

目 录

1 大模型全家桶	1	1.7 百川大模型	46
1.1 多模态大模型基本概念	3	1.7.1 预训练	47
1.1.1 多模态	4	1.7.2 对齐	51
1.1.2 大模型和基础模型	4	1.8 本章小结	52
1.1.3 多模态大模型	5	2 多模态大模型核心技术	53
1.2 BERT 技术详解	6	2.1 预训练基础模型	54
1.2.1 模型结构	6	2.1.1 基本结构	55
1.2.2 预训练任务	10	2.1.2 学习机制	56
1.2.3 下游应用场景	13	2.2 预训练任务概述	58
1.3 ViT 技术详解	14	2.2.1 自然语言处理领域的预训练任务	58
1.3.1 模型结构	15	2.2.2 计算机视觉领域的预训练任务	58
1.3.2 预训练任务	17	2.3 基于自然语言处理的预训练关键技术	59
1.4 GPT 系列	19	2.3.1 单词表征方法	60
1.4.1 GPT-1 结构详解	20	2.3.2 模型结构设计方法	62
1.4.2 GPT-2 结构详解	23	2.3.3 掩码设计方法	62
1.4.3 GPT-3 结构详解	24	2.3.4 提升方法	63
1.5 ChatGPT 简介	28	2.3.5 指令对齐方法	64
1.5.1 InstructGPT	28	2.4 基于计算机视觉的预训练关键技术	66
1.5.2 ChatGPT	32	2.4.1 特定代理任务的学习	67
1.5.3 多模态 GPT-4V	37	2.4.2 帧序列学习	67
1.6 中英双语对话机器人 ChatGLM	40	2.4.3 生成式学习	68
1.6.1 ChatGLM-6B 模型	41	2.4.4 重建式学习	69
1.6.2 千亿基座模型 GLM-130B 的结构	43		

2.4.5	记忆池式学习	70	3.1.2	选择有效的预训练方法	122
2.4.6	共享式学习	71	3.1.3	选择和扩展模型	123
2.4.7	聚类式学习	73	3.1.4	预训练	123
2.5	提示学习	74	3.2	BLIP	124
2.5.1	提示的定义	75	3.2.1	模型结构	124
2.5.2	提示模板工程	77	3.2.2	预训练目标函数	125
2.5.3	提示答案工程	80	3.2.3	标注过滤	126
2.5.4	多提示学习方法	81	3.3	BLIP-2	127
2.6	上下文学习	84	3.3.1	模型结构	128
2.6.1	上下文学习的定义	85	3.3.2	使用冻结的图像编码器进行视觉与语言表示学习	128
2.6.2	模型预热	85	3.3.3	使用冻结的 LLM 进行从视觉到语言的生成学习	129
2.6.3	演示设计	87	3.3.4	模型预训练	130
2.6.4	评分函数	89	3.4	LLaMA	131
2.7	微调	90	3.4.1	预训练数据	131
2.7.1	适配器微调	91	3.4.2	网络结构	132
2.7.2	任务导向微调	94	3.4.3	优化器	133
2.8	思维链	97	3.4.4	高效实现	133
2.8.1	思维链的技术细节	98	3.5	LLaMA-Adapter	133
2.8.2	基于自洽性的思维链	99	3.5.1	LLaMA-Adapter 的技术细节	135
2.8.3	思维树	102	3.5.2	LLaMA-Adapter V2	136
2.8.4	思维图	105	3.6	VideoChat	139
2.9	RLHF	109	3.6.1	VideoChat-Text	141
2.9.1	RLHF 技术分解	110	3.6.2	VideoChat-Embed	142
2.9.2	RLHF 开源工具集	113	3.7	SAM	145
2.9.3	RLHF 的未来挑战	114	3.7.1	SAM 任务	148
2.10	RLAIF	114	3.7.2	SAM 的视觉模型结构	149
2.10.1	LLM 的偏好标签化	115	3.7.3	SAM 的数据引擎	150
2.10.2	关键技术路线	117	3.7.4	SAM 的数据集	151
2.10.3	评测	117	3.8	PaLM-E	152
2.11	本章小结	118	3.8.1	模型结构	154
3	多模态基础模型	119			
3.1	CLIP	121			
3.1.1	创建足够大的数据集	121			

目 录

3.8.2	不同传感器模态的输入与 场景表示	156	5.1.2	模型设计准则模糊	231
3.8.3	训练策略	157	5.1.3	多模态对齐不佳	232
3.9	本章小结	157	5.1.4	领域专业化不足	232
4	多模态大模型的应用	158	5.1.5	幻觉问题	234
4.1	视觉问答	158	5.1.6	鲁棒性威胁	234
4.1.1	视觉问答的类型	159	5.1.7	可信性问题	236
4.1.2	图像问答	160	5.1.8	可解释性和推理能力问题	240
4.1.3	视频问答	177	5.2	因果推理	244
4.1.4	未来研究方向	188	5.2.1	因果推理的基本概念	245
4.2	AIGC	189	5.2.2	因果的类型	249
4.2.1	GAN 和扩散模型	190	5.2.3	LLM 的因果推理能力	250
4.2.2	文本生成	192	5.2.4	LLM 和因果发现的关系	252
4.2.3	图像生成	196	5.2.5	多模态因果开源框架 CausalVLR	253
4.2.4	视频生成	201	5.3	世界模型	255
4.2.5	三维数据生成	202	5.3.1	世界模型的概念	256
4.2.6	HCP-Diffusion 统一代码 框架	202	5.3.2	联合嵌入预测结构	259
4.2.7	挑战与展望	207	5.3.3	Dynalang: 利用语言预测 未来	262
4.3	具身智能	207	5.3.4	交互式现实世界模拟器	264
4.3.1	具身智能的概念	208	5.3.5	Sora: 模拟世界的视频生成 模型	265
4.3.2	具身智能模拟器	210	5.4	超级智能体 AGI Agent	269
4.3.3	视觉探索	214	5.4.1	Agent 的定义	270
4.3.4	视觉导航	217	5.4.2	Agent 的核心组件	272
4.3.5	具身问答	221	5.4.3	典型的 AGI Agent 模型	273
4.3.6	具身交互	223	5.4.4	AGI Agent 的未来展望	282
4.3.7	存在的挑战	226	5.5	基于 Agent 的具身智能	284
4.4	本章小结	229	5.5.1	具身决策评测集	285
5	多模态大模型迈向 AGI	230	5.5.2	具身知识与世界模型嵌入	285
5.1	研究挑战	231	5.5.3	具身机器人任务规划与 控制	287
5.1.1	缺乏评估准则	231	5.6	本章小结	294