训练模型通常采用无监督学习的方法进行大规模训练，预训练数据来自互联网上大量的多模态数据，例如网页、视频等，无需人工标注，从而具有良好的拓展性和通用性。在不微调或采用少量数据微调的情况下，多模态预训练模型就可直接用于解决不同类型的多模态数据处理问题，例如为视频自动配上字幕、声音，输入声音和文本自动生成图像或视频片段等。多模态数据相比单模态数据更具

有研究意义，但同时也存在更多的困难与挑战。

具体而言，平治 AI 数字人的多模态预训练模型的研究面临以下挑战:

(1)模型构建不完善

现有的多模态预训练模型往往忽略了视觉内容、特别是理解图像、视频的人体动作、表情、情绪并在数字人驱动参数方面的语义编码，对视觉内容常使用离线训练的目标检测模型进行编码，然后进行“图像文本”的匹配，而“目标-语义”才是实际任务中真正的需求所在。另一方面,现有模型的训练方式和优化机制沿用了语言预训练模型的范式,相对于 NLP 任务中的自由文本,多模态对齐数据的获取难度大，面临着数据噪声大、不同模态缺失等挑战。为此，需围绕基于全注意力机制的跨模态关联建模,对 Transformer 模型在视觉中的应用进行改进;充分发挥多模态预训练模型对不同模态数据间关联互补特性的有效建模能力，以及模态信息缺失情形下的鲁棒分析能力;设计与多模态预训练模型网络结构尽可能兼容的多任务学习机制，优化模型参数的学习机制。

(2)知识利用不充分

目前，多模态预训练领域通过增加训练数据和计算资源来提升模型的性能,而这种粗放耗能方式和人类大脑的集约计算方式是完全不一样的，并且这种粗放的学习方式随着数据量和计算机资源的增长，所带来的收益会越来越小，而且很快会达到性能瓶颈。另外，数据≠知识，数据是未经组织和处理的文本、声音、图像和视频，本身没有

意义;而知识与内容相关，是可表示、可计算、可生长、有逻辑、可推理的更高层智能载体，是有意义的。为此，需构建知识引导的适用AI数字人的多模态预训练模型，研究多模态数据与知识统一表征的预训练模型构建、知识嵌入的多模态预训练模型构建方法、大规模预训练模型的可解释性理论，提升模型的可信可解释性。

(3)理解与生成不统一

多模态内容的理解与表达是人类智能中的最重要能力之一。然而，现有的预训练模型通常极少同时具备理解与生成能力，他们要么仅仅关注理解类而缺乏生成的能力，或者仅关注生成任务却对理解类任务表现欠佳。在实际应用中也通常需要同时对多模态数据进行理解和生成。为此，需研究多模态内容理解与生成任务统一建模，面向理解与生成任务相关的多种下游任务，包括语音识别、动作识别、表情识别、情绪识别、语音合成、视觉描述视觉问答、视觉内容生成、跨模态检索、情绪生成等具体应用，设计与多模态预训练模型尽可能兼容的多任务学习机制，使得各个任务的学习能够互相兼容、互为增强，各个任务可以从其他任务中所学到的技能和知识中收益，在提升模型性能和泛化能力的同时，赋予预训练模型更强的通用性。

(4)应用部署代价高

大规模预训练模型通常包含数以亿计的参数,海量的参数量使得预训练大模型在下游任务微调和推理解码时速度慢，对计算资源要求高，计算成本大，并且无法满足线上系统的实时性要求，导致无法在实际场景开展高效部署应用。因此，需研究面向应用部署的模型推

理加速、面向特定任务的模型泛化与迁移学习，实现高效的预训练学习算法，同时保障预训练模型泛化性和鲁棒性。

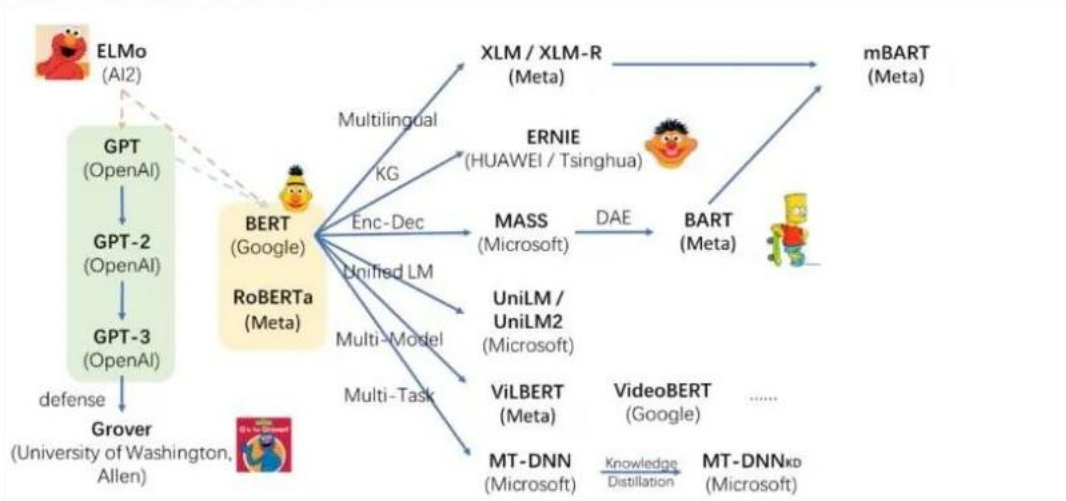
2 多模态大模型国内外发展现状

2.1 语言大模型

近年来,随着 ELMo、GPT-1、BERT、GPT-2 与 GPT-3 等预训练语言模型的发布，预训练技术这场革新正在自然语言处理领域悄然展开，并迅速影响到它的各个子领域之中。

顾名思义，预训练指的是使用通用性的任务和大规模的无标注数据进行第一阶段的训练,让机器学习模型学习到具有较强泛化性的参数。接下来，对于特定的下游任务，模型仅需对学习好的参数进行微小的调整(或训练)就能够完成高效迁移，达到显著的性能表现。

上述方法被我们称作“预训练+微调”，该范式逐步覆盖自然语言处理的各大任务并带来了显著的改进，如文本分类、阅读理解、序列标注和文本生成等。



预训练语言模型发展脉络图

预训练语言模型发展迅速，发展脉络如图所示。我们对近年来国内外的相关研究做了简单的分类，其中大部分研究可以被归纳到以下三类:

》预训练方法优化与改进;

》面向目标任务的预训练方法;

》超大规模模型高效训练框架(M6、华为、抖音);

2.2 视觉大模型

在过去的十年中，深度监督学习取得了巨大的成功。但是，由于它对手动标签的强依赖以及泛化性不足的缺点，促使人们寻求更好的解决方案。与此同时，在硬件快速增长的帮助下，今天的模型可以轻松地训练上百万张图像并开始尝试训练数亿张图像数据。然而，有监督方法对数据有着人工标注的要求，从而造成了获取成本过高，因此通过大规模有标签的数据来训练大规模通用预训练模型不现实。此外:即使是耗费大量人力物力获取标签,但是有标签的监督方法仍然会因为人为疏漏造就错误标签(即使是 ImageNet 这种高质量数据集也存在错误标签和混淆概念)。

图灵奖获得者 Yann LeCun 在演讲时表示，如果智能是一块蛋糕那么蛋糕的主体是无监督学习，蛋糕上的糖衣是监督学习，蛋糕上的樱桃是强化学习,而人类对世界的理解主要来自于大量未标记的信息。而同时不可忽视的是，无监督/自监督学习这类方法已经革新了自然语言处理的通用范式，如 BERT、GPT 系列在大规模语料上进行无监

督预训练，在各类下游任务中均取得了令人惊艳的效果。因此，无监督/自监督学习将是实现人类智能的关键，被广泛认为是通往通用人工智能的重要途径之一。

近些年自监督学习越来越受到广大研究人员的关注，其设计与思想天生就适合训练视觉大模型:利用大量的无标记数据训练模型构建通用的视觉表征，以此来使得所有类型的下游任务受益。

自监督学习常用方法是提出不同的上游任务(pretext task)网络可以通过学习上游任务的目标函数来训练，视觉特征也在这一过程中获得。如图所示，在自监督的上游任务训练阶段，自监督方法首先根据数据的某些属性自动生成该前置任务的伪标签，以此来训练神经网络获得预训练模型。在自监督的训练完成之后，可以将学习到的视觉特征迁移到下游任务(downstreamtask)，使用少量带标签的数据进行微调，以提高性能并克服过度拟合的情况。

自监督方法框架图

本章节，我们全面回顾了现有的经验方法，并根据代理目标的不同将其概括为四个主要类别:生成式、上下文式、对比式以及多任务式。我们将进一步研究相关的理论分析工作，以提供有关自监督学习如何工作的更深层次的思想。

2.3 语音大模型

语音大模型概况

随着语音技术的发展，在有标注的训练数据充足的情况下，使用深度神经网络训练语音识别等任务上已经能够取得非常好的性能;但是现实中有标注的训练数据获取的成本很高，在一些低资源的任务场景也很难获取足够多的有标注训练数据。基于此背景，近些年来研究者们致力于从大量没有标签的数据中预先学习有效的语音特征，使模型学习到语音更深层次的特征表达，从而在低资源的下游语音任务中摆脱对训练数据量的依赖，并且获取更好的性能。近些年来研究者们常用的语音预训练方法主要有以下几类:基于自回归重构自监督特征学习方法、掩码重构方法、对比预测编码方法、掩码预测方法、多任务学习方法等。

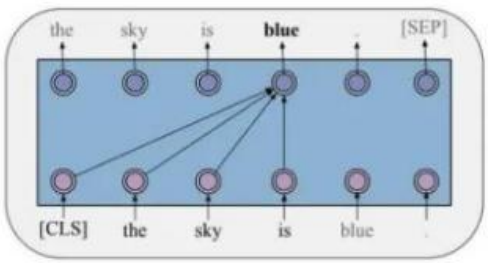
3 平治 AI 数字人的多模态大模型关键技术

3.1 预训练方法优化与改进

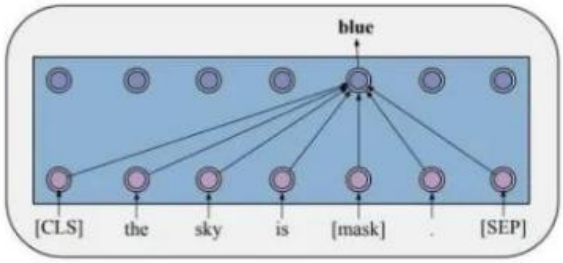
(1)自回归与自编码的预训练方法

自回归方法是传统条件语言模型常用的训练方法，例如 GPT，它旨在训练模型根据给定的上文预测当前词汇，如图所示，模型根据语句部分“theskyis”来预测即将出现的词汇“blue”。对于句子 x，自回归模型按照最大化对数似然损失的方式进行参数优化，公式如下:

$$L(x)=\sum_{i=1}^{n+1}\log P(x_i \mid x_{i-k},\cdots,x_{i-1};\theta)$$



自回归预训练示意图



自编码预训练示意图

自编码的预训练方法代表性的工作是掩码语言模型，如 BERT. 简单来说，模型需要通过对于遮盖数据的预测进行参数优化。如图所示，我们将输入中“blue”用掩码[mask]代替，让模型在输出层预测出“blue”

对于句子 x，假设其中存在 m 个需要预测的词汇，那么其损失

函数的形式化表示可以写作:

$$L(x)=\sum_{i=1}^{m}\log P([\text{mask}]_i=y_i\,|\,\hat{x},\boldsymbol{\theta})$$

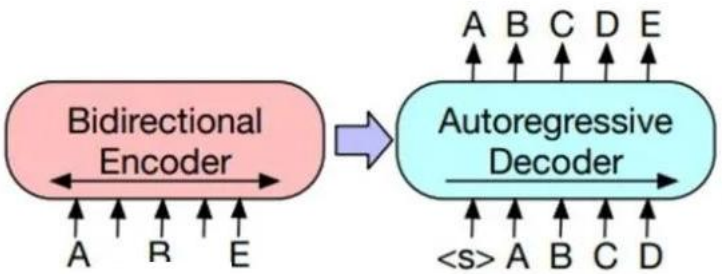
自回归预训练模型擅长于文本生成，能够进行文本创作和新闻撰写等;自编码预训练模型侧重于文本理解，能够迁移到多种下游任务中获取强大的性能改进。在这两类预训练方法的基础上，各种不同的模型相继被提出，预训练的研究如雨后春笋般发展起来。

(2)自回归与自编码的统一框架

自回归和自编码两类预训练方法各有优劣，大量的研究讨论如何将他们的优势进行统一。

在各种改进的预训练方法中，排列语言模型和序列到序列的方法影响最大，它们间接地将自回归和自编码统一在同一个框架中，同时发挥文本生成和文本理解的作用。

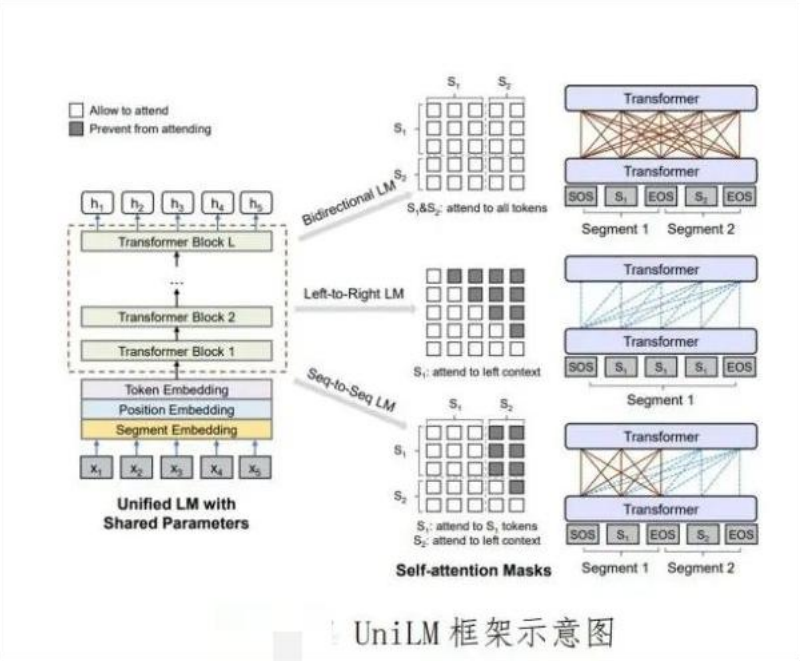
XLNet 是一类排列语言模型。它打乱文本序列的顺序，但仍然预测原始文本中的下一个词汇。通过这种方式，XLNet 可以同时利用左侧和右侧文本的信息，模拟自编码的形式实现自回归的任务，在一系列下游任务上取得了显著的性能改进。



自回归训练模型

自回归训练方法虽然可以进行文本生成，但是没有对于给定前缀和生成内容进行区分。序列到序列的预训练方法应运而生，这类方法对于条件生成任务非常有效。常见的模型如基于掩码的序列到序列预训练框架 MASS、基于序列到序列模型的降噪自编码器 BART 等:MASS 能够对编码器和解码器进行联合训练提高特征抽取和语言模型的表征能力，对基于少量样本的语言生成任务，如机器翻译、文本摘要和对话生成等，进行参数微调能够获得较好的性能。BART 通过先损坏文档再解码恢复的策略实现从损坏文档到原始文档之间的鲁棒映射。

另一类方法使用注意力机制将理解和生成任务结合起来，如 UniLM 等。对于理解任务，模型打开双向注意力，同时关注上文和下文;对于生成任务，模型在计算注意力分布的时候将后续文本进行遮盖;当然，模型也能够结合二者来实现条件生成任务，如图所示。



UniLM 框架示意图

(3)预训练的经验性训练方法

除了基本训练方法和模型结构的设计，研究者们也开始探讨如何设置训练的参数来达到更好的效果。

Meta 的研究者通过调整 BERT 训练过程中的批处理大小和学习率发现，更大的批样本能够显著改进预训练的效果，发布了 RoBERTa 模型华为基于 BERT 和 GPT 提出 NEZHA 模型，提供了更加丰富的训练技巧;百度提出的 ERNIE1.0 和 ERNIE2.0，在 BERT 的基础上进行了全词掩码，在中文任务上取得了更好的效果;ZEN 模型融入了 n-gram 信息,取得了相比于 BERT 模型更好的收敛速度和性能表现;MacBERT 则使用相似词替换[MASK]字符，以缓和预训练和下游任务的不一致性。

这些研究表明，预训练语言模型的训练不仅是科学研究问题，也是工程问题,因此如何更好地设计训练方法发挥模型最大的效用也是工业界主要探索的问题之一。

3.2 面向目标任务的预训练语言模型

(1)多语言融合的预训练模型

融合多语言的预训练模型旨在使用同一套参数处理多种不同的语言文本。该方法可以捕捉不同语言之间的相似性规律，从而达到互相帮助的效果。

常见的多语言预训练模型往往包含两种，一种是仅使用单语数据训练，让模型自动捕捉不同语言之间的相似性规律，如 XLM-MLM 模型等;另一种则同时使用单语数据和平行数据，充分发挥训练语料的能力，如 XLM-TLM 模型等。

多语言融合的预训练模型往往能够在跨语言任务上取得不错的性能表现，如机器翻译、跨语言问答等，也具备很好的语言迁移能力!但是,该类模型的训练方法与单语训练方法基本保持一致,我们认为如何根据语言之间的规律设计训练方法是该领域未来研究的核心。

(2)融入外部知识的预训练模型

预训练语言模型通过对大规模文本数据统计规律的捕捉，蕴含了一定程度知识信息，如语言学知识和事实性知识等。但是，这种知识仍然存在噪声，模型仍需要具备见多识广的能力。

一些工作希望将知识图谱中的事实三元组信息融入到预训练模型中，清华和华为提出 ERNIE 模型，使用额外的知识融合层次将知识信息在预训练阶段注入到模型之中。当然，也有研究探讨直接在推理阶段使用知识对于文本表示进行增让，如 K-BERT，它使用树状结构约束自注意力分布，减少结构化知识影响的同时，增强文本表示的知识

性。

除了将外部知识融入预训练之中，将预训练本身蕴含的知识进行抽取也是当前研究的热点之一，常见的方法如基于提示的方法，虽然可以抽取到一定程度的事实性知识，但是仍然存在较多噪声，如何将噪声去除也是当前该方向上亟待探索的关键问题之一。

(3)面向对话生成的预训练模型

人机对话模型旨在让机器理解人类语言，并根据人类语言的意图执行特定任务或做出相应回答，从而实现与人类进行自由交流与沟通的目标。目前对话模型大致可分为两类，一类为任务型对话模型，例如智能客服、问答系统等,主要是针对特定领域内具体任务而设计的;另一类是开放域对话模型，以达到人与人之间自由、无约束的闲聊水平为目的。随着深度学习与神经网络的不断发展，端到端的对话模型已经成为主流的设计方案。

端到端的对话模型一般由编码器和解码器两部分构成的,编码器接收历史信息并对其进行编码，解码器根据编码结果生成对应的回复。模型能够通过编码器学习到语言信息的良好表征，并从中获取语义等更深层次的信息，从而在后续的解码阶段模型才能够生成语义通顺、逻辑一致的高质量回复。此外，为了学习到能够美人类的语言表征和理解能力，对话模型往往有着大规模的参数且在海量的人类对话数据上进行训练。近几年来，国内外在对话模型方面的研究也是成果斐然，越来越多大规模、性能卓越的对话模型的发布也是不断刷新了人机对话的各类评价指标。

百度的 PLATO 系列模型在全球独树一帜，最新发布的 PLATO-XL 也成为了全球最大的对话生成模型。2019 年百度发布了通用领域的对话模型 PLATO，该模型首次提出了将离散隐变量与 Transformer 的编码器-解码器结构相结合，离散隐变量的每一个取值都与一个回复意图相对应，从而对上文历史信息与回复之间“一对多”的关系进行有效的建模。在编码过程中使用双向的编码器，达到充分理解上下文信息的目的，在解码过程中使用单向的解码器，结合隐变量对应的意图自回归式地生成合理的回复。2020 年百度在 PLATO 的基础上进一步扩展优化提出了 PLATO-2,该模型通过扩展网络层数，增加训练集数据，将模型增加到 16 亿参数。此外，PLAT0-2 将具体训练过程分为两个阶段:第一阶段，在一对一映射的简化框架下，训练粗粒度生成模型来学习回复生成;第二阶段，进一步训练细粒度生成模型和评估模型，其中细粒度生成模型显式地建模一对多关系，评估模型则用来学习回复的一致性从而选择最合适的回复。PLATO-XI 作为 2021 年发布的全球首个百亿参数的模型，也在 PLATO、PLATO-2 模型的基础上进一步增强了模型对话理解和回复生成的能力。由于训练语料?大多是社交媒体对话，涉及多个角色且对话话题较广，因此模型难以区分历史信息中不同角色的观点与信息，从而会产生前后不一致的回复。针对此类问题，PLATO-XL 进行了多角色感知的预训练，以生成更加流程一致的回复。PLATO-XL 凭借其千亿级的训练语料和百亿级的参数规模已经在各类评估指标上显著超越了目前主流的对话模型。

Facebook 提出的 Blender 对话模型具有仅次于 PLATO-XL 的 94 亿

参数，是一个强大的综合人工智能聊天机器人框架。该模型结合了三个子模型:检索模型、生成模型、检索+生成模型。检索模型以对话历史信息为输入，在候选集中对每一个响应进行评分，并从中选择最合适的作为回复;生成模型采用经典的端到端的 Transformer 结构根据历史信息直接生成回复;检索+生成模型采用了“检索和提炼“的方式，首先检索出候选的回复，再将该候选传入生成器中作为参考，进一步产生更加高质量的回复。Blender 也是凭借上述模型提供的先进的解码生成策略和混合技能，集移情、知识、个性于一体，在各领域内都达到了能与人自然交互对话的水平。

尽管现有的对话模型已经达到了媲美人类的对话水平，但是依旧日存在不少问题亟待解决:

1)模型会经常重复对方的说话内容，产生较为普遍的迎合式或者附和式回复，缺乏个性化、新颖的对话内容;

2)模型无法记住所有的历史信息，也无法根据对话内容建立逻辑上的联系，因此很容易产生自相矛盾、前后不一致的情况;

3)模型缺乏对知识和客观事实的理解，除非针对特定领域精心挑选训练数据，否则模型很容易产生与客观事实相悖等错误。

想要完美解决上述问题是十分困难的,如何构建一个真正像人类一样智能的对话模型仍然是当今人工智能领域的一大挑战。

(4)预训练语言模型的蒸馏与压缩

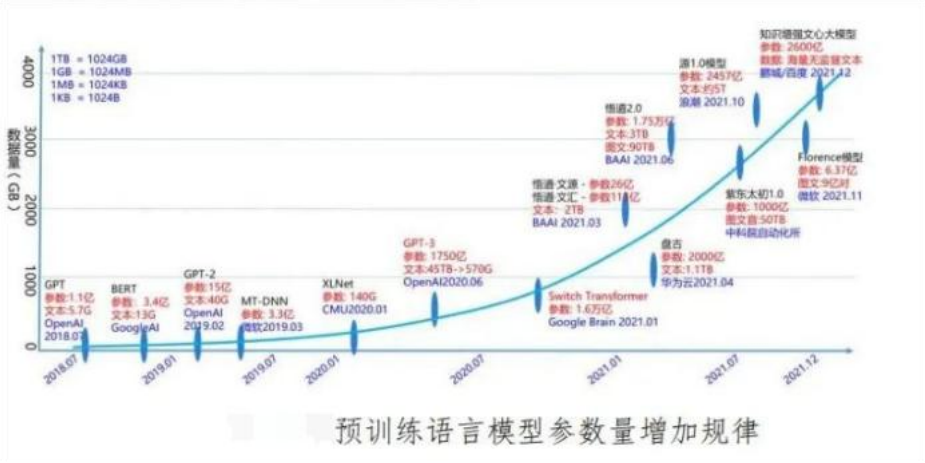
预训练语言模型虽然在自然语言任务中取得了很好的效果，但是这类模型参数量较大，难以满足实际应用中的时间和空间需求。为了

缓解该问题,研究者们开始讨论使用知识蒸馏的方法来进行预训练模型压缩和加速，代表性工作如 DisilBERT、MobileBert，在原有的损失基础上增加了蒸馏损失项和词向量余弦损失，让小模型也具备大模型的能力。

除了基础的蒸馏方法,现在的工作也开始探讨针对特定任务的蒸馏、渐进式的知识迁移等，更大程度上保留模型原本能力的同时减小体积、提升解码速度。

2.1.3 超大规模预训练模型的训练方法

(1)超大规模预训练模型的发展现状



预训练语言模型参数量增加规律

近年来，预训练语言模型的发展呈指数型增强，参数规模的增长也呈现出这种规律。GPT-3 是首个发布的超大规模语言模型，使用自回归的方法和超大规模的数据进行训练，呈现出了强大的通用性和少样本学习的能力，为通用人工智能的实现打开了一个窗口。

在国内，清华大学和智源研究院合作发布悟道大模型，是中文超大规模预训练的排头兵;华为云由底层向上逐步研发，开源了盘古大模型;中科院自动化研究所提出千亿规模的多模态预训练模型，应用

场景广泛。参数量的增加不仅显著提升了模型通用能力，也彰显了中国人工智能发展的速度和水平。

(2)超大规模预训练模型的训练方法

超大规模的预训练模型往往具备参数量巨大的特点，如 GPT-3 足足有 1750 亿参数,这样的体量和我们平时使用的模型有着天壤之别。因此，大规模预训练往往需要大量的计算资源，采取模型并行和数据并行技术。

除了惊人的性能表现,超大规模预训练模型也带来了显著的计算开销，训练时间也大幅度增加。因此，研究者开始从减少模型复杂度的角度进行设计，在增大模型参数的同时减小其复杂程度，如稀疏注意力、混合专家模型(MixtureofExpert)等。最近，混合专家型受到了广泛的关注。一方面，它能够显著增加参数但基本不改变计算复杂度，另一方面，它能够让模型自主学习稀疏通路从而完成多种任务。但是每个专家模型处理样本的数量不均衡的问题仍然存在，如何改善这个问题是未来大模型训练亟待解决的问题之一。

(3)超大规模预训练模型的能力探索

由于超大规模预训练模型本身大量的参数难以进行全参数的微调，因此研究者们主要从两个角度来探索大模型的能力:一种方法是利用语言模型的能力来进行零样本和少样本的学习，另一种方法是使用参数高效的微调方法。

1)零样本和少样本学习能力探索

GPT-3 模型采用零样本或者少样本学习的方法进行能力探索，如

下图所示，零样本指的是给定 GPT-3 任务描述，让模型自主生成目标内容;少样本则是给定任务描述之后，再给出若干样例帮助模型确定任务和目标输出形式。研究者发现，超大模型具备强大的少样本学习能力，在一些任务上甚至可以达到最佳的性能表现。

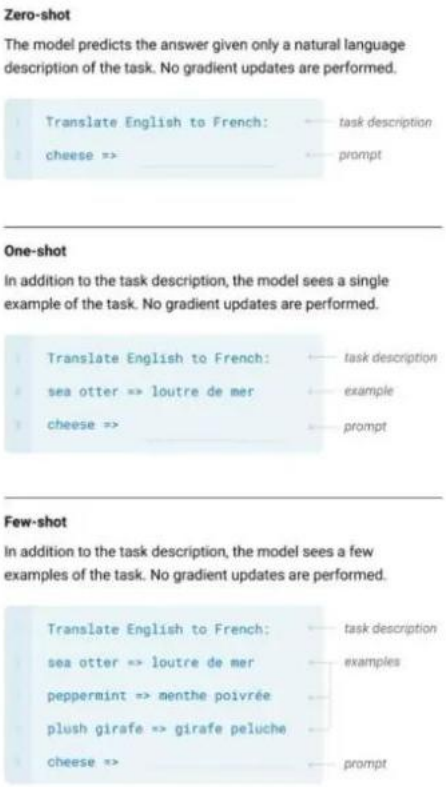


图 零样本和少样本学习的样例

2)参数高效的微调方法

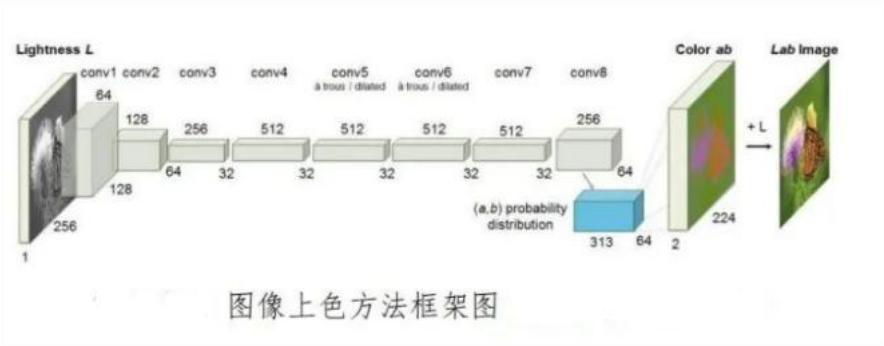
由于大规模预训练模型无法进行全参调整,研究者们开始探索参数高效的微调方法，常见的如基于提示(Prompt)学习的方法和基于适配器(Adapter)的方法。这两种思路均在原有模型的基础上增加少量可学习的参数,在大模型上仅调整这部分参数就可以达到与微调模型相当的效果。另一方面，研究者们也开始思考模型本身是否可以进行拆解或组合，来达到高效微调的目的。

我们认为，充分发挥大模型的能力是一个非常重要的问题，这个方向上仍然具备着巨大的探索空间。

3.3 视觉大模型生成式方法

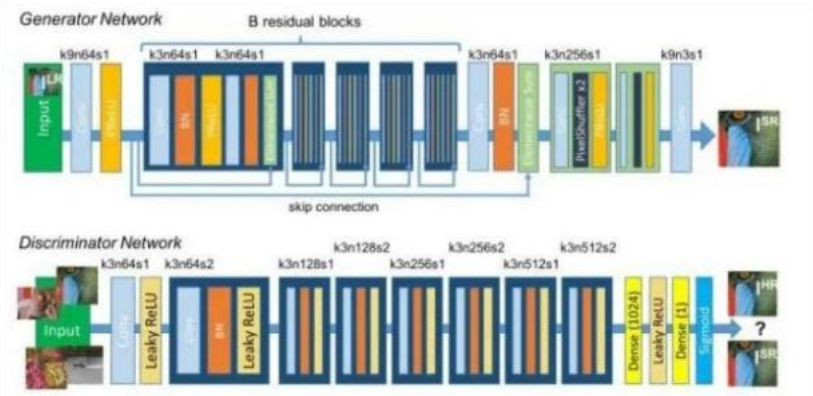
生成式方法的典型代表如图像上色、图像超分辨率等，方法各式各样，但本质都是生成式方法。

图像上色:图像上色任务是将彩色图像转化为灰度图像，此灰色图像通过神经网络，并使得上游训练任务为预测原本的彩色图像逼迫网络来学习图像的结构和上下文信息，框架图如图所示。



图像上色方法框架图

图像超分辨率:图像超分辨率任务是将输入图像的尺寸缩小，并使得缩小后的图像通过神经网络，并使得上游训练任务为预测原本的图像，是以此生成对抗的思想逼迫网络来学习图像的结构和上下文信息，框架图如图所示。

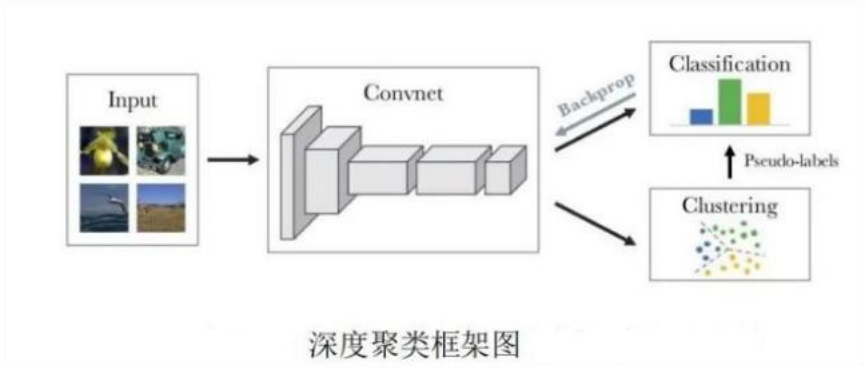


图像超分辨率方法框架图

这些方法是视觉自监督方法最早的启蒙思想,思想的本质是损坏输入图像的部分特征，让神经网络重建回原本的特征。在早期的领域中有着重要的贡献，然而正是因此，这些方法也有很多缺陷，如无法从图像中提取多粒度的特征，仅在某个或某几个任务中有效，很难泛化到各类下游任务。

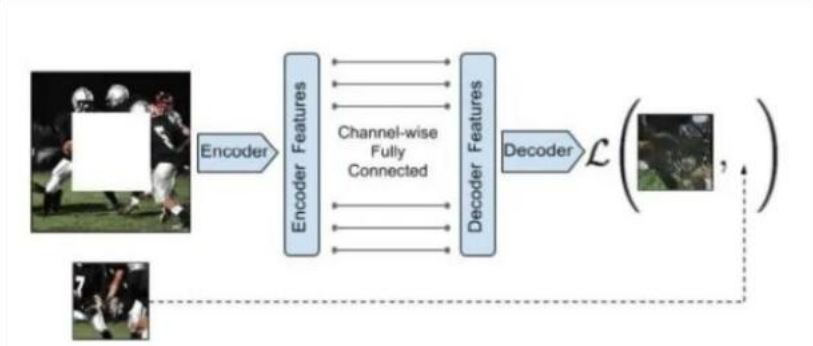
3.4 视觉大模型上下文式方法

基于上下文的上游任务的设计主要利用图像的上下文特征，如上下文相似性、空间结构等。基于上下文的相似性的典型方法是根据图像的上下文相似性设计的，这种类型的方法主要是图像深度聚类的方法。基于空间结构任务用于训练基于图像块之间的空间关系的神经网络，这种方法以上下文编码器思想的方法流行,深度图像聚类:深度图像聚类是将聚类与深度结合的方法，这种方法可以学习到一些有用的通用特征，这个框架如图所示整个过程包含对特征进行聚类，然后基于聚类的结果作为伪标签，更新网络的参数，让网络预测这些伪标签，这两个过程依次进行。



深度聚类框架图

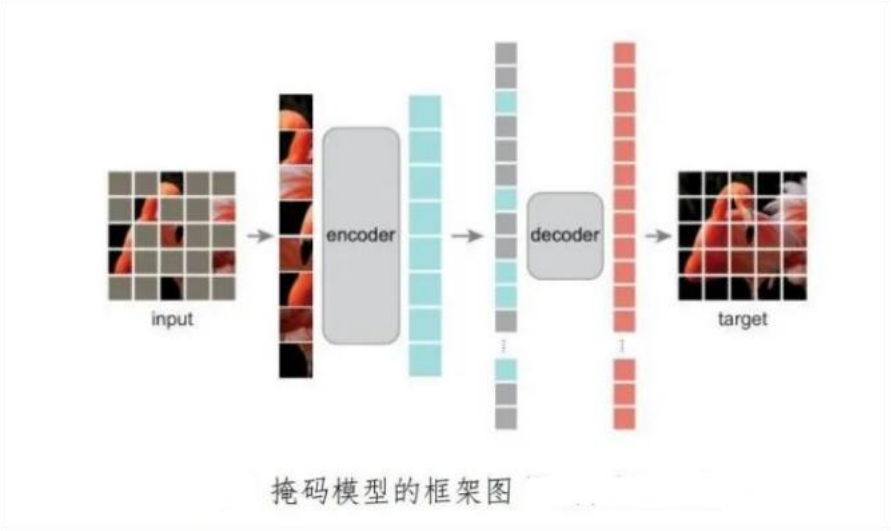
上下文编码器:上下文编码器将图像中的随机区域丢弃，丢弃填补 0 像素值，之后利用卷积神经网络的编码-解码结构和图像先天就具备的上下文信息来学习恢复被丢弃的随机区域，如图所示。在训练完成后，将编码-解码模块的部分参数作为预训练模型应用于其他的任务。然而，基于卷积神经网络的上下文编码器方法并没有取得让人印象深刻的性能。但是这种思想却成为自然语言处理自监督学习的标准范式，说明了上下文式方法的极大潜力。



上下文编码器框架图

掩码模型:掩码模型也是视觉上下文编码器的一种，但是却取得了极大的突破，其成功基于两方面:1).视觉 Transformer 的提出，为计算机视觉和自然语言处理的预训练统一奠定了基础;2).分析了视觉信息和语言信息的不同，证明了视觉信息具备更冗余的特性，以此提出更进一步的上下文编码模型。掩码模型是自然语言处理预训练一种流

行的标准范式，在视觉中采用此种预训练方式有助于统一不同模态的预训练方法，发展通用的人工智能大模型。掩码模型中的代表 MAE 的框架图如图所示。图像经过线性层映射成词条(token)集合，并被随机掩码。将没有被掩码的 token 输入进编码器与此同时解码器是轻量级的，并且从经过编码器输出的潜在表示和被掩码的 token 中重建输入。此种训练方法可以提取较好的表征，适合在密集预测的下游任务中使用。此外微软亚研院提出的 BEiTSimMIM 也是此类方法中的一种。



掩码模型的框架图

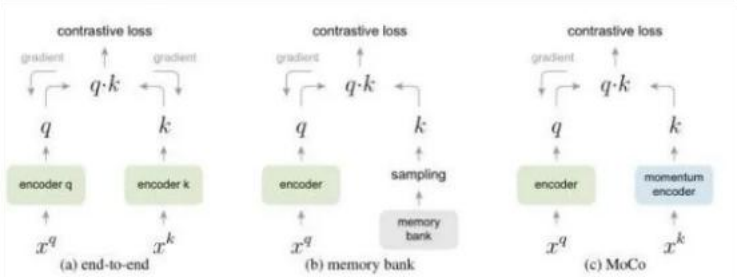
这些方法虽然在密集预测的下游任务上表现效果好，然而在流行的线性评估方面效果较差，同时有训练效率低的问题。

(1)对比学习方法

对比学习是近年来自监督研究社区最热门的研究方向，其建立在语义一致性的假设之上:对于同一图像的不同视角(叫做正样本，通常由数据增广获得)，网络应该提取相似的特征，对于不同图像(叫做负样本，从数据集中重采样获得)网络提取的特征要尽量远离。

此类方法典型代表为 MoCo 系列、SimCLR 系列、SimSiam、BYOL 等。

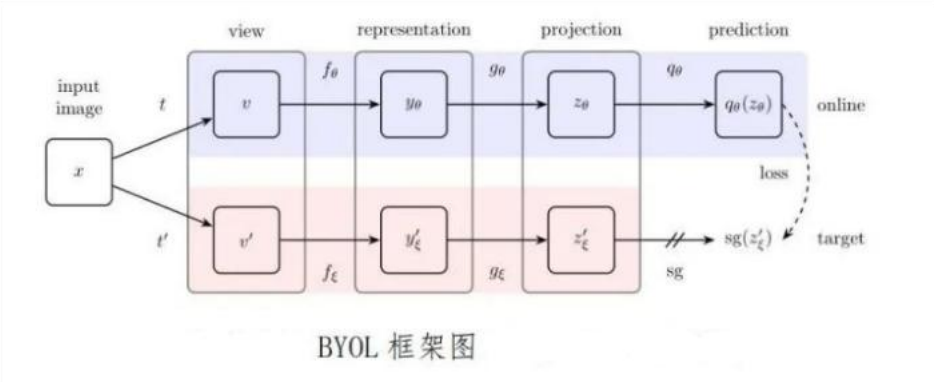
MoCo: 此方法基于对比学习的基础上提出了记忆池(memorybank)，该方法使用一个在内存中的 memory bank 保存以往样本的特征，并且动量网络在训练过程中会以滑动平均的方式更新，网络中直接输入的样本均为正样本，负样本从记忆池中取，所以不需要很大的批次大小，如图所示。



MoCo 框架图

SimCLR:此方法基于对比学习的基础上，将同一个批次(batch)中的其他样本视为负样本，所以需要非常大的批次大小(批次大小)，使用非常多的 GPU 来训练网络。

BYOL:提出一种无须负样本就能学习的自监督方法，本质还是对比学习的思想，即对于同一图像的不同视角仍然是一个类别。BYOL 训练框架如图所示。



BYOL 框架图

由上可知,对比学习的成功是建立在语义一致性假设的基础之上的，公开的单目标数据集 ImageNet 保证了这种假设，然而若是使用

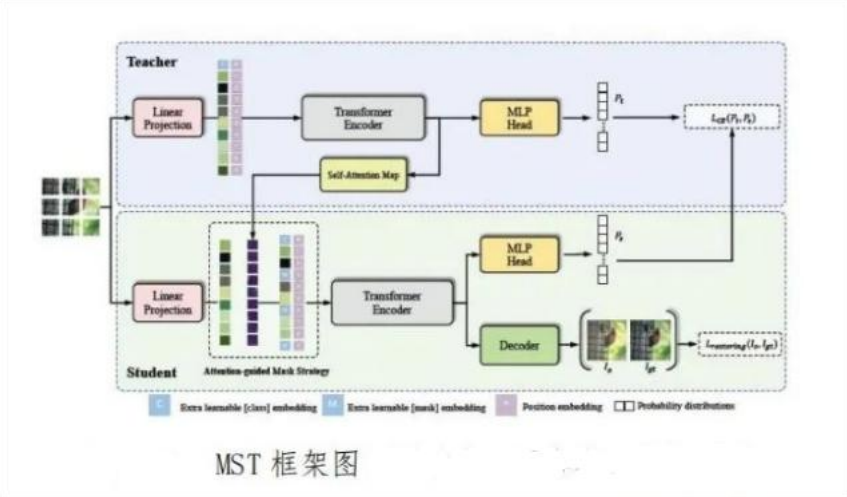
多目标数据集训练神经网络，则会不符合假设，无法提取图像中有用的特征。因此对比学习方法目前不适合在现实场景中训练，同样地，这些方法存在训练效率低的问题。

(2)多任务式方法

上述的 3 类方法都是以一种简单的方式预训练神经网络，然而若是训练一个通用的大模型，简单的任务容易使得模型过拟合当前任务，造成迁移性能变差。因此构建一个通用的上游任务使得网络避免过拟合特定任务,提升网络在下游任务的迁移性能,也是目前的研究方向。多任务学习就是组合多个自监督任务，提升网络表征的方法。典型代表方法:DeepMind 提出的多任务学习框架，中科院自动化所提出的 MST,以及字节跳动的 iBoT。

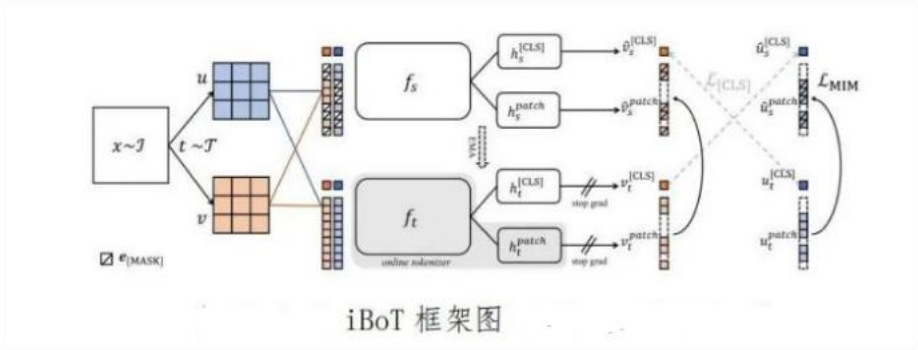
DeepMind 提出的方法由 4 个自监督任务组成:(a)切片的相对位置预测;(b)图像上色;(c)对比学习;(d)预测视频中下一帧的像素变化。通过组合多种自监督任务来学习通用的特征表达，证明了多任务学习的优越性。

中科院自动化所提出的 MST 采用了掩码模型和对比学习的方式。在对比学习的基础上，提出了一种注意力引导的掩码策略，利用教师模型产生的注意力指导学生模型掩码，并通过一个全局的解码器对图像进行重建，学习语义特征的同时保留图像的空间分辨率。整体框架如图所示。



MST 框架图

此外，字节跳动也提出了掩码学习和对比学习结合起来的多任务学习工作 iBoT，参考 MST 利用教师模型产生的信息给予学生模型以指导，其对学生模型的 token 进行掩码，并利用教师模型的完整 token 当做标签信息进行训练学习。框架图如图所示。



iBoT 框架图

多任务自监督学习涉及到了多个任务间的兼容、系数调整等问题当前对于多任务自监督算法的探究还非常少。

3.5 语音大模型基于自回归重构自监督特征学习方法 (Autoregressive Reconstruction)

自回归预测编码是一种生成式模型，即给定一段输入序列，来预测未来时段的频谱，通常使用 L1 损失函数。它构建了一种自我监督

目标函数,能够从海量无标签数据集中学习丰富的特征表达并应用于下游任务。2020 年麻省理工学院(MIT)人工智能实验室的 Yu-An Chung 等人提出了一种生成式的自监督语音特征学习方法 APC(Autoregre-ssive Predictive Coding)，该方法受到文本生成模型的启发,主要通过对未来时刻的语音帧特征进行预测，从语音浅层的表达方式中学习到更高层次的语音特征。同年 MIT 人工智能实验室又在 APC 自回归编码器的基础上提出了 APC 的离散化进阶版本矢量量化的自回归预测编码 VQ-APC(Vector-Quantized Autoregressive PredictiveCoding)，通过在自回归层之间添加额外的矢量量化网络层，实现了从连续表征到离散化表征的转换。

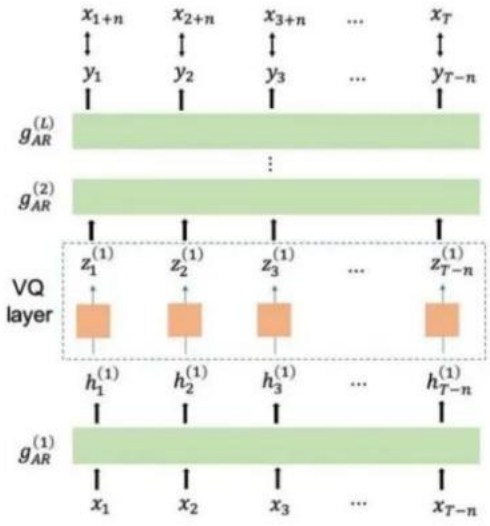
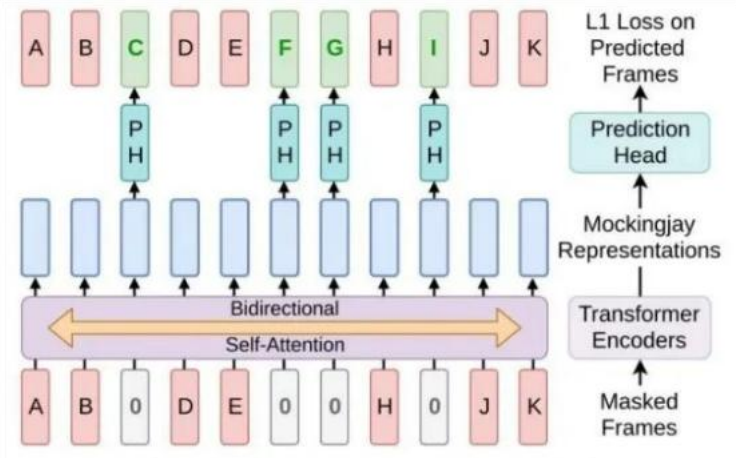


图　矢量量化的自回归预测编码框架示意图

3.6 掩码重构(Mask Reconstruction)

2019 年台湾大学 Andy T.Liu 等人提出了一种新的语音表示学习方法 Mockingjay，该方法使用双向的 Transformer Encoder 在大规模无

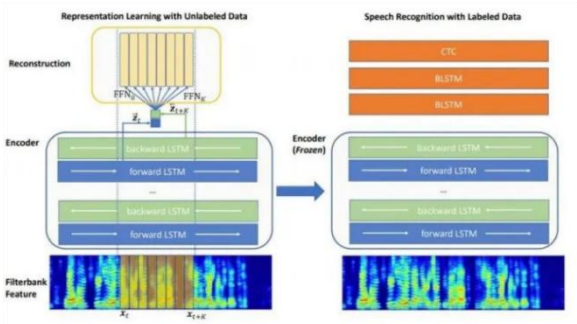
标签语音数据集上进行预训练，该方法通过掩码(mask)的方式，根据过去和未来的帧来共同预测当前的帧。通过这个方法得到的语音表示广泛地提高了下游任务的表现，例如音素分类、语音识别、基于语音的情感分析等。除此之外，实验证明用 Mockingjay 方法进行预训练，然后再用其他下游任务进行微调，能够极大得提高表现;在 Mockingjay 模型的基础上，该团队又提出了两种进阶优化版的模型 Audio Albert 与 TERA，其中 Audio Albert 是在 Mockingjay 的基础上将多层 Transformer 都参数共享,通过这样的改进方式有效降低了模型需要训练的参数量,且性能能够和 Mockingjay 保持一致;另一个 Mockingjay 的改进版 TERA 主要在输入端做了一系列改进,分别在时间维度和特征维度上都进行了掩码，此外还对整个输入增加了高斯白噪声，这一系列改进进一步提升了预训练模型的效果。



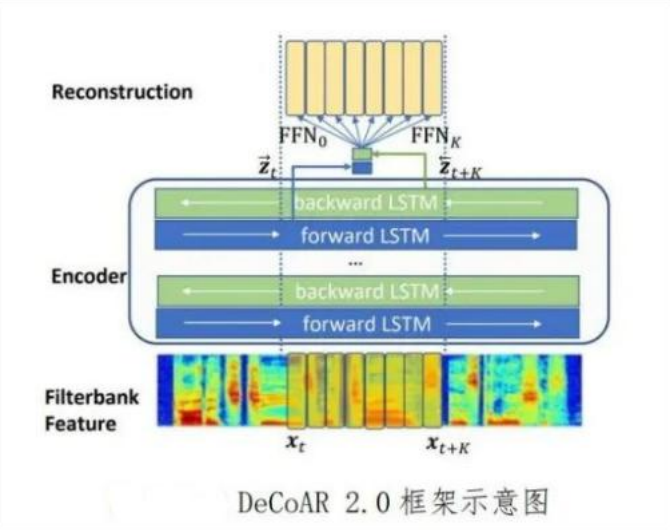
Mockingjay 框架示意图

2019 年 Amazon 团队 Shaoshi Ling 等人也提出基于半监督上下文学习的语音预训练模型 DeCOAR，如图所示，DeCOAR 结合了 ELMo 的双向性和 APC 的重构目标,首先在编码网络,根据过去和将来的帧重建特征，得到的特征再送入基于 CTC 的说话人识别网络。相比

Mockingjay 使用 Transformer 作为编码网络，DeCOAR 使用双向 LSTM 作为编码网络。2020 年该团队提出了优化版本的 DeCoAR2.0 模型,如图所示,将原先的 LSTM 网络替换成了 Transformer 网络，并且加入了一个 VQ 模块学习离散化的特征，在下游任务上取得了更好的性能。

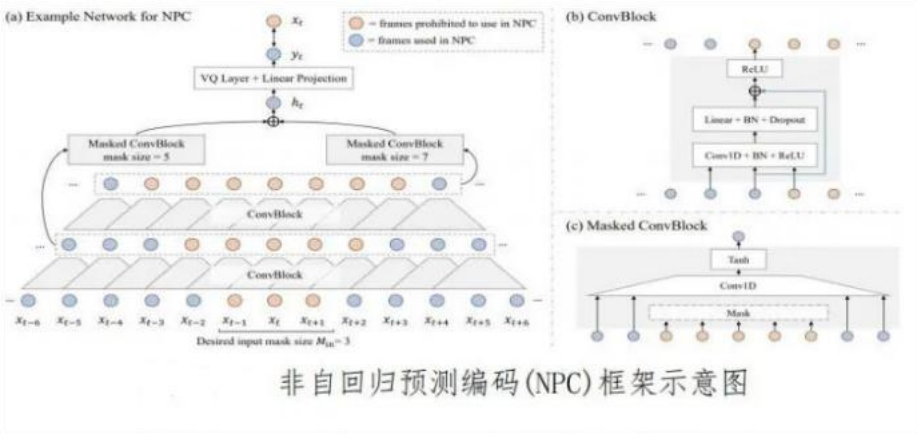


DeCoAR 框架示意图



DeCoAR 2.0 框架示意图

除了上述介绍的一系列基于掩码重构的预训练方法，AlexanderH.Liu 等人提出一种使用两侧帧的信息预测中间帧信息的预训练方法,非自回归预测编码 NPC(Non-Autoregressive Predictive Coding),通过感受野限制信息的前向传递过程来确保重建过程只依赖于目标语音的周围信息。NPC 与前述方法的区别是模型的输入为被掩码(mask)的帧前后的帧，从而可以进一步提升模型的速度。



非自回归预测编码(NPC)框架示意图