《多模态大模型原理与应用》

Lecture 7
多模态基础模型

刘阳
中山大学
人机物智能融合实验室（HCP Lab）
liuy856@mail.sysu.edu.cn

多模态基础模型概述

- 基础模型定义（Bommasani 等，Stanford HAI）
  - 在自监督或半监督方式下训练、基于大规模数据的基础模型，可用于多个下游任务
  - 一次训练即可快速适应多种应用，具有广泛泛化与迁移能力
- 多模态基础模型（Multimodal Foundation Model）
  - 能同时理解图像、文本、视频、音频等多种模态
  - 无需重新训练，通过提示修改：边框分割、视觉问答、语言指令控制等
- 四大类模型演进时间线
  - 传统模型：AlexNet → ResNet → ViT → DINO（1999—2021）
  - 文本提示模型：CLIP、ALIGN、BLIP、Flamingo、GPT-4（2021—2023）
  - 视觉提示模型：SAM、CAT、SAM-Track（2023）
  - 多模态模型：ImageBind、VOYAGER、PaLM-E、MineDojo（2023）
- 核心价值
  - 取代多个狭窄的任务特定模型，提供更广泛通用的基础，大幅降低开发成本

基础模型：AI的范式转变

核心定义

CONCEPT / 概念引入

由斯坦福HAI研究院的Bommasani等人正式提出，是当前AI领域最核心的研究方向之一。

DEFINITION / 本质特征

在自监督或半监督方式下，利用海量无标注数据进行预训练。具备极强的泛化能力，能够快速适应多个下游任务。

模型通用性

打破“一个任务一个模型”的局限，用单一、广泛的通用模型，覆盖并解决多个垂直领域的特定问题。

开发效率跃升

实现“一次预训练，多次微调”。大幅降低模型开发门槛，缩短应用落地周期，快速响应业务需求。

全域性能优化

得益于海量数据的训练，在已知的域内任务和未知的域外场景中，均能提供显著优于传统模型的预测精度。

涌现智能特性

模型规模突破临界点后，自发产生零样本学习、思维链推理等新兴智能属性，具备了类人类的基础认知能力。

多模态基础模型：连接视觉与语言的桥梁

核心定义

多模态基础模型是人工智能领域的重要突破，它能够通过深度学习技术，同时学习并缩小图像、文本、语音等不同模态数据之间的语义鸿沟，实现跨模态的统一理解与生成。

关键能力

上下文推理
理解视觉场景中对象间的复杂逻辑关系

环境泛化能力
对现实世界中的歧义、变化有极强适应性

动态提示能力
无需重训，通过文本/视觉提示灵活修改输出

交互示例

视觉问答 (VQA)
基于图像内容进行自然语言的交互式问答

视觉指令分割
根据语言指令在图像中精准分割特定对象

智能机器人控制
将语言指令转化为机器人的具体执行动作

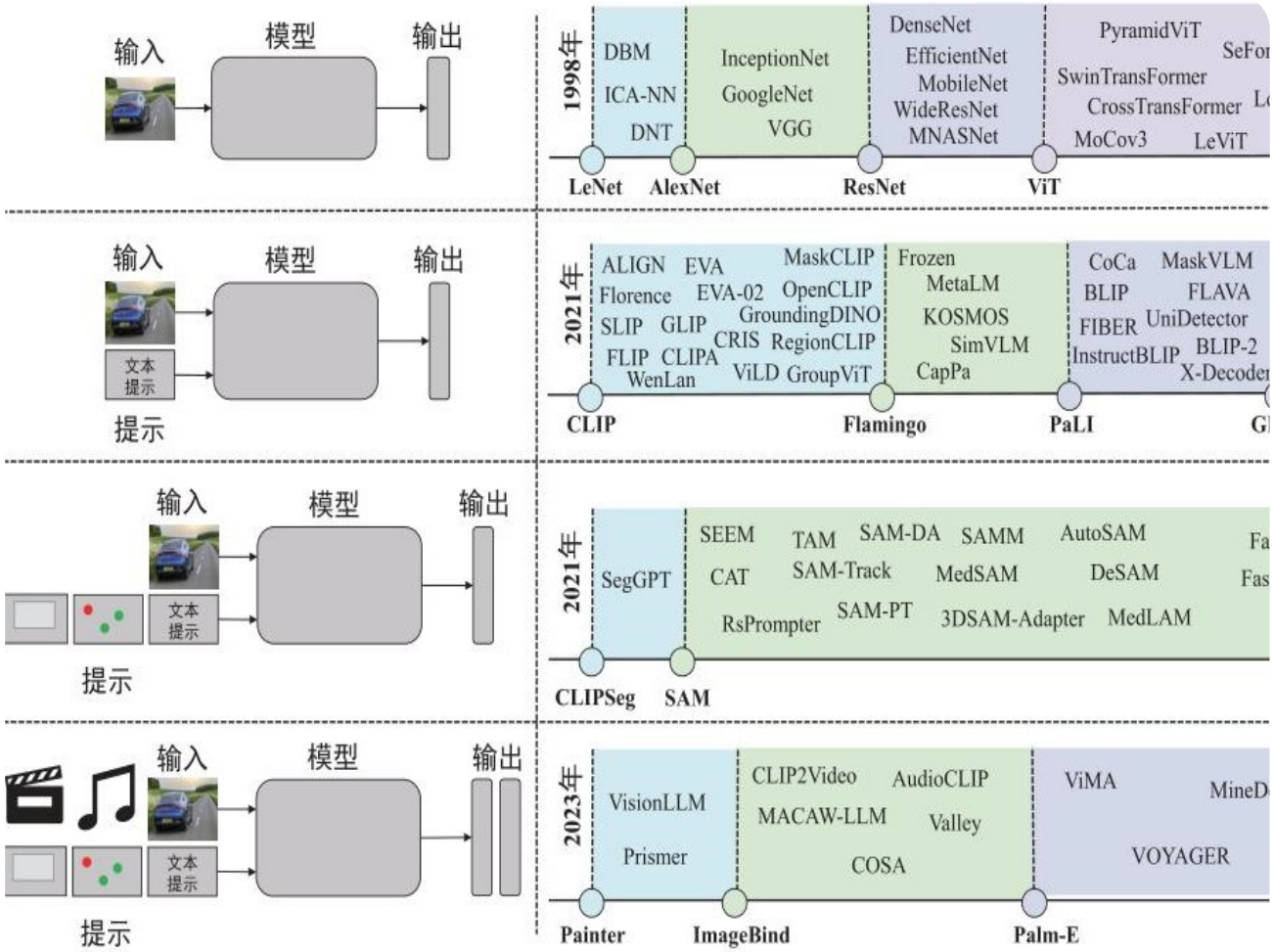
多模态基础模型的演进路径

四个核心演进阶段

本页展示了多模态基础模型从2021年至今的发展脉络。演进路径清晰地体现了模型架构从**单模态向多模态**融合，以及任务处理能力从**任务特定向通用化**不断拓展的核心趋势。

能力持续增强

交互方式日益丰富，应用边界不断扩大



演进路径解读

传统模型

Pre-2021

核心代表

CNN (AlexNet, VGG, ResNet), 早期 Vision-Language 模型。

阶段特点

模型设计面向特定任务，跨领域的泛化能力较为有限。

文本提示

2021 - 2022

核心代表

CLIP, ALIGN, CoCa, PaLM-E, GPT-4 等多模态模型。

阶段特点

通过自然语言指令实现零样本学习，正式开启了语言监督的新时代。

视觉提示

2022 - 2023

核心代表

SAM, SEEM, FastSAM, GroundingDINO 等分割/定位模型。

阶段特点

引入点、框等视觉提示实现通用分割，极大地拓展了人机交互的边界。

多模态融合

2023 Onwards

核心代表

VisionLM, ImageBind, MineDojo 等大一统感知模型。

阶段特点

深度融合图像、文本、音视频等多种模态，具备复杂的综合感知与交互能力。

本节内容

CONTENTS

CLIP的诞生背景与核心思想

核心思想： 机器无需依赖人工标注，而是直接从自然语言中存在的视觉概念中学习感知，建立图像与文本之间的深层语义关联。

传统方法的局限性

依赖固定标签： 严重依赖人工标注的类别标签（如ImageNet-1k），标注成本极高且难以覆盖长尾类别。

泛化能力薄弱： 模型被限制在预定义的封闭集合中，在面对训练集外的新类别时，识别能力会显著下降。

CLIP 的突破性创新

语言作为监督信号： 利用互联网中海量的“图像-文本”对作为训练数据，彻底摆脱了对人工固定标签的依赖。

对比学习驱动训练： 通过对比图像与文本的匹配关系，让模型学会判断“哪张图对应哪段描述”，从而习得通用视觉概念。

CLIP的核心训练范式

对比图像-语言预训练

Contrastive Language-Image Pre-training

本质定义

一种基于对比学习框架的多模态大模型，打破单一视觉模态的局限。

训练数据

海量的图像-文本对数据，即一张图像对应一段准确描述它的自然语言文本。

核心目标

在特征空间中拉近匹配的图文对，推远不匹配的负样本对，建立跨模态对齐。

范式差异：VS 传统CV

Different from Traditional Contrastive Learning

SimCLR / MoCo (传统视觉)

对比对象局限在单一视觉模态内。正样本为同一图像的不同数据增强视图，负样本为批次中的其他图像。

CLIP (多模态对比)

打破视觉模态壁垒。对比图像与文本的跨模态匹配关系，让模型理解图像内容与语义描述的深层联系。

CLIP模型架构与流程

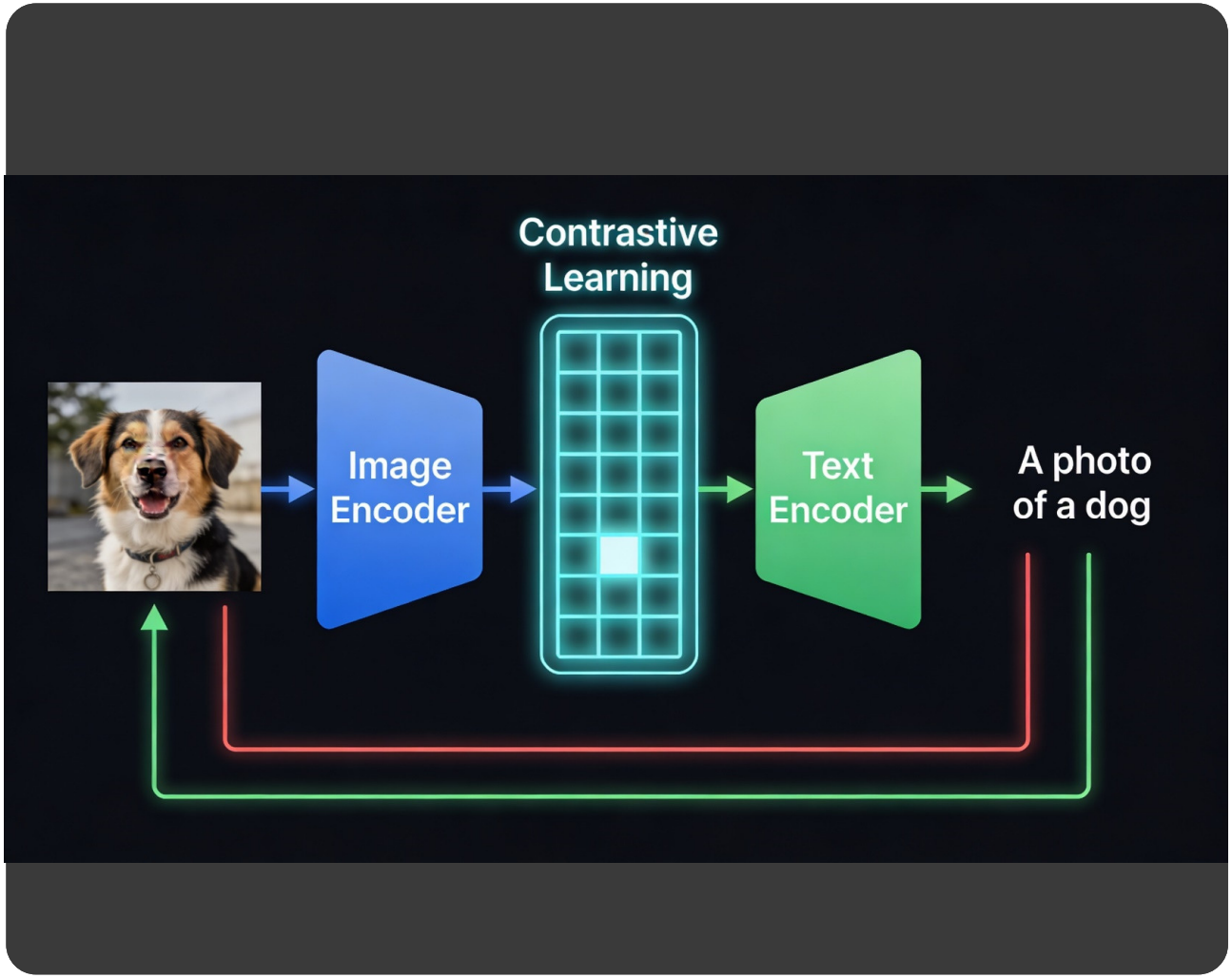
图像编码器 (Image Encoder)

核心架构：ViT (Vision Transformer) 或 ResNet。负责将原始像素转换为高维特征向量。

文本编码器 (Text Encoder)

核心架构：Transformer。负责将文本描述转换为与图像同维度的特征向量，实现语义对齐。

模型通过“对比学习”联合训练，在推理阶段利用“零样本分类”直接匹配图像与文本特征。



CLIP的基石：创建足够大的数据集

现有数据集的局限性

MS-COCO / Visual Genome

优势是数据标注质量极高，但核心痛点是规模太小（约10万张），难以支撑百亿级参数模型的训练需求。

YFCC100M (亿级数据集)

虽然拥有海量规模，但元数据稀疏、描述随意，且存在大量低质量、噪声数据，无法直接用于高质量对齐任务。

CLIP 的破局之道：WIT 数据集

核心目标

构建一个“规模足够大、概念覆盖足够广”的高质量图像-文本对齐数据集。

实施策略

从互联网海量公开资源中广泛抓取，并通过特定的筛选算法进行严格的去噪与平衡。

WIT 数据集成果

最终构建了包含4亿+个有效图像-文本对的超大规模数据集。

CLIP的核心：选择有效的预训练方法

早期尝试

早期研究（如VirTex）致力于共同训练图像和文本编码器，试图让模型直接预测图像的完整描述。

关键发现

基于对比的表征学习研究表明：**对比目标函数的性能显著优于等效的预测目标函数。**

CLIP的核心决策

代理任务重构：放弃生成文本，预测“图文配对关系”。

零样本迁移提速：

4X

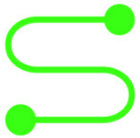
CLIP的简化与优化：训练细节

大道至简：CLIP的训练简化策略

由于数据集规模巨大，过拟合不是主要问题，因此研究者大胆简化了训练过程，反而提升了模型效率。

从头开始训练

摒弃预训练权重初始化，直接在海量数据上从零开始训练。



线性投影层

移除复杂的非线性投影网络，仅保留线性层映射到多模态空间。

极简数据增强

去除大部分复杂的视觉变换，仅保留最基础的随机裁剪策略。

动态温度参数

将对比损失中的温度系数设为可学习参数，随训练动态调整。

文本处理简化

移除复杂的文本转换函数，对句子进行均匀采样输入模型。

CLIP的模型架构：图像编码器

ResNet-based 架构

基础：基于经典的 ResNet50 网络结构进行改进。

改进：引入 ResNetD 抗锯齿池化，并将全局平均池化替换为**注意力池化机制**，以更好地聚合全局特征。

ViT-based (Vision Transformer)

基础：完全遵循 ViT (Vision Transformer) 的核心实现逻辑。

微调：添加额外的层归一化 (Layer Norm) 层，并使用了针对对比学习优化的初始化方案。

CLIP的模型架构：文本编码器

核心架构原理

采用标准的 Transformer 编码器堆叠结构，通过“自注意力机制+前馈网络”的经典组合，实现对输入文本序列的深度语义特征提取。

关键模型参数配置

堆叠层数	隐含层维度	注意力头数
12 Layers	**512 Dim**	**8 Heads**

特征映射与输出

选取 Transformer 最高层在**EOS ([SEP])**标记处的输出作为文本特征；经层归一化与线性投影后，映射到多模态统一嵌入空间。

CLIP的模型扩展策略

协同缩放：宽度、深度、分辨率的平衡

多维协同 EfficientNet

突破单维度扩展瓶颈，在宽度、深度和分辨率所有维度上**均衡分配**额外的计算资源，实现整体性能的最优增长。

图像编码器 (ResNet)

采用ResNet变体结构，将额外算力**平均分配**到模型的宽度、深度和分辨率三个维度上，最大化视觉特征提取能力。

文本编码器 (Transformer)

策略聚焦于**仅按比例扩展宽度**。研究实验表明，模型性能对文本编码器的“深度”变化不敏感，扩展宽度性价比更高。

CLIP的预训练与模型变体

模型体系架构

ResNet 系列

ResNet50 / ResNet101
RN50x4 / RN50x16 /
RN50x64

Vision Transformer 系列

ViT-B/32 / ViT-B/16
ViT-L/14 (高性能大模型)

海量算力投入

RN50x64 (最大ResNet)

基于**592**块 V100 GPU

连续训练时长：**18 天**

ViT-L/14 (最大ViT)

基于**256**块 V100 GPU

连续训练时长：**12 天**

性能优化技巧

FixRes 优化策略

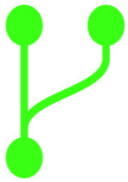
对ViT-L/14在336px高分辨率下进行额外预训练，显著提升图像特征提取能力。

最终模型：ViT-L/14@336px

CLIP的革命性影响

开创语言监督新时代

证明了通过大规模图像-文本对进行对比学习，可使模型获得强大的零样本泛化能力，摆脱对人工标注的依赖。



推动视觉范式转变

实现了从依赖人工标注类别标签，向利用更丰富、更灵活的自然语言作为监督信号的根本性范式转变。

多模态模型的基石

其构建的对比学习框架，成为了BLIP、ALIGN等后续大量多模态模型进行改进和创新的核心基础。

CLIP的关键特性与局限性

关键特性 KEY FEATURES

强大的零样本性能

在多个下游视觉任务上无需微调即可达到不错的性能，具备极强的泛化基础。

良好的鲁棒性

对数据分布迁移和跨领域泛化场景具有较高的稳定性，抗干扰能力强。

高效的线性微调

在基于线性探测的下游微调任务中表现优异，训练成本低且效果显著。

局限性 LIMITATIONS

细粒度理解能力不足

模型对图像中非常细微的局部特征（如小物体、精细纹理）的捕捉和理解相对较弱。

缺乏主动生成能力

CLIP 核心侧重于图像与文本的双向匹配和语义理解，并不擅长生成全新的文本或图像内容。

CLIP的应用案例

零样本图像分类

打破传统模型对标注数据的依赖，无需额外训练，直接使用CLIP对全新的类别进行高效分类识别。

视觉搜索 (Text-to-Image)

利用CLIP的跨模态对齐能力，根据用户输入的自然语言文本描述，在海量图片库中精准检索出语义匹配的图像。

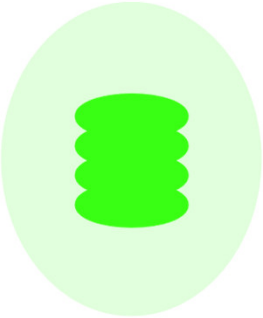
可控风格迁移

结合Diffusion等生成模型，通过CLIP引导，根据文本提示生成特定艺术风格、特定内容的高质量图像。

智能数据增强

自动为现有的图像生成多样化、高质量的文本描述，作为辅助标注数据，有效提升下游视觉模型的训练效果。

CLIP的后续发展



ALIGN (Google)

采用类似CLIP的对比学习框架，但使用了更大规模的数据集（18亿图像-文本对），进一步显著提升了模型的泛化性能。

CoCa (Google)

创新性地结合了对比学习和生成式学习，成功统一了图像文本检索和图像字幕生成两大核心任务，实现了多模态的融合。

PaLM-E (Google)

将CLIP的视觉表征能力与大型语言模型PaLM深度结合，赋予了AI强大的具身推理能力，使其能理解物理世界并进行规划。

CLIP — 数据集构建与预训练策略

- 数据集构建挑战
  - 现有数据集（MS-COCO / Visual Genome / YFCC100M）规模不足或质量参差
  - YFCC100M 拥有 1 亿张图像，但元数据稀疏，质量差异大
- WIT 数据集（WebImageText）
  - 从互联网公开获取 4 亿图像-文本对
  - 每个查询最多包含 2 万个图文对，保持类别平衡
  - 总词数与用于训练 GPT-2 的 WebText 数据集相似
- 预训练方法选择
  - 最初类似 VirTex：共同训练编码器以预测图像描述
  - 改用对比学习目标：效率比预测目标高约 4 倍
  - 简化处理：移除非线性投影、文本 Transformer 仅均匀采样一句话
  - 数据增强：仅随机裁剪正方形，无额外增强
  - 温度参数 τ 在训练过程中直接优化，不作为超参数固定

影响力

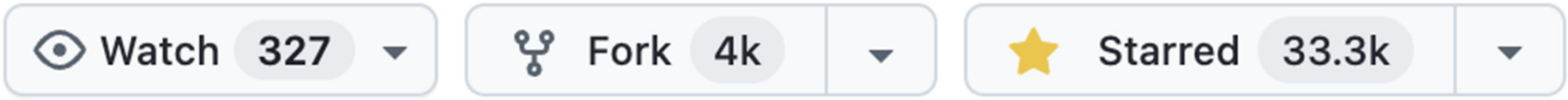
arXiv

https://arxiv.org › cs · 翻译此页

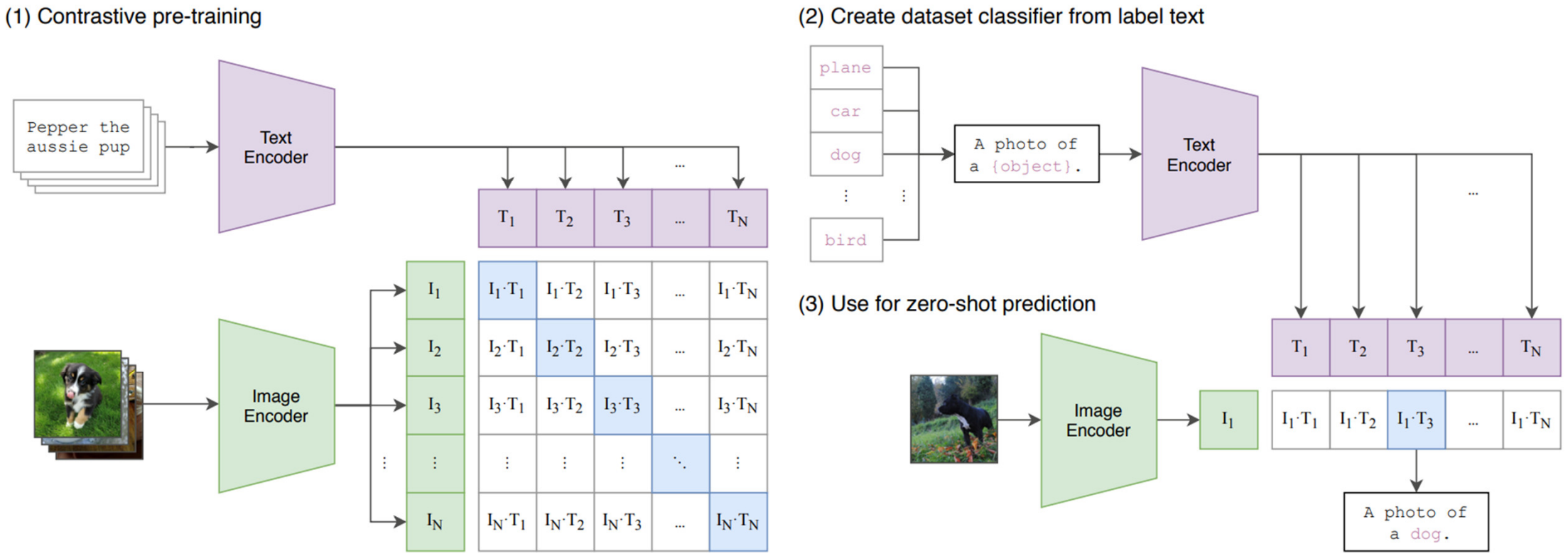
Learning Transferable Visual Models From Natural ...

作者：A Radford · 2021 · 被引用次数：59259 — Access Paper: View a PDF c

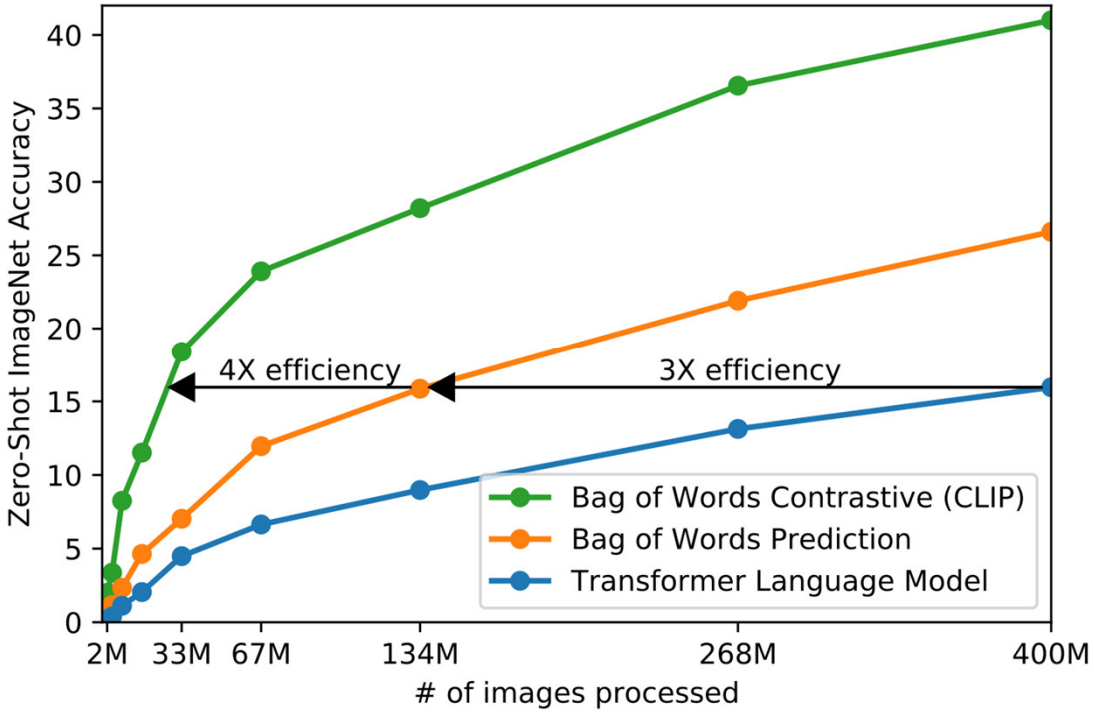
Transferable Visual Models From Natural Language Supervision, by Alec Rad

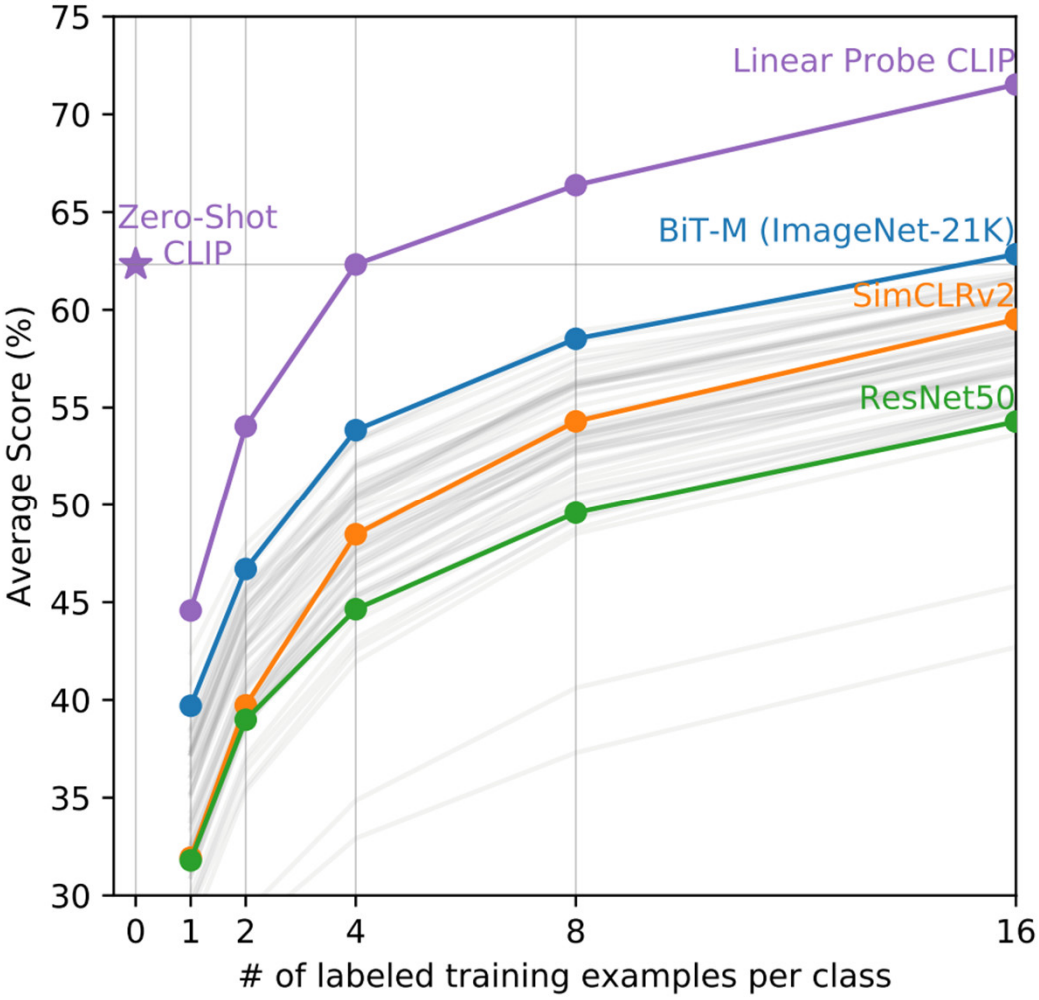


方法


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

效果


Zero-Shot ImageNet Accuracy
40
35
30
25
20
15
10
5
0
4X efficiency
3X efficiency
Bag of Words Contrastive (CLIP)
Bag of Words Prediction
Transformer Language Model
2M
33M
67M
134M
268M
400M
# of images processed


Average Score (%)
75
70
65
60
55
50
45
40
35
30
Linear Probe CLIP
Zero-Shot
CLIP
BiT-M (ImageNet-21K)
SimCLRv2
ResNet50
0
1
2
4
8
16
# of labeled training examples per class

代码

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

CLIP 核心总结

核心思想

突破传统监督学习范式，通过**对比学习**机制，直接从海量的自然语言监督信号中高效学习通用的视觉概念。

关键创新点

- 以**自然语言**作为监督信号
- 构建**图文对比学习**训练目标
- 利用WIT等**大规模数据集**预训练

主要贡献

开创了**多模态预训练**的全新范式，充分证明了**零样本迁移学习**在视觉任务中的巨大潜力与可行性。

本节内容

CONTENTS

一、CLIP

二、BLIP 和 BLIP-2

三、LLaMA 和 LLaMA-Adapter

四、VideoChat

BLIP的提出背景与核心贡献

现有 VLP 模型的局限

任务割裂：理解与生成难以兼顾

大多数模型仅在单一任务（如检索或字幕）表现出色，缺乏统一的多模态建模能力。

数据低效：噪声数据利用率低

单纯扩大网络爬取的带噪图文对规模，是一种次优的监督方式，难以充分挖掘数据价值。

BLIP 的两大核心贡献

MED 混合架构：统一理解与生成

提出 Encoder-Decoder 的混合模态结构，打破任务壁垒，实现了多模态理解与生成任务的统一建模。

CapFilt 框架：高效利用噪声数据

引入“Captioning and Filtering”策略，主动从海量带噪网络数据中提炼高质量监督信号，提升学习效率。

BLIP的性能突破

SOTA 性能表现

在图像-文本检索、图像字幕生成、视觉问答(VQA)等多项视觉-语言跨模态任务上，刷新了当时的最先进（SOTA）成果。

AIGC 核心应用

广泛用于生成高质量图像提示词(Prompt)，是 AIGC 生态的重要上游组件。例如 ControlNet 中的 Automatic Prompt 功能，正是基于 BLIP 模型实现的。

图像文本检索

平均召回率 (Recall) 提升 **+2.7%**

图像字幕生成

CIDEr 评估分数提升 **+2.8%**

视觉问答 (VQA)

VQA Accuracy 准确率提升 **+1.6%**

BLIP模型结构：MED混合架构

统一理解与生成的MED架构

Multimodal Mixture of Encoder-Decoder

统一的模型框架

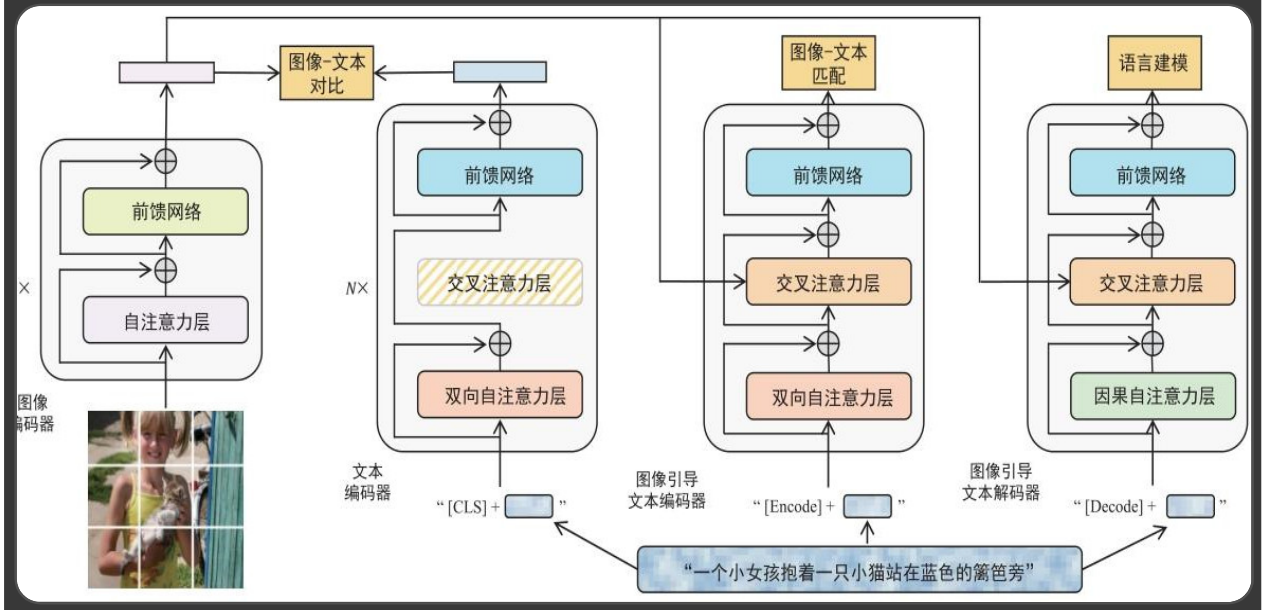
一套架构灵活适配多种任务，避免重复建模，降低系统复杂度。

支持三种核心模式

集成单模态对比、图像引导匹配与文本生成，覆盖理解与生成任务。

极致的效率与性能

共享底层视觉与语言特征，显著提升模型在下游任务中的表现。



BLIP — 统一理解与生成框架（Salesforce, 2022）

核心贡献

在多项视觉-语言任务上取得 SOTA，实现了理解与生成的统一。

- 技术贡献1：MED（多模态编码器-解码器混合）

灵活的多任务预训练架构，可作为单模态编码器、图像引导文本编码器或图像引导文本解码器运行

- 技术贡献2：CapFilt（字幕生成与过滤）

从噪声图像-文本对中学习的数据引导新方法，提升数据质量

在 AIGC 中广泛应用（如 ControlNet 中 Automatic Prompt 由 BLIP 生成）

三个预训练目标

- ITC（图像-文本对比学习）

激活单模态编码器，对齐视觉和文本特征空间

- ITM（图像-文本匹配损失）

激活图像引导文本编码器，细粒度图文多模态表示学习，使用硬负向挖掘策略

- LML（语言建模损失）

激活图像引导文本解码器，在给定图像条件下生成文本描述

三个目标联合训练，共享参数提高训练效率，LML 标签平滑设为 0.1

BLIP — MED 模型结构

- 单模态编码器（Unimodal Encoder）
  - 分别对图像和文本进行编码
  - 文本编码器在输入开头添加 [CLS] 标记以汇总整句
- 图像引导文本编码器（Image-Grounded Text Encoder）
  - 在文本编码器每个 Transformer 块的双向自注意力层与前馈网络之间插入 Cross-Attention（CA）层
  - CA 层将视觉信息注入文本流
  - 任务特定的 [Encode] 标记附加到文本上，其输出嵌入用作图文多模态表示
- 图像引导文本解码器（Image-Grounded Text Decoder）
  - 将双向自注意力层替换为因果自注意力层
  - [Decode] 标记标记序列开始，[EOS] 标记序列结束
- 参数共享策略
  - 编码器和解码器共享除自注意力层外的全部参数
  - 原理：编码任务使用双向自注意力，解码任务使用因果自注意力，差异最好由自注意力层捕获
  - 嵌入层、交叉注意力层和前馈网络在编码和解码任务之间功能类似，故共享

BLIP — CapFilt 数据引导方法

CapFilt 的问题背景

注释成本高 → 有限的高质量人工标注数据（如 COCO）

大量网络图像-文本对存在噪声：

- 图像文件名作为"标题"（如 "20160716 113957.JPG"）
- 包含相机曝光设置的"描述"文字
- 经过筛选，YFCC100M 仅剩 1,500 万张图像

目标：从嘈杂的网络数据中学习，提高数据质量，同时保留大规模数据优势

CapFilt 两个核心模块

- 字幕生成器（Captioner）

  图像引导文本解码器

  在 COCO 上微调后，为网络图像生成合成字幕 Ts
- 过滤器（Filter）

  图像-文本编码器（对比+匹配目标微调）

  判断文本是否与图像匹配

  从原始网络文本和合成文本中删除噪声
- 效果提升

  图文检索平均召回率 +2.7%

  图像字幕 CIDEr 指标 +2.8%

  视觉问答 +1.6%

影响力

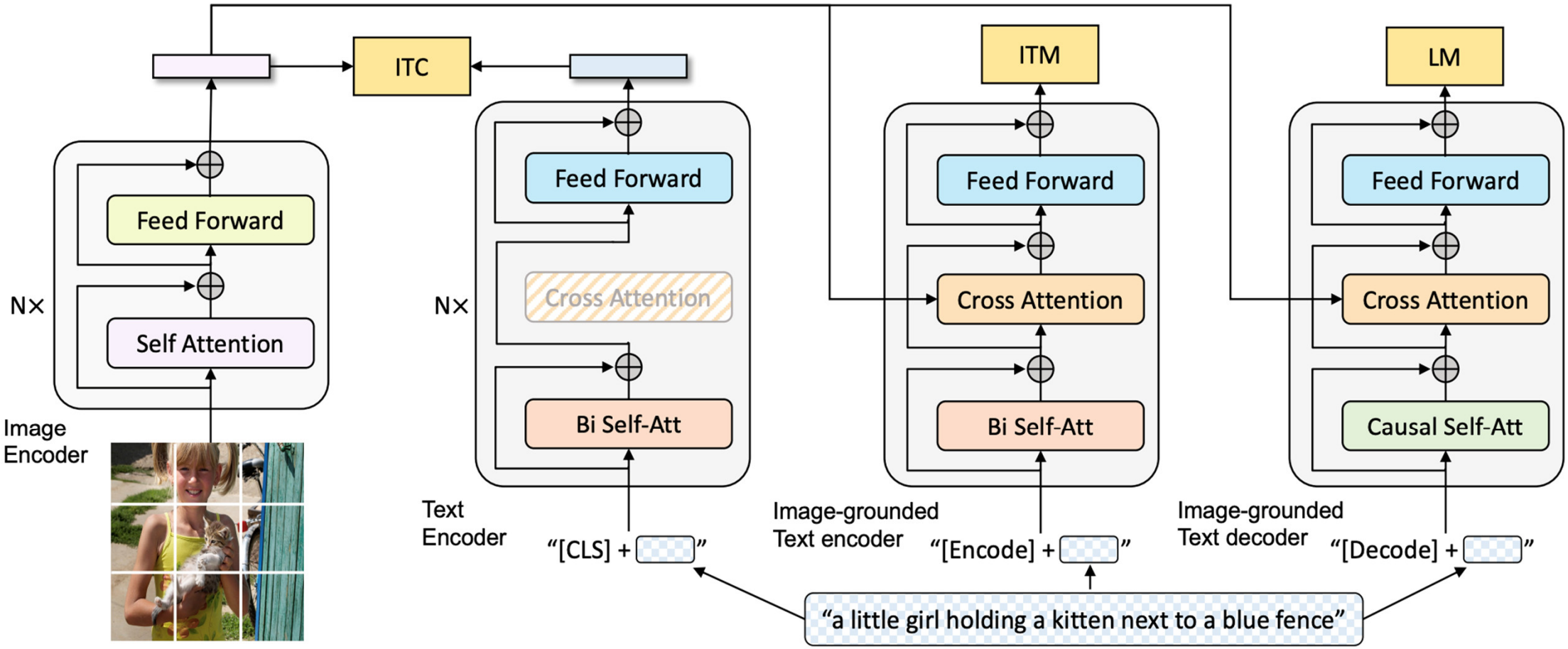
arXiv

https://arxiv.org › cs · 翻译此页

BLIP: Bootstrapping Language-Image Pre-training for ...

作者： J Li · 2022 · 被引用次数： 8802 — In this paper, we propose BLIP, a new transfers flexibly to both vision-language understanding and generation tasks.

方法


ITC
ITM
LM
Feed Forward
Self Attention
N×
Image
Encoder
Feed Forward
Cross Attention
Bi Self-Att
N×
Text
Encoder
"[CLS] +
Feed Forward
Cross Attention
Bi Self-Att
Image-grounded
Text encoder
"[Encode] +
Feed Forward
Cross Attention
Causal Self-Att
Image-grounded
Text decoder
"[Decode] +
"a little girl holding a kitten next to a blue fence"

数据

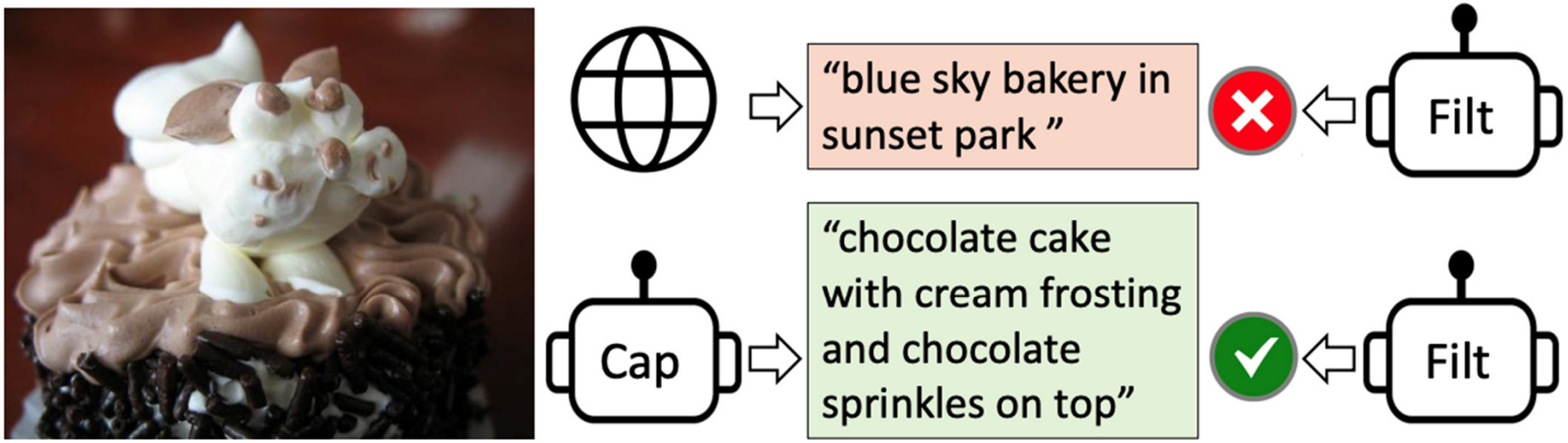
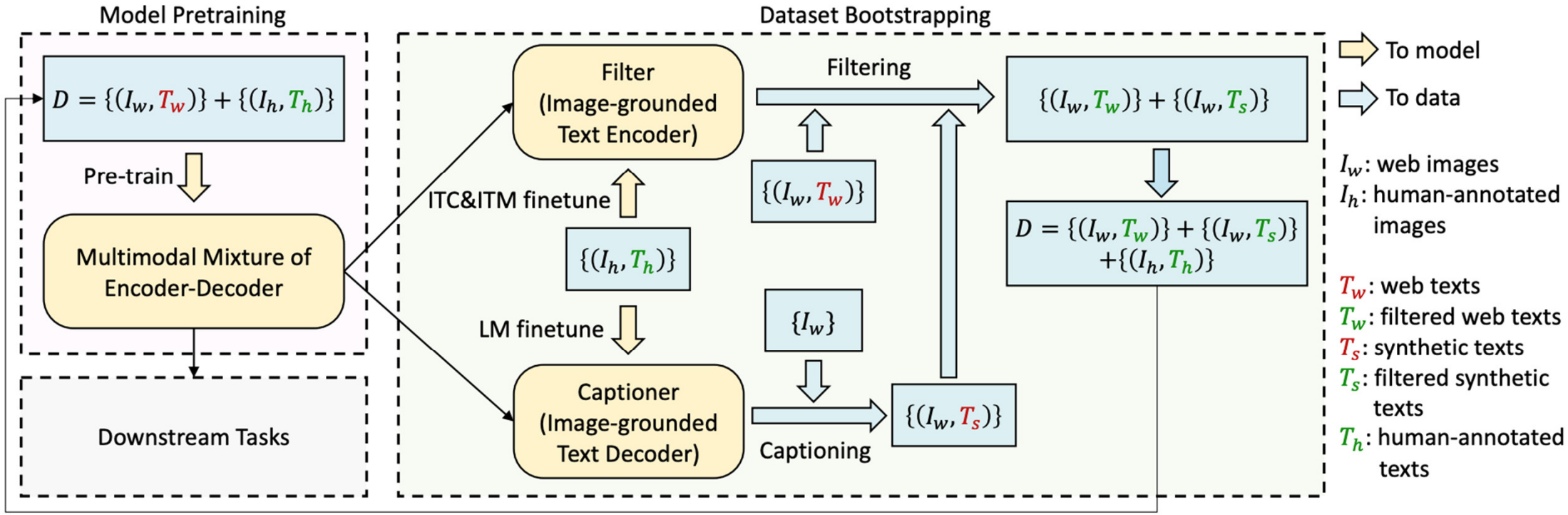


Figure 1. We use a Captioner (Cap) to generate synthetic captions for web images, and a Filter (Filt) to remove noisy captions.

方法

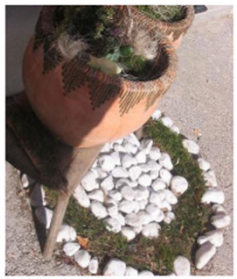

Model Pretraining
Dataset Bootstrapping
$D = \{(I_w, T_w)\} + \{(I_h, T_h)\}$
Pre-train
Multimodal Mixture of Encoder-Decoder
Downstream Tasks
Filter (Image-grounded Text Encoder)
ITC&ITM finetune
$\{(I_h, T_h)\}$
LM finetune
Captioner (Image-grounded Text Decoder)
Filtering
$\{(I_w, T_w)\}$
$\{I_w\}$
Captioning
$\{(I_w, T_s)\}$
$\{(I_w, T_w)\} + \{(I_w, T_s)\}$
$D = \{(I_w, T_w)\} + \{(I_w, T_s)\} + \{(I_h, T_h)\}$
To model
To data
I_w: web images
I_h: human-annotated images
T_w: web texts
T_w: filtered web texts
T_s: synthetic texts
T_s: filtered synthetic texts
T_h: human-annotated texts

效果

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	✓$_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	✓$_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	✓$_B$	✓$_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	✓$_B$	✓$_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	✓$_L$	✓$_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	✓$_L$	✓$_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"



T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

T_w: "the current castle was built in 1180, replacing a 9th century wooden castle"

T_s: "a large building with a lot of windows on it"

Figure 4. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

BLIP-2 — 高效视觉-语言预训练（Salesforce, 2023）

- 研究背景与动机
  - 视觉-语言预训练成本随大模型端到端训练急剧增高
  - BLIP-2 提出：从现成的冻结预训练图像编码器和冻结 LLM 中引导视觉语言预训练
- 核心组件：Q-Former（查询 Transformer）
  - 轻量级 Transformer，通过可学习的查询向量从冻结图像编码器中提取视觉特征
  - 充当冻结图像编码器和冻结 LLM 之间的信息瓶颈
  - 188M 参数，32 个查询，每个查询维度 768
  - BERT_base 预训练权重初始化，交叉注意力层随机初始化
- 两阶段预训练
  - 阶段一：使用冻结图像编码器进行视觉与语言表示学习（ITC + ITG + ITM）
  - 阶段二：使用冻结 LLM 进行从视觉到语言的生成学习（全连接层投影）
- 核心优势
  - 可训练参数远少于现有方法，但性能最先进
  - 可与任意强大 LLM 组合，复用成熟 LLM 能力，无需端到端训练

BLIP-2 — Q-Former 架构

Q-Former 组成

两个 Transformer 子模块，共享自注意力层：

- 图像 Transformer

  与冻结图像编码器交互，提取视觉特征

- 文本 Transformer

  既可作为文本编码器，也可作为文本解码器

查询通过自注意力层彼此交互，通过交叉注意力层（每隔一个 Transformer 块一个）与图像特征交互，同时通过共同的自注意力层与文本交互。

输出 Z（32 × 768）远小于冻结图像特征（ViT-L/14：257 × 1,024），形成有效信息瓶颈。

阶段一预训练目标

- ITC（图像-文本对比学习）

  对齐图像与文本表示；单模态自注意力掩码，查询不与文本 Token 交互

- ITG（图像引导文本生成）

  Q-Former 以图像为条件生成文本；多模态因果自注意力掩码；[DEC] 标记替换 [CLS]

- ITM（图像-文本匹配）

  细粒度对齐，二元分类；双向自注意力掩码，所有查询和文本可相互关注；硬负向挖掘策略

不同掩码策略灵活控制查询与文本的交互，一个 Q-Former 高效支持三个目标

BLIP-2 — 第二阶段：从视觉到语言的生成学习

- 目标：利用 LLM 的语言生成能力
  - 将 Q-Former 的输出查询表示 Z 通过全连接层线性投影到与 LLM 文本表示相同维度
  - 在输入文本之前添加投影的查询表示，作为软的视觉提示，将 LLM 置于视觉特征之上
- 基于解码器的 LLM（如 OPT）
  - 使用语言建模损失预训练
  - 冻结 LLM：在图像条件下生成文本，冻结任务是基于 Q-Former 提取的视觉特征生成文本
- 基于编码器-解码器的 LLM（如 FlanT5）
  - 使用前缀语言建模损失：前缀文本与视觉特征拼接作为 LLM 编码器输入
  - 后缀文本作为 LLM 解码器的生成目标
- 预训练配置
  - 训练数据：129M 张图像（COCO / Visual Genome / CC3M / CC12M / SBU / LAION400M 中 115M）
  - 使用 CapFilt 合成标题，保留每张图像前两个标题
  - 阶段一 250K 轮次（批量 ViT-L:2320, ViT-g:1680），阶段二 80K 轮次
  - 冻结 ViT 参数转换为 FP16，提升计算效率

影响力

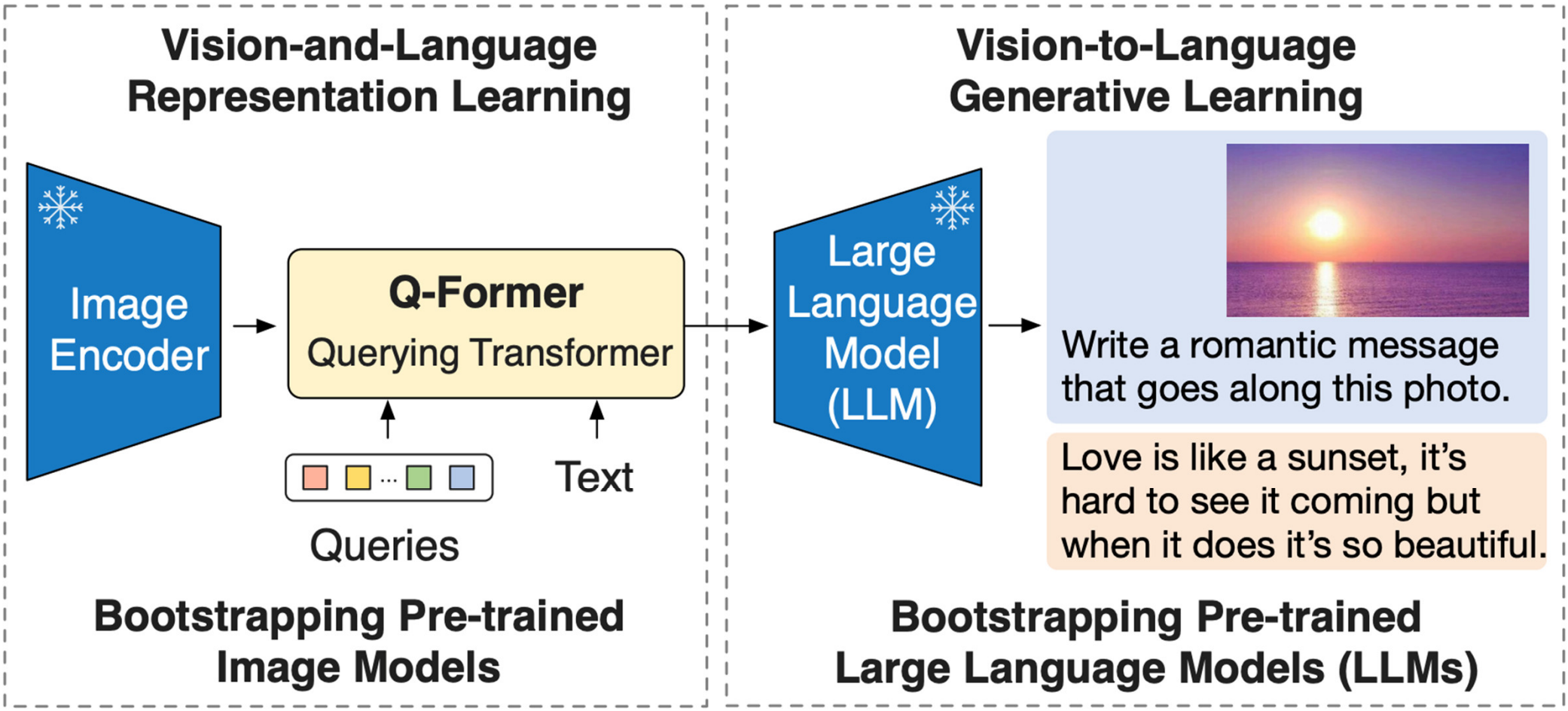
arXiv

https://arxiv.org › cs · 翻译此页

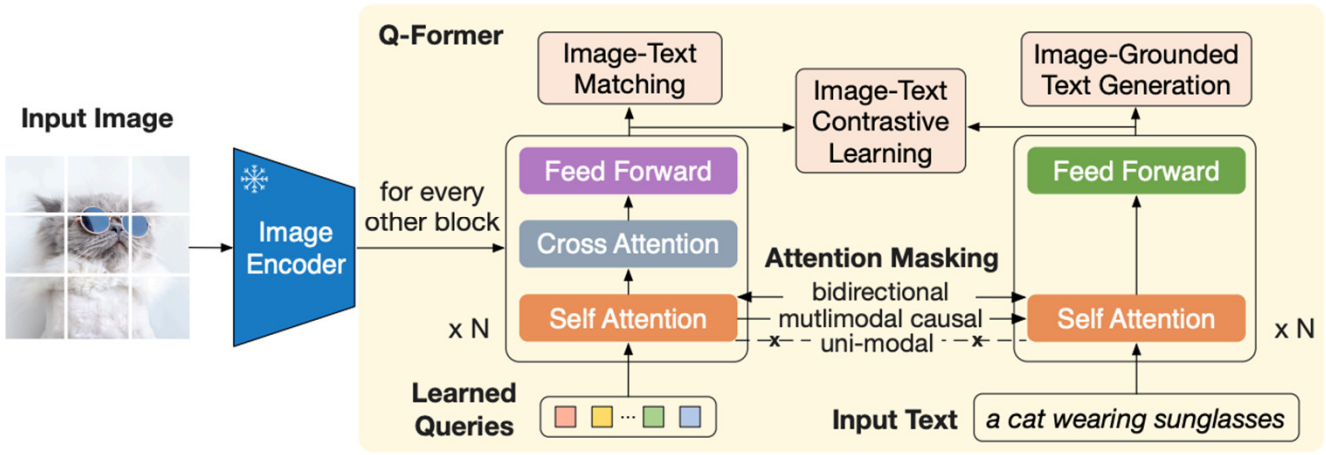
BLIP-2: Bootstrapping Language-Image Pre-training with ...

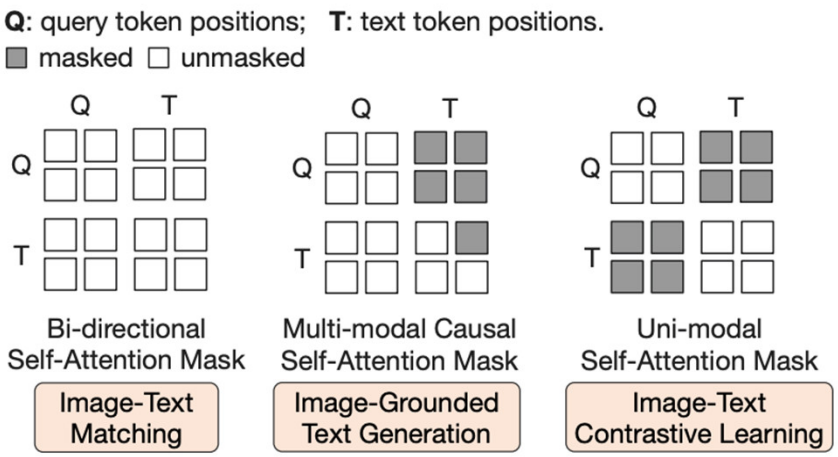
作者： J Li · 2023 · 被引用次数： 11357 — This paper proposes BLIP-2, a generic anc training strategy that bootstraps vision-language pre-training from off-the-shelf frozen

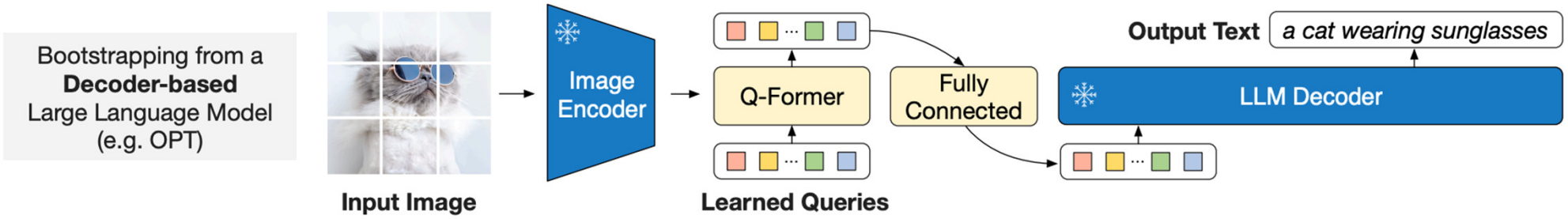
方法

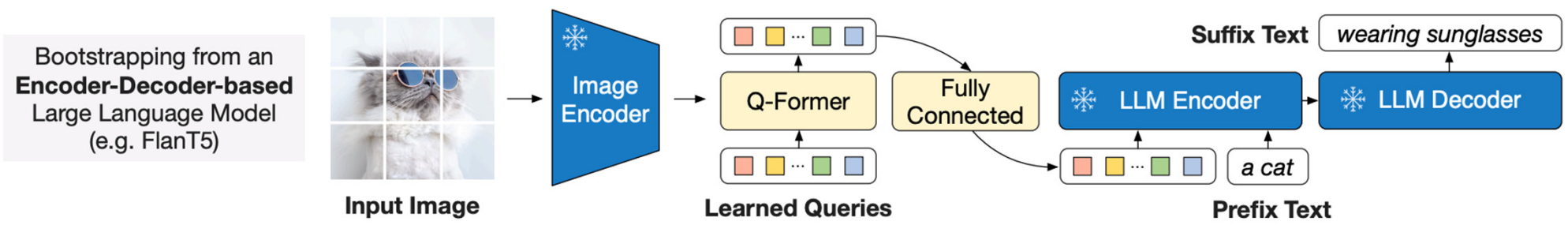

Vision-and-Language
Representation Learning
Image
Encoder
Q-Former
Querying Transformer
Text
Queries
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large
Language
Model
(LLM)
Write a romantic message
that goes along this photo.
Love is like a sunset, it's
hard to see it coming but
when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

方法


Input Image
Image Encoder
for every other block
Q-Former
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
Feed Forward
Cross Attention
Self Attention
x N
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Feed Forward
Self Attention
x N
Learned Queries
Input Text
a cat wearing sunglasses


Q: query token positions; T: text token positions.
masked unmasked
Q T
Q
T
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning


Bootstrapping from a **Decoder-based** Large Language Model (e.g. OPT)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
Output Text
a cat wearing sunglasses
LLM Decoder


Bootstrapping from an **Encoder-Decoder-based** Large Language Model (e.g. FlanT5)
Input Image
Image Encoder
Q-Former
Learned Queries
Fully Connected
Suffix Text
wearing sunglasses
LLM Encoder
LLM Decoder
a cat
Prefix Text

效果

Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.

What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.

What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

效果

Models	#Trainable Params	Open-sourced?	Visual Question Answering VQAv2 (test-dev) VQA acc.	Image Captioning NoCaps (val) CIDEr	Image Captioning NoCaps (val) SPICE	Image-Text Retrieval Flickr (test) TR@1	Image-Text Retrieval Flickr (test) IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	**65.0**	**121.6**	**15.8**	**97.6**	**89.7**

Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{no\text{-}vqa}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

本节内容

CONTENTS

一、CLIP

二、BLIP 和 BLIP-2

三、LLaMA 和 LLaMA-Adapter

四、VideoChat

LLaMA — 开源大型语言基础模型（Meta AI）

LLaMA 核心特点

参数规模：7B / 13B / 33B / 65B

- 训练策略

  在数十亿标记上训练，仅使用公开可获取数据

  使用更多标记（而非更大模型）以在推理预算下实现最佳性能

- 关键性能

  LLaMA-13B 在大多数基准测试中优于 GPT-3（175B），体积仅为其 1/10

  LLaMA-65B 与 Chinchilla 和 PaLM-540B 媲美

- 开源优势

  可在单个 GPU 上运行，降低访问和研究门槛

  仅使用公开数据，与开源兼容，便于社区复现

训练数据构成（共 ~1.4TB 标记）

- English CommonCrawl（67%）

  2017-2020 年 5 个 CommonCrawl 数据集

  CCNet 流程：行级去重，FastText 语言识别，n-gram 过滤低质量内容

- C4（15%）：公开可获取的 C4 数据集
- GitHub（4.5%）：Apache / BSD / MIT 许可的公共代码
- Wikipedia（4.5%）：20 种语言，2022年6-8月数据
- Gutenberg + Books3（4.5%）：公共领域图书
- arXiv（2.5%）：科学数据，清除注释
- Stack Exchange（2%）：高质量问答

LLaMA — 网络结构改进

- 基于标准 Transformer，融合三项关键改进
- 改进1：预归一化 + RMSNorm（受 GPT-3 启发）
  - 为提高训练稳定性，归一化每个 Transformer 子层的输入（而非输出）
  - 使用 RMSNorm 归一化函数，计算效率更高
- 改进2：SwiGLU 激活函数（受 PaLM 启发）
  - 将 ReLU 替换为 SwiGLU，提高性能
  - 将维度从 4d 扩大到 34d/3（约 2.67d）
- 改进3：旋转位置嵌入 RoPE（受 GPTNeo 启发）
  - 删除绝对位置嵌入，在每一层添加旋转位置嵌入
  - 更好地建模相对位置关系，支持更长上下文
- 优化器与训练配置
  - AdamW 优化器（beta1=0.9, beta2=0.95），余弦学习率调度，权重衰减 0.1，梯度剪裁 1.0
  - 2,000 步预热，预热后学习率最终衰减至最大值的 10%
  - 65B 模型：2,048 块 A100 GPU（80GB 内存），约 21 天，约 380 标记/秒

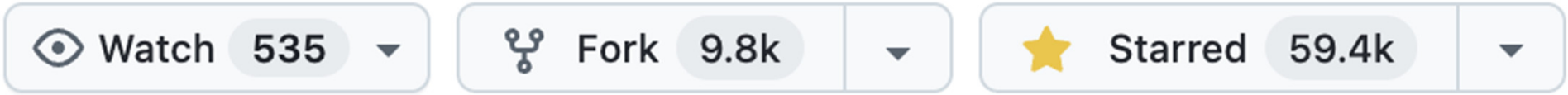
影响力

arXiv

https://arxiv.org › cs · 翻译此页

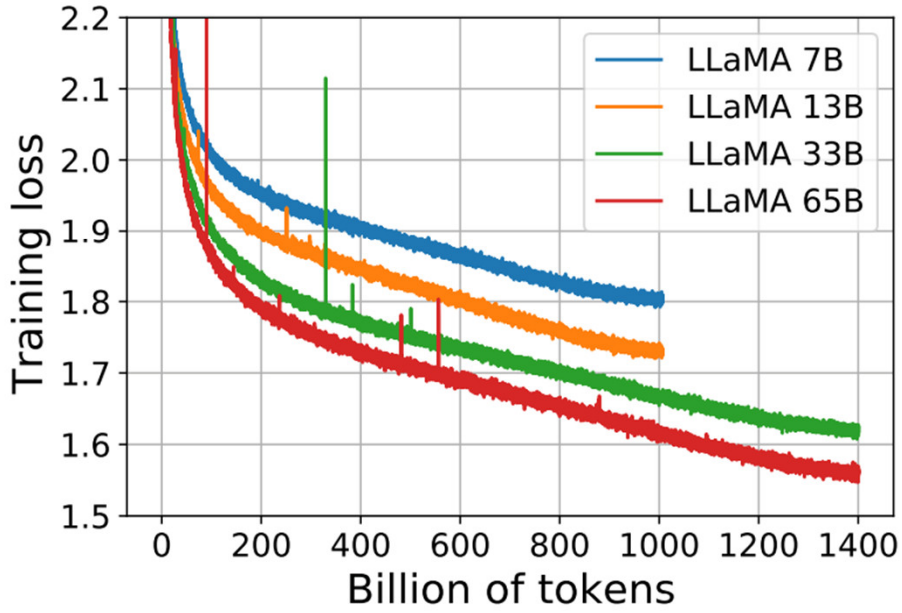
LLaMA: Open and Efficient Foundation Language Models

作者： H Touvron · 2023 · 被引用次数： 26001 — We introduce LLaMA, a collection language models ranging from 7B to 65B parameters. We train our models on trillio



尺寸

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T



效果

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

LLaMA-Adapter — 高效多模态指令跟随

研究背景

将 LLM 转化为指令跟随器是热门方向（Alpaca、Vicuna 等）。

MiniGPT-4 和 LLaVA 引发多模态研究热潮：将语言指令模型扩展为多模态模型。

LLaMA-Adapter V2 以冻结的指令跟随 LLaMA-Adapter 为起点，通过在图像-文本对上优化视觉投影层进行微调。

■ 核心优势

仅引入 1,400 万个参数就实现了强大的多模态推理

仅占整个模型参数的 0.04%

结合专家模型无需大量多模态指令数据

LLaMA-Adapter V2 三大改进

■ 改进1：调整线性层的偏置

为所有 Transformer 线性层添加偏置 b 和比例因子 s

初始化：b = 0, s = 1

将指令跟随知识分布到整个 LLaMA 中

参数量仅占整体 0.04%

■ 改进2：不相交参数的联合训练\

字幕数据（500K 图像-文本）和指令数据（50K 指令）

分别优化不相交的可学习参数组

解决两种微调目标之间的干扰

■ 改进3：视觉知识的早期融合

将视觉标记分配给前 K 层（K < N−L）

防止图像提示影响后期指令跟随能力

LLaMA-Adapter — 技术细节

- 零初始化注意力（Zero-init Attention）
  - 冻结整个 LLaMA，引入轻量级适配器模块（120 万参数）
  - 适配器层应用于 LLaMA 较高的 Transformer 层
  - 通过学习零初始化的门控因子，自适应控制适应提示对词标记的贡献
  - 门控幅度在训练中逐渐增加，稳定注入指令跟随能力
- 简单的多模态变体
  - 处理图像时，使用预训练视觉编码器（如 CLIP）提取多尺度视觉特征
  - 特征聚合为全局视觉特征，通过可学习投影层对齐视觉与语言嵌入空间
  - 全局视觉特征逐元素添加到较高 Transformer 层每个适应提示中
- 与专家模型集成（第4项技术）
  - 引入额外专家模型增强图像理解能力（如字幕专家、OCR 专家）
  - 默认实现：COCO Caption 上微调的 LLaMA-Adapter 作为字幕专家
  - 任何图到文本模型都可作为专家模型，增强灵活性
  - 无需大量图像-文本对训练数据，大幅降低数据需求

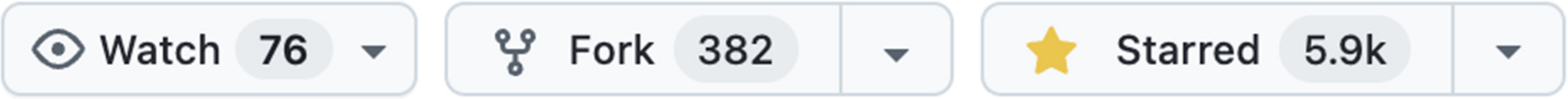
影响力

arXiv

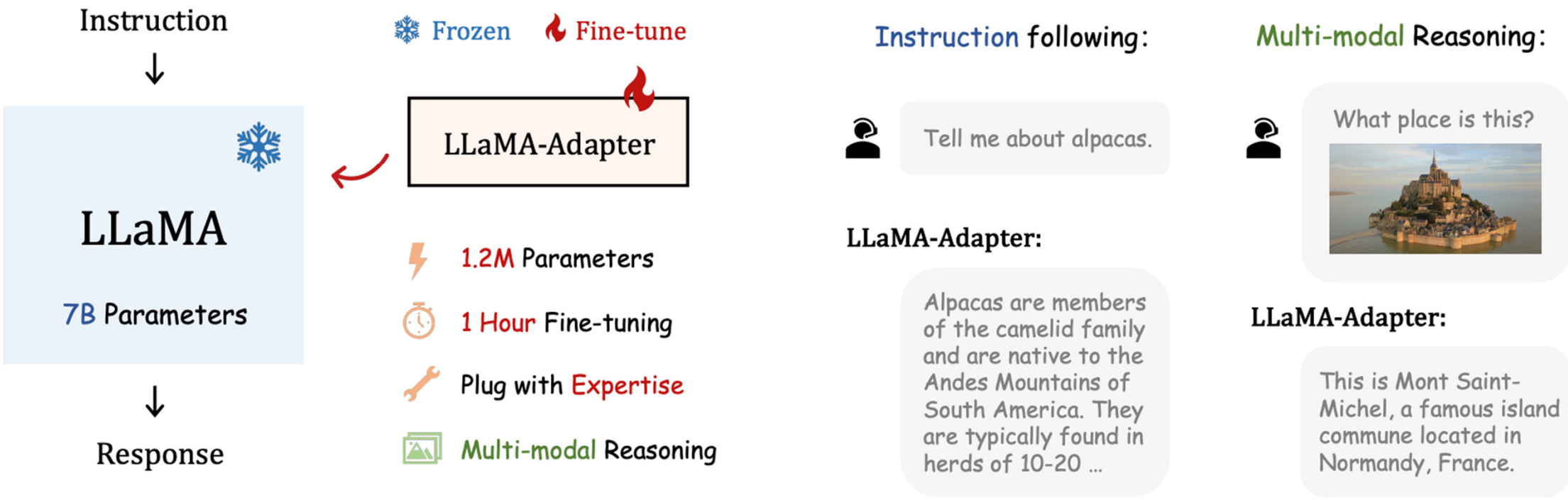
https://arxiv.org › cs · 翻译此页

LLaMA-Adapter: Efficient Fine-tuning of Language Models

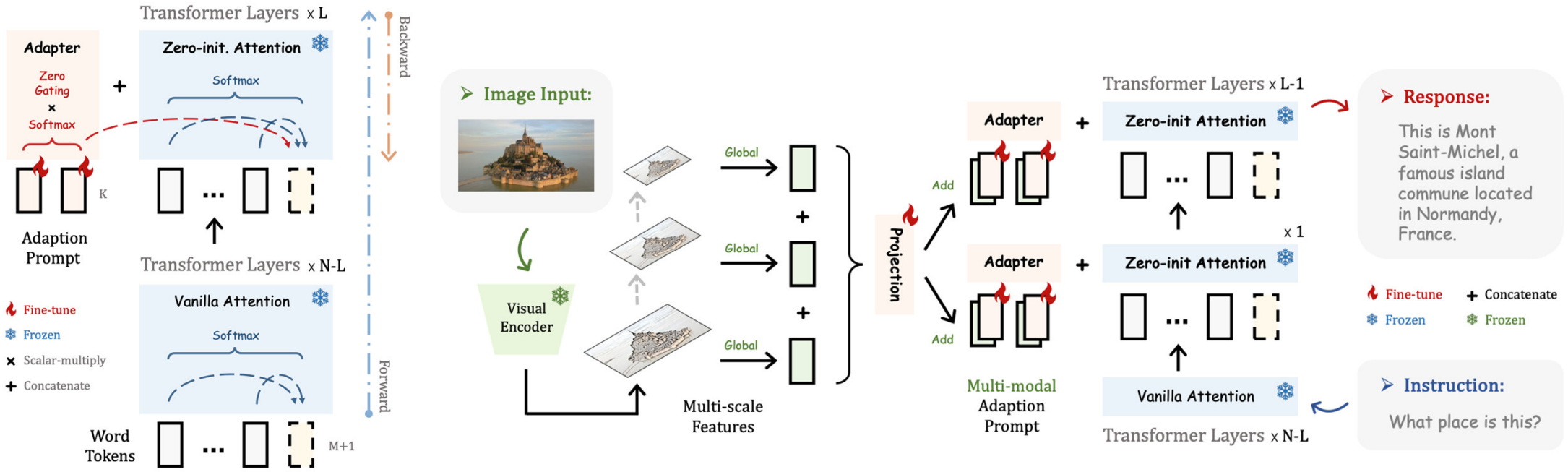
作者：R Zhang · 2023 · 被引用次数：1110 — Abstract:We present LLaMA-Adapter, adaption method to efficiently fine-tune LLaMA into an instruction-following model.



方法


Instruction
LLaMA
7B Parameters
Response
Frozen
Fine-tune
LLaMA-Adapter
1.2M Parameters
1 Hour Fine-tuning
Plug with Expertise
Multi-modal Reasoning
Instruction following:
Tell me about alpacas.
LLaMA-Adapter:
Alpacas are members of the camelid family and are native to the Andes Mountains of South America. They are typically found in herds of 10-20 ...
Multi-modal Reasoning:
What place is this?
LLaMA-Adapter:
This is Mont Saint-Michel, a famous island commune located in Normandy, France.

方法


Transformer Layers x L
Adapter
Zero Gating
×
Softmax
+
Zero-init. Attention
Softmax
K
Adaption Prompt
Transformer Layers x N-L
Vanilla Attention
Softmax
Word Tokens
M+1
Backward
Forward
Fine-tune
Frozen
Scalar-multiply
Concatenate
Image Input:
Visual Encoder
Global
Global
Global
Multi-scale Features
Projection
Add
Add
Adapter
Adapter
Multi-modal Adaption Prompt
Transformer Layers x L-1
Zero-init Attention
Zero-init Attention
x 1
Vanilla Attention
Transformer Layers x N-L
Response:
This is Mont Saint-Michel, a famous island commune located in Normandy, France.
Fine-tune
Concatenate
Frozen
Frozen
Instruction:
What place is this?

效果

Figure 5: **GPT-4 Evaluating Benchmark (Chiang et al., 2023)** for LLaMA-Adapter, Alpaca and Alpaca-LoRA.

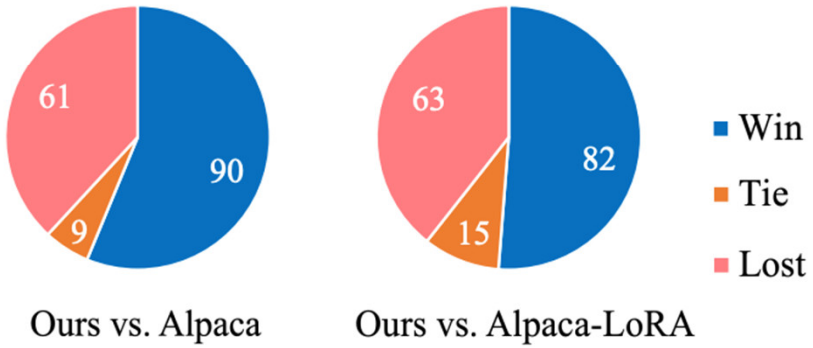


Table 1: **Efficiency Comparison.** The training time is tested on 8 A100 GPUs.

Model	Tuned Params	Storage Space	Training Time
Alpaca	7B	13G	3 hours
Alpaca-LoRA	4.2M	16.8M	1.5 hours
LLaMA-Adapter	**1.2M**	**4.7M**	**1 hour**

效果

Model	Tuned Params	Avg	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
Random Choice (Lu et al., 2022)	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human (Lu et al., 2022)	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
ChatGPT$_{CoT}$ (OpenAI, 2023a)	0M	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
GPT-4$_{CoT}$ (OpenAI, 2023b)	0M	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
MCAN (Yu et al., 2019)	95M	54.54	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72
VisualBERT (Li et al., 2019a; 2020)	111M	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA (Khashabi et al., 2020)	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA$_{CoT}$	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
MM-COT$_T$ (Zhang et al., 2023e)	223M	70.53	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68
MM-COT	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
LLaMA-Adapter$_T$	1.2M	78.31	79.00	73.79	80.55	78.30	70.35	83.14	79.77	75.68
LLaMA-Adapter	1.8M	85.19	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05

本节内容

CONTENTS

VideoChat — 以视频为中心的对话系统

研究动机

视频是最接近人类持续感知世界的表达方式，对人机交互、自动驾驶、智能监控等至关重要。

- 当前挑战

  现有视频理解受限于针对特定任务调整基础模型，无法满足通用时空理解需求

  将视频转文本不可避免丢失视觉信息，过度简化时间复杂性

  难以进行时间推理、事件定位和因果关系推断

- VideoChat 方案

  可学习的神经接口将视频基础模型与 LLM 结合

  两阶段轻量级训练（对齐 + 指令调优）

  在时间推理、事件定位和因果关系推断方面表现出色

VideoChat 两种系统

- VideoChat-Text

  感知工具：InternVideo / Whisper / Tag2Text / GRiT

  以 1 FPS 将视频文本化（约 2s 处理 10s 视频）

  生成带时间戳的综合视频文本描述

  局限：文本媒介限制感知模型的表示能力

- VideoChat-Embed（端到端）

  将视频和语言基础模型通过可学习 VLTI 结合

  基于 BLIP-2 和 StableVicuna

  两阶段训练：对齐阶段 + 指令调优阶段

  在更高级时间任务分配中性能更强

VideoChat-Embed — 端到端视频理解

- 模型结构
  - 预训练的 ViT-G 与全局多头关系聚合器（GMHRA）结合
  - GMHRA：InternVideo 和 UniFormerV2 中使用的时间建模模块
  - 通过额外线性投影补充，添加额外查询 Token 建模视频上下文
- VLTI（视频-语言令牌接口）
  - 采用 Q-Former 压缩视频 Token，减少冗余
  - Video Token、用户查询、对话上下文统一输入 LLM
- 两阶段训练
  - 阶段一（对齐）：2500 万视觉-文本对（1000 万视频 + 1500 万图像-文本）
  - 阶段二（指令调优）：7,000 个详细视频描述 + 4,000 个视频对话，共训练 3 轮
- 指令数据特点
  - 视频对话由 ChatGPT 基于 VideoChat-Text 生成的视频文本生成
  - 时序采样提示："该视频包含在 t0, t1, ..., tT 秒处抽样的 T 帧图像"
  - 包含时序推理的抽样信息，增强时空理解能力

影响力

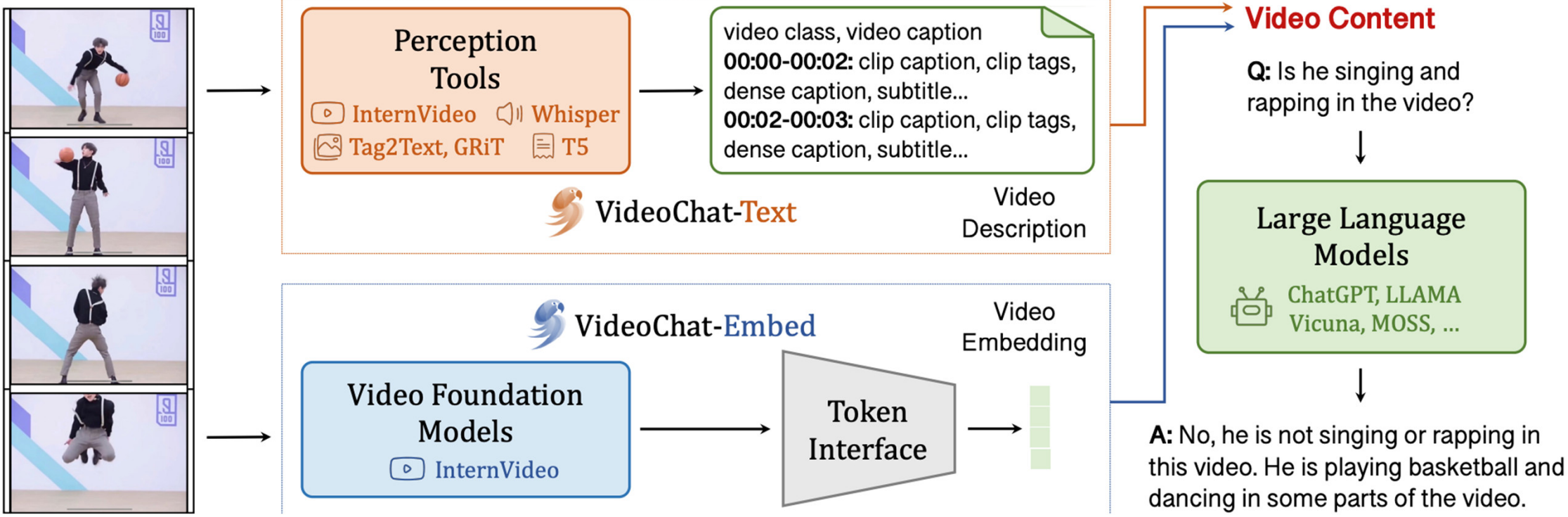
arXiv

https://arxiv.org › cs · 翻译此页

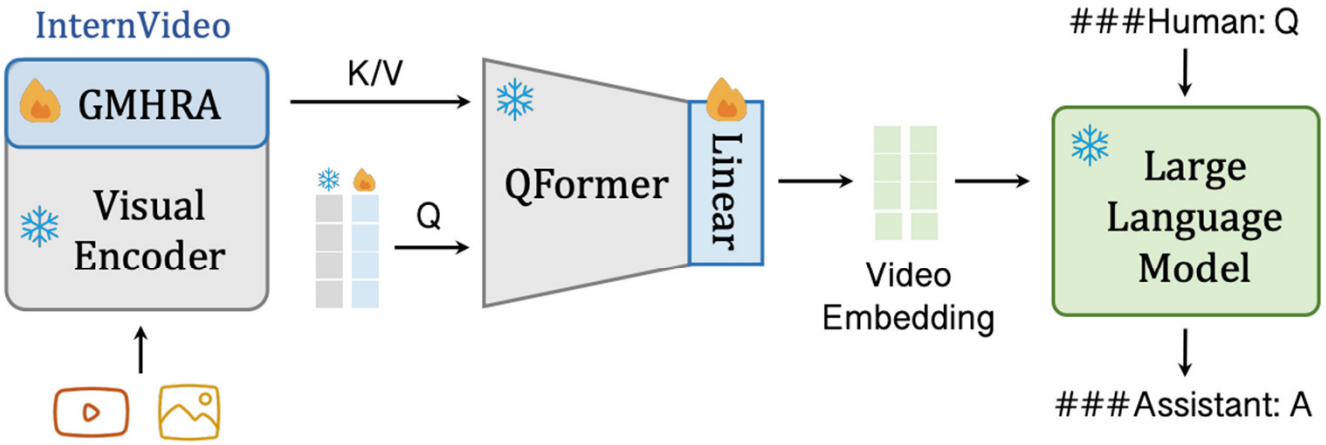
[2305.06355] VideoChat: Chat-Centric Video Understanding

作者：KC Li · 2023 · 被引用次数：1234 — Abstract:In this paper, we initiate an attemp end-to-end chat-centric video understanding system, coined as VideoChat.

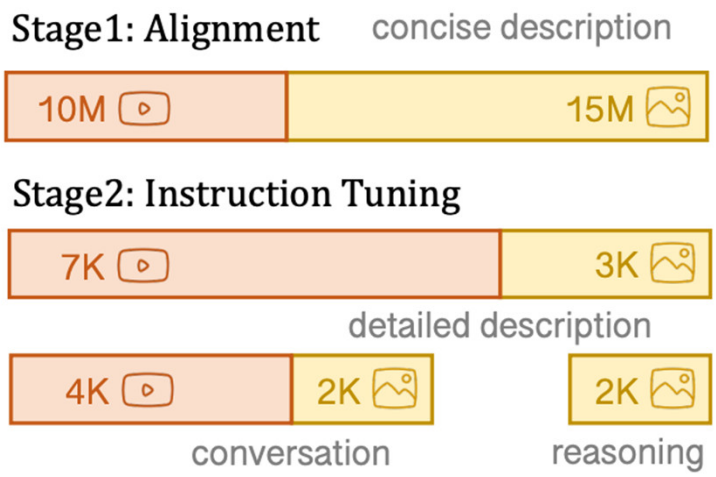
方法


Perception Tools
InternVideo
Whisper
Tag2Text, GRiT
T5
video class, video caption
00:00-00:02: clip caption, clip tags, dense caption, subtitle...
00:02-00:03: clip caption, clip tags, dense caption, subtitle...
VideoChat-Text
Video Description
VideoChat-Embed
Video Embedding
Video Foundation Models
InternVideo
Token Interface
Video Content
Q: Is he singing and rapping in the video?
Large Language Models
ChatGPT, LLAMA
Vicuna, MOSS, ...
A: No, he is not singing or rapping in this video. He is playing basketball and dancing in some parts of the video.

方法



(a) Architecture



(b) Data

模板

Video Class, Video Caption
00:00-00:02 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...
00:02-00:03 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...
00:03-00:06 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...
...

answering questions, a man and a woman sitting on a couch in a living room with a table in front of them.
00:00-00:11 a man and a girl sitting on a couch in a living room.
a lamp with a white shadea woman sitting at a table: [446, 155, 710, 476]; man wearing a plaid shirt: [361, 44, 581, 337]; man sitting on couch: [10, 63, 324, 350]; the tie is grey: [441, 150, 486, 280]; a glass of beer: [38, 305, 77, 367]; a stack of magazines: [28, 350, 180, 394]; a white tablecloth: [0, 334, 626, 476]; stainless steel oven: [1, 55, 150, 142]; a brown tie on a man: [144, 168, 191, 270]; the couch is white: [0, 119, 730, 472]; a gray binder: [0, 377, 157, 411]; a white couch: [768, 350, 848, 477]; a lamp with a white shade: [582, 26, 713, 195];
00:00-00:02: Hey, Pheebs, you gonna have the rest of that Pop-Tart?
00:02-00:03: Pheebs?
00:03-00:09: Does anyone want the rest of this Pop-Tart?
00:09-00:11: Hey, I might.

系统提示词

You are a chatbot that conducts conversations based on video contexts. You mainly answer based on the given contexts, and you can also modify the content according to the tag information, and you can also answer the relevant knowledge of the person or object contained in the video. **The timing description is a description every $1/FPS$ second, so that you can convert it into time. When describing, please mainly refer to the timing description. Dense caption is to give content every five seconds, you can disambiguate them in timing.** But you don't create a video plot out of nothing.
Begin!
Video contexts in temporal order: `textualizing_videos`
Question: `question`

提示词

Give you a video of `origin_caption`. The content of the video in temporal order is: `textualizing_videos`. Please use the sequence adverbs "first", "next", "then" and "finally" to describe this video in detail, but don't mention the specific time. Give as many details as possible. Say everything you see. The description should be more than 150 words and less than 200 words.

Table 3: **Prompts for detailed video descriptions.** The `origin_caption` is generated from VideoChat-Text.

Fix the error in the given paragraph. Remove any repeating sentences, meaningless characters, not English sentences, and so on. Remove unnecessary repetition. Rewrite any incomplete sentences. Return directly the results without explanation. Return directly the input paragraph if it is already correct without explanation.

Table 4: **Prompt for post-processing.** Copied from miniGPT-4 [58].

效果

Language Assistants(3)

00:01:09,836 --> 00:01:11,201
(EXHALES)

00:01:15,442 --> 00:01:18,843
(Eve:)Do you think the guy who did this to me feels bad?

What was House sit on when Eve asked if the guy felt bad? According to the subtitle of a video, and choose the best answer matches the question.
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.

3 **C: A couch in a living room.**

4 00:01:15,442--> 00:01:18.843
(Eve:)Do you think the guy who did this to me feels bad?

Answer: **B: A desk in an office.**

Image Assistants(5)

Choose the best answer matches the question: What was House sit on when Eve asked if the guy felt bad?
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.

M The best answer is **A: The floor in an office.**

The correct answer is **B: A desk in an office.**

A couch in a living room.

D: A park bench.

Video Assistants(3)

Choose the best answer matches the question: What was House sit on when Eve asked if the guy felt bad?
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.

D: A park bench.

D: A park bench. In the video, House and Eve are sitting on a park bench in a park when Eve asks if the guy feels bad. There is a lake and a road nearby in the background.

效果

What's video of this?

This is a video of a young man singing with a microphone, wearing a suit and tie. He is standing in front of a white curtain and a stained glass window. The video lasts 10 seconds.

Usually a person watching this video will feel cheated, do you know why?

It depends on their expectations. If they were expecting a longer or more engaging video, they may feel cheated if they don't get that. However, if they just wanted a brief glimpse of the man singing and it satisfies their curiosity, they may not feel cheated at all.

Describe the video.

The video shows a young man dancing in a kitchen, surrounded by colorful cartoon characters. He wears a white shirt and blue pants and is dancing in front of the refrigerator. The video also shows the kitchen's interior with a wooden floor and appliances. In some frames, the man is dancing in front of the cereal box, cookies, and a can of peas. There are also other items, like a juice bottle, in the kitchen. The man dances joyfully, with a big smile on his face, and enjoys the music. Overall, the video captures the energy and enthusiasm of the dancer and creates a fun and lively atmosphere.

Explain why this video is funny.

The video is funny because the man's dancing is random and spontaneous, and he adds an element of humor to the ordinary kitchen setting. His dancing is unpredictable and silly, and it adds an element of surprise to the video. The man's joyful facial expressions and childlike gestures make the video amusing and entertaining to watch. Additionally, the juxtaposition of the man's dancing with the serious and mundane objects in the kitchen creates an interesting contrast and adds to the humor. The fact that the man is dancing in front of food items in a kitchen is also a nod to our daily lives and the fact that we often find ourselves in the kitchen dancing and singing along to music, making the video relatable and funny to watch.

问题和讨论