多模态预训练模型调研记录

目录

视频和文本模态数据集

文本-视频常用数据集（英文）：

MSR-VTT: 该数据集包含 10000 个视频片段（video clip），每个视频 10 s 到 32s，被分为训练，验证和测试集三部分。每个视频片段都被标注了大概 20 条英文句子。此外，MSR-VTT 还提供了每个视频的类别信息（共计 20 类），这个类别信息算是先验的，在测试集中也是已知的。同时，视频都是包含音频信息的。该数据库共计使用了四种机器翻译的评价指标，分别为：METEOR, BLEU@1-4,ROUGE-L,CIDEr。

数据集地址： http://ms-multimedia-challenge.com/2017/dataset

MSVD：包含 1970 个视频，每个视频剪辑包含 40 个句子，每个视频 1s 到 62s，平均时长为 9s，视频包含不同的人，动物，动作，场景等。每个视频由不同的人标注了多个句子，大约 41 annotated sentences per clip, 共有 80839 个 sentences, 平均每个句子有 8 个 words, 这些所有的句子中共包含近 16000 个 unique words。caption 中包括多国的语言进行描述，部分论文中采取只选用 laguage = english 的 caption 进行训练和测试。

数据集地址：https://www.cs.utexas.edu/users/ml/clamp/videoDescription/

LSMDC： 由 118081 个视频组成，每个视频的长度从 2 秒到 30 秒不等。视频是从 202 部电影中提取的。验证集包含 7408 个视频，测试集包含 1000 个电影视频

数据集地址：https://sites.google.com/site/describingmovies/download

ActivityNet：是目前视频动作分析方向最大的数据集，包含分类和检测两个任务。目前版本为 v1.3，包括 20000 个 Youtube 视频（训练集包含约 10000 个视频，验证集和测试集各包含约 5000 个视频），共计约 700 小时的视频，平均每个视频上有 1.5 个动作标注（action instance）。ActivityNet 涵盖了 200 种不同的日常活动，例如：’ walking the dog’ ， ‘long jump’ , and ‘vacuuming floor’ 等。数据量分布：train（~50%）, validation（~25%）, test（~25%）。

数据集地址：http://activity-net.org/download.html

DiDeMo： 包含 10000 个视频和 40000 个句子。为了便于注释，每个视频被分成 5 秒的时间块。第一个时间块对应视频中的 0-5 秒，第二个时间块则对应 5-10 秒，等等。

数据集地址：https://github.com/LisaAnne/LocalizingMoments

VATEX： 全称 Video And TEXt，是一个大规模、多语言视频描述数据集，该数据集包含超过 41250 个视频和 82.5 万中英文视频描述，其中包括超过 20.6 万描述是中英平行翻译对。它包含 600 种人类活动和不同的视频内容。每个视频具备 10 个英文描述和 10 个中文描述，分别来自 20 个人类标注者。。

数据集地址：https://eric-xw.github.io/vatex-website/download.html

文本-视频常用数据集（中文）：

Youku-mPLUG：含来自 45 个不同类别的 1000 万个中文视频-文本对。视频标题长度在 5 到 30 个字之间，而且至少包含 5 个汉字，长度在 10 到 120 秒之间。
数据集地址：https://modelscope.cn/datasets/modelscope/Youku-AliceMind/summary

Tencent-MVSE Dataset： 中文视频多模态相似度数据集，包含视频对，和视频文本对，每一个视频都提供了丰富的信息,包括中文标题、ASR 文本(视频语音转化成文本)、视频 frame 特征、人工标注的 tag 以及视频所属的 category（垂类)。
数据集地址：https://tencent-mvse.github.io/

其他训练可用数据集（英文）：

Howto100M 数据集： https://opendatalab.org.cn/HowTo100M/download (质量比较差)
LAION-5B 数据集：https://laion.ai/blog/laion-5b/
Web-Vid-10M 数据集：https://m-bain.github.io/webvid-dataset/
HD-VILA-100M 数据集 :https://github.com/microsoft/XPretrain/tree/main/hd-vila-100m
UCF-101 数据集：https://www.crcv.ucf.edu/data/UCF101
YT-Temporal-180M 数据集： https://opendatalab.com/YT-Temporal-180M

图像-文本可用数据集：
LAION-400M 图文数据集：https://laion.ai/blog/laion-400-open-dataset/
COYO-700M 图文数据集： https://github.com/kakaobrain/coyo-dataset
MMC4-1000M 图文数据集：https://github.com/allenai/mmc4
SBU-图文数据集：https://opendatalab.org.cn/SBU_Captions_Dataset/download
Conceptual Captions-12M：https://opendatalab.org.cn/Conceptual_Captions/download
COCO 数据集：http://images.cocodataset.org/
GRIT- 20M 数据集 :https://github.com/microsoft/unilm/tree/master/kosmos-2 (文本描述包含物体的位置坐标，可以增强模型的位置感知能力，参考论文 kosmos-2)

1. 图像文本模态方法

（1）**CLIP** ： Learning Transferable Visual Models From Natural Language Supervision

Paper：https://arxiv.org/pdf/2103.00020.pdf

（2）**FLIP** ：Scaling Language-Image Pre-training via Masking

Paper：https://arxiv.org/pdf/2212.00794.pdf

亮点：通过随机 mask 掉图像的 patch,来极大的提高训练效率

（3）**ViLT**：Vision-and-Language Transformer Without Convolution or Region Supervision

Paper：https://arxiv.org/pdf/2102.03334.pdf

亮点：直接把 text tokens 和 image patches 怼到一块过 Transformer，大力出奇迹

（4）**ALBEF**：Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Paper：https://arxiv.org/pdf/2107.07651.pdf

亮点：做 N 多个任务，多个网络，一次迭代要前传很多次

（5）**BLIP**：BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

Paper：https://arxiv.org/pdf/2201.12086.pdf

亮点：利用训练好的 LM 生成伪标签、训练好的 ITM 清洗数据，然后再训一轮

（6） **BLIP2**：BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Paper：https://arxiv.org/pdf/2301.12597.pdf

亮点：利用 Q-former 桥接视觉模型和语言模型（LLM）

（7）**CoCa**：Contrastive Captioners are Image-Text Foundation Models

Paper：https://arxiv.org/pdf/2205.01917.pdf

亮点：对比学习+captioning

（7）**VLMO**: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts

Paper：https://arxiv.org/pdf/2111.02358.pdf

亮点：魔改 transformer layer，self-attention 部分不动；FFN 部分分成三： V-FFN 处理视觉模态，L-FFN 处理文本模态，VL-FFN 处理多模态

（7）**BEiT-V3**：Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks

Paper：https://arxiv.org/pdf/2208.10442.pdf

亮点： backbone 用 VLMO 里提出的 Multiway Transformer，训练目标只有 Masked Data Modeling，没有对比学习了

（7）**Kosmos-1**：Language Is Not All You Need: Aligning Perception with Language Models

Paper：https://arxiv.org/pdf/2302.14045.pdf

亮点：一个多模态的 LLM，把所有模态的输入都展平成序列按规则拼接到一起输入到 Transformer Decoder 中。

（8）**Kosmos-2**：Grounding Multimodal Large Language Models to the World

Paper：https://komosarxiv.org/pdf/2306.14824.pdf

亮点 :提出了一个 Large-Scale Training Corpus of Grounded Image-Text Pairs（GRIT）数据集，大模型模型具备了感知物体位置的能力。

简要描述可参考：https://zhuanlan.zhihu.com/p/614964205?utm_id=0

2. 视频文本模态方法

目前已有的视频文本模态对齐方法有 CLIPBERT，CLIP4Clip，CLIP2Video，CLIPTV、TACo 等方法，在这里我们将简单的对他们做一些总结性的介绍。

2.1 TACo

TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment

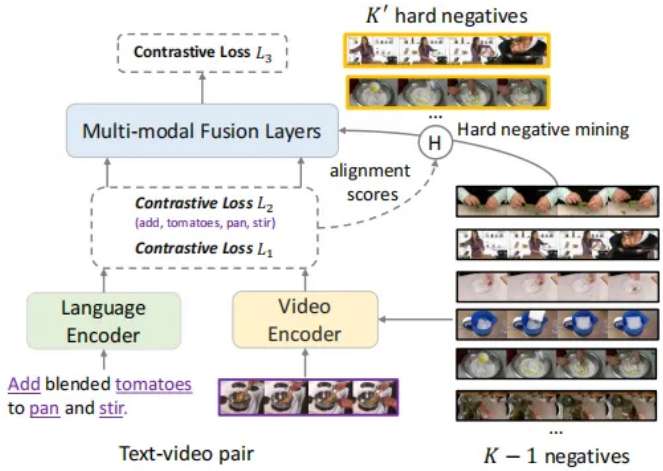
Jianwei Yang
Microsoft Research
jianwyan@microsoft.com

Yonatan Bisk
Carnegie Mellon University
ybisk@cs.cmu.edu

Jianfeng Gao
Microsoft Research
jfgao@microsoft.com

论文名：TACo: Token-aware Cascade Contrastive Learning for Video-Text Alignment
论文链接：https://arxiv.org/abs/2108.09980
代码链接: 无
论文机构：微软
论文创新：训练优化创新

本文提出了一种 token 感知级联对比学习(TACo) 算法，该算法利用两种新技术改进了目前的对比学习。 **第一个是 token 感知的对比损失** ，它是通过考虑单词的句法类来计算的。这是由于对于视频-文本对，文本中的内容词，如名词和动词，比功能词更有可能与视频中的视觉内容对齐。**第二是采用了级联采样方法生成一组少量 hard negative 样本**，以有效地估计多模态融合层的损失。



本文的方法主要有三个模块组成：**Video encoding module、Language encoding module、Multi-modal fusion module**

【Video encoding module】视频编码模块由 θ 参数化的自注意层实现。输入的视频特征使用一些预先训练的模型提取，如 2D CNN 或 3D CNN。给定输入的视频嵌入，视频编码器从一个线性层开始，将它们投射到与自注意层相同的维度 d 上。作者用 m 个特征的序列来表示视频编码器的输出，。特征的数量 m 取决于采样帧率的选择和视频特征提取器的选择。

【Language encoding module】作者分别使用预训练的 tokenizer 和 BERT 对输入文

本进行 tokenize 和提取文本特征。给定一个原始句子，分别在开头和结尾追加一个“[CLS]”和“[SEP]”。在模型顶部，可以得到一个由 n 个文本特征组成的序列。这里保证了视频编码器的输出特征维数与语言编码器的特征相同。在训练过程中，更新语言编码器中的参数 θ，以适应特定域的文本。

【Multi-modal fusion module】多模态融合模块由具有可学习参数 θ 的自注意层组成。它将两种独立模态的视频特征和文本特征作为输入，然后输出特征。

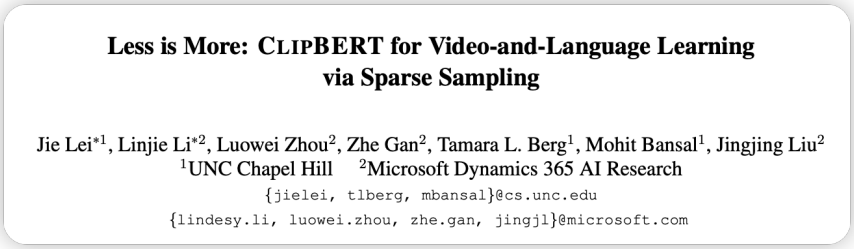
上述三个模块组成了本文的视频-文本对齐模型，然后对该模型使用所提出的 token 感知级联对比损失进行训练。

sentence-level contrastive loss：对整个视频帧编码后的特征求平均，对视频相应文本编码后的特征求平均，计算相似性损失

token-level contrastive loss：不求平均，计算损失，让文本特征和视频帧的特征对齐

Token of interest loss：挑选一些名词和动词的对应的 **Token** 出来计算损失，为了减少计算，只挑选 k 个最困难的负样本参与计算。

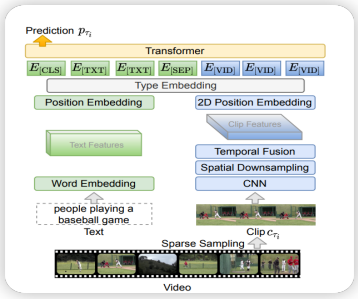
2.2 CLIPBERT

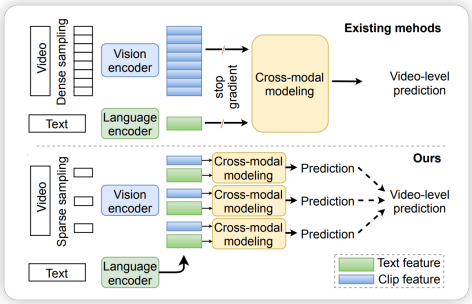


Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling

Jie Lei*[1], Linjie Li*[2], Luowei Zhou[2], Zhe Gan[2], Tamara L. Berg[1], Mohit Bansal[1], Jingjing Liu[2]
[1]UNC Chapel Hill [2]Microsoft Dynamics 365 AI Research
{jielei, tlberg, mbansal}@cs.unc.edu
{lindesy.li, luowei.zhou, zhe.gan, jingjl}@microsoft.com

论文名：CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval
论文链接：https://arxiv.org/pdf/2102.06183.pdf
代码链接：https://github.com/jayleicn/ClipBERT
论文机构：微软
论文创新：结构小改





来自 CVPR2021。作者的 motivation 来自于，目前大部分预训练模型都使用提前提取好的特征提取器，然而 1）固定的特征对于不同的下游任务来说不是最优的，且不同的模态的特征相互独立。2）密集的视频特征的计算量要求较高，以原视频作为输入太慢了，因此特征提取器很难参与到微调中。

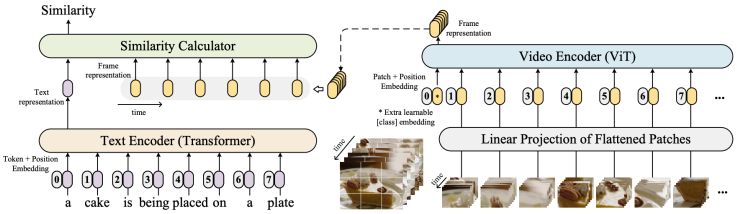
CLIPBERT，通过稀疏采样，即只使用一个或几个稀疏采样的视频短片段来代替整个视频，以 less-is-more 的原则使模型可以负载端到端学习。如上图所示，该模型仅仅使用少量的短片段即可，然后对多个短片段的预测进行融合如平均池化，以得到最终在整个视频级上的预测。这种先稀疏训练后密集推理的策略可以大大减少内存需求和计算量。

2.3 CLIP4Clip

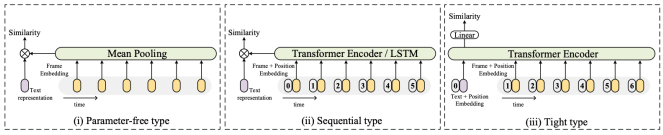
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval

Huaishao Luo[1]*, Lei Ji[2], Ming Zhong[3], Yang Chen[3], Wen Lei[3], Nan Duan[2], Tianrui Li[1]
[1]Southwest Jiaotong University, Chengdu, China
huaishaoluo@gmail.com, trli@swjtu.edu.cn
[2]Microsoft Research Asia, Beijing, China
[3]Microsoft STCA, Beijing, China
{leiji,minzhon,emchen,wen.lei,nanduan}@microsoft.com

论文名：CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval
论文链接：https://arxiv.org/pdf/2104.08860.pdf
代码链接: https://github.com/ArrowLuo/CLIP4Clip
论文机构：微软
论文创新：增加了用于视频和文本的相似度计算模块



(a) Main structure



(b) Similarity calculator

CLIP4Clip 基于预训练好的图文模型，引入 Similarity Calculator 来计算多帧特征和文本特征相似度，根据模块是否引入新参数进行学习，分为三类：无参数方法、序列型和紧凑型，结构分别如上图的下半部分。其中无参数方法直接使用平均池化直接融合视频表示。序列型的视频文本处理采用两个单独的分支。而紧凑型直接用 Transformer 学多模态交互。

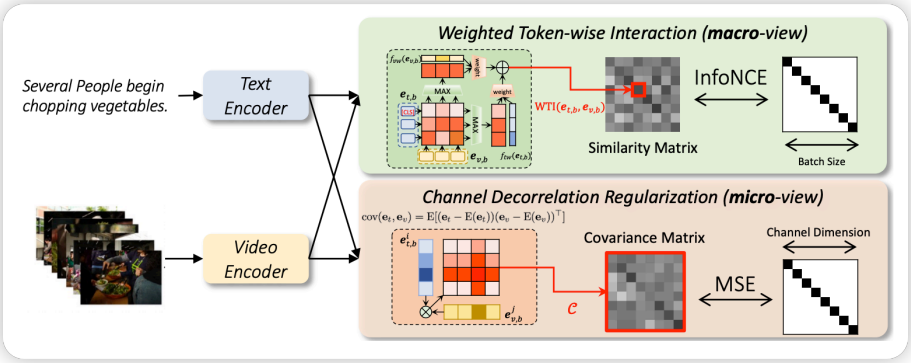
2.4 DRL

Disentangled Representation Learning for Text-Video Retrieval

Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, Xian-Sheng Hua

DAMO Academy, Alibaba Group
{qishi.wq, yanhao.zyh, zhengyun.zy, panpan.pp, xiansheng.hxs}@alibaba-inc.com

论文名：Disentangled Representation Learning for Text-Video Retrieval
论文链接：https://arxiv.org/pdf/2203.07111.pdf
代码链接: https://towhee.io/video-text-embedding/drl
论文机构：达摩院
论文创新：相似度计算改进



对 CLIP4Clip 的改进工作，之前的 CLIP4Clip 是对两个模态的总体表征计算相似度，缺少细粒度的交互，比如，描述可能只对应了视频的一部分帧，如果抽取整体特征可能其它帧的信息会占主导。DRL 使用提出两个重要改进，一个是 Weighted Token-wise Interaction，进行相似度的稠密预测，通过 max 操作找到潜在的激活的 token。另一个是 Channel Decorrelation Regularization，通道去相关正则可以减少通道间信息的冗余和竞争，使用协方差矩阵度量通道上的冗余。DRL 在大量视频检索数据集上取得了优于 CLIP4Clip 的效果。

2.5 CLIP2Video

CLIP2Video: Mastering Video-Text Retrieval via Image CLIP

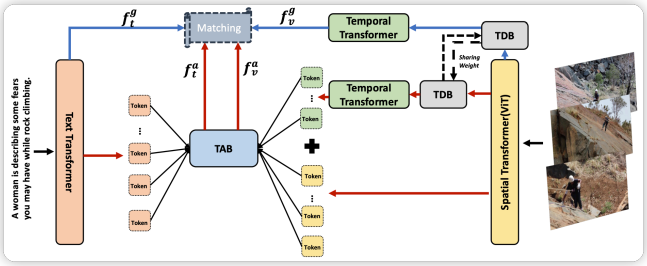
Han Fang* Pengfei Xiong*† Luhui Xu Yu Chen

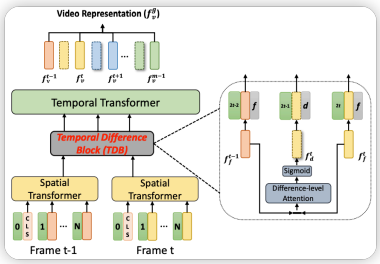
PCG, Tencent

fanghan@bupt.edu.cn, xiongpengfei2019@gmail.com, {lukenxu, andyyuchen}@tencent.com

https://github.com/CryhanFang/CLIP2Video

论文名：CLIP2Video: Mastering Video-Text Retrieval via Image CLIP
论文链接：https://arxiv.org/abs/2106.11097
代码链接: https://github.com/CryhanFang/CLIP2Video
论文机构：腾讯 PCG
论文创新：特征提取模块的改进





作者提出了两个模块： 时间差分块（Temporal Difference Block，TDB）和时间对齐块（Temporal Alignment Block，TAB）。**时间差分块。**在序列中加入图像帧的差分来模拟运动变化。具体来说，以相邻时间戳之间帧嵌入的变换差来表示，即使用 sigmoid 和差异的注意力来表示，最后全局拼接得到视频表征。**时间对齐块。**利用文本上下文和关键帧内容之间的对齐，以增强视频片段和短语之间的相关性。具体实现是使用共享的聚类中心来联合对齐帧和单词嵌入，即计算不同模态特征和共享中心的相关度作为不同 cluster 中心的权重。

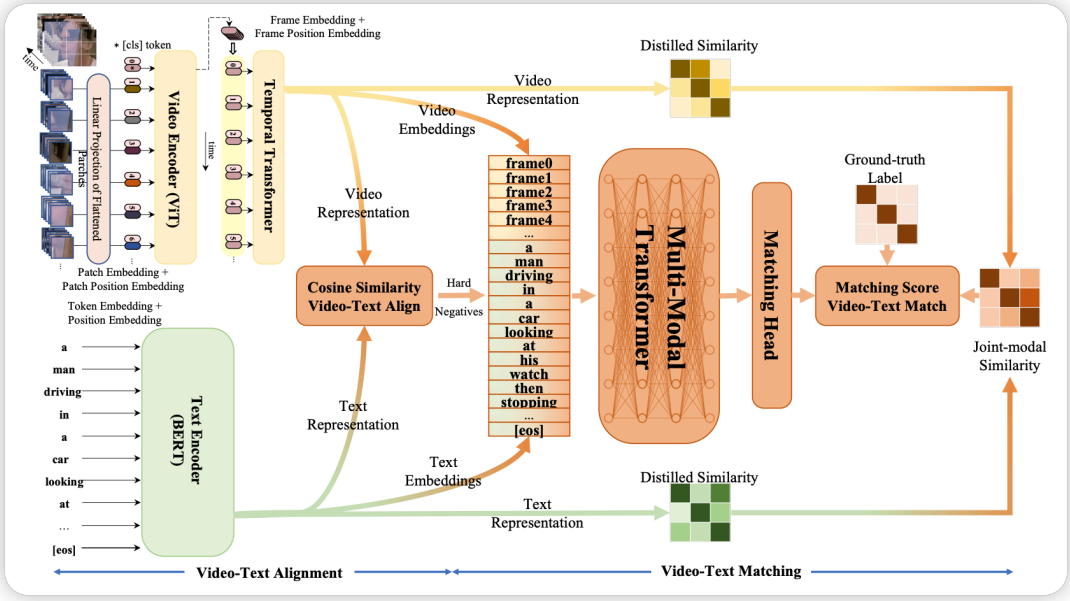
2.6 CLIP2TV

CLIP2TV: Align, Match and Distill for Video-Text Retrieval

Zijian Gao[1], Jingyu Liu[1], Weiqi Sun[1],
Sheng Chen[1], Dedan Chang[1], and Lili Zhao[1]

OVBU, PCG, Tencent
gzjbupt2016@bupt.edu.cn, sunweiqi@buaa.edu.cn,
{messijyliu, carlschen, dedanchang, lilillzhao}@tencent.com

论文名：CLIP2TV: Align, Match and Distill for Video-Text Retrieval
论文链接：https://arxiv.org/pdf/2111.05610.pdf
代码链接: 无
论文机构：腾讯 PCG
论文创新：特征提取模块的改进



结合 CLIP 和动量蒸馏来做视频文本检索。主要的贡献是在 CLIP4CLIP 的基础上，在推理阶段引入动量蒸馏。动量蒸馏的引入是为了解决图像文本的弱相关性，即标题不完全覆盖视频，视频片段又不包含文本描述。整体的结构如上图，主干部分和 CLIP4CLIP 一样，值得注意的点主要有：

Contrastive learning：由于帧特征和标题特征都被投影到了多模态共享空间中，作者试图结合余弦相似性和对比性损失，计算标准化帧表示和标准化标题表示之间的余弦相似度。

Momentum Distillation：动量蒸馏处理图像文本对之间的弱相关性。

3. 多模态方法(至少三种模态)

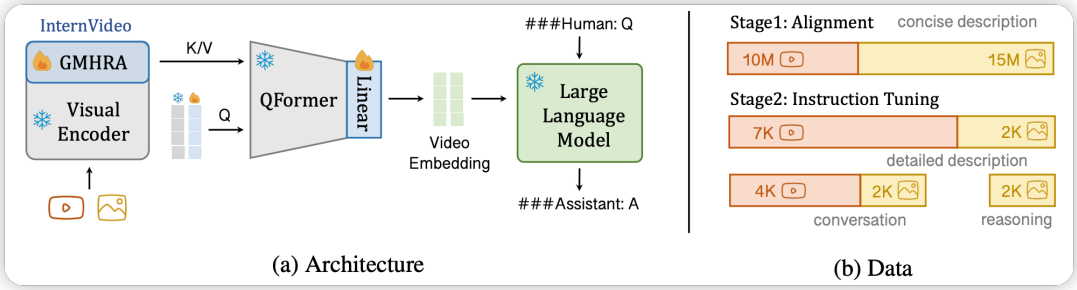
3.1 VideoChat(image、video、text 三模态)

VideoChat : Chat-Centric Video Understanding

KunChang Li[*1,4], Yinan He[*1], Yi Wang[*†1], Yizhuo Li[1,3], Wenhai Wang[1]
Ping Luo[3], Yali Wang[4,1], Limin Wang[2,1], Yu Qiao[1]
[1]OpenGVLab, Shanghai AI Laboratory [2]Nanjing University [3]The University of Hong Kong
[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

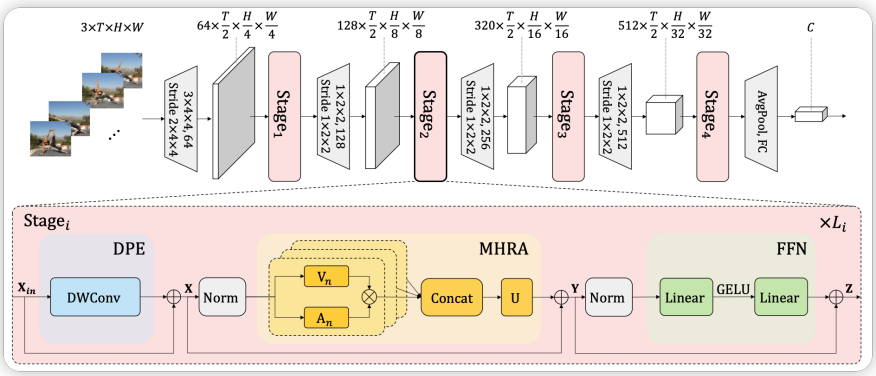
https://github.com/OpenGVLab/Ask-Anything

论文名：VideoChat : Chat-Centric Video Understanding
论文链接：https://arxiv.org/pdf/2305.06355.pdf
代码链接: https://github.com/OpenGVLab/Ask-Anything
论文机构：上海人工智能实验室
支持模态：视频、图像、文本



VideoChat 通过 BLIP2 (BLIP+UniFormerV2)和 LLM (Vicuna)构建了一个支持图像和视频输入的多模态对话系统。作者以一种极其简洁的方式统一了视频和图像模态的特征提取，从而实现了 BLIP 对图像、视频、文本三模态的支持。

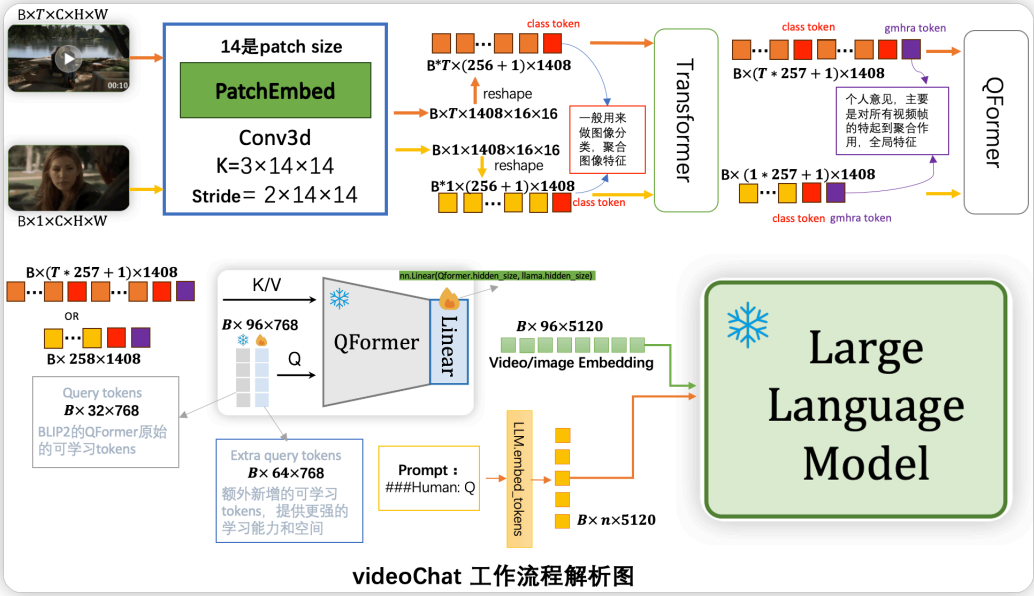
具体来说， 作者通过将 BLIP 的特征提取网络（普通 VIT）替换成了兼容图像和视频特征提取的 VIT 神经网络 **UniFormer，**其结构如下：



UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning

以下是 VideoChat 中图像和视频的处理流程：

1、对输入的图像和视频进行预处理，尺寸缩放到 224×224，颜色进行归一化处理。
2、利用 VIT-G (UniFormer) 网络进行图像和视频的特征提取，具体操作如下

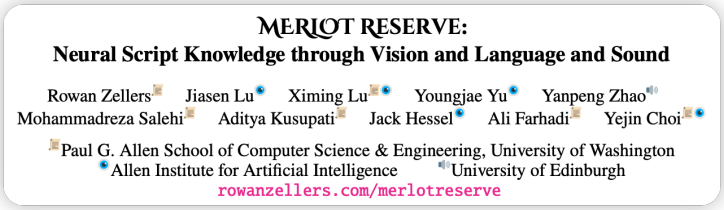


videoChat 工作流程解析图

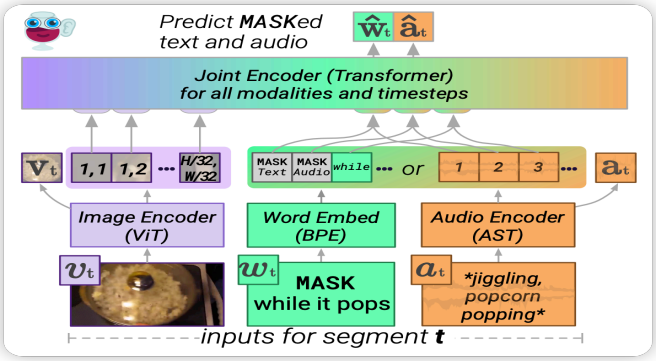
3、将 QFormer 编码后的图像或者视频特征与文本经过编码的特征拼接到一起送入 LLM 中

通过以上流程，VideoChat 以一种优雅的方式统一了图像和视频模态，从而基于训练好的 BLIP2 和 LLM 权重构建了一个视频、图像和文本的三模态对话系统。

3.2 MERLOT Reserve（视频、文本、音频）



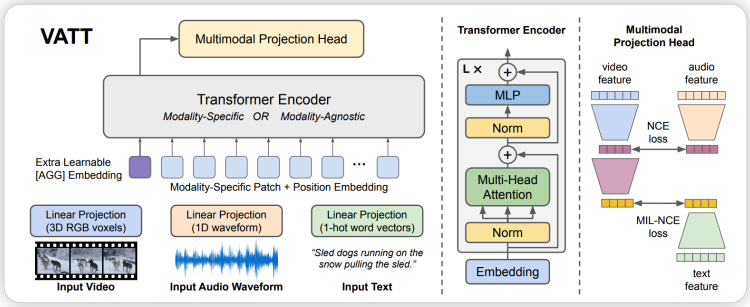
论文名：MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound
论文链接：https://arxiv.org/pdf/2201.02639.pdf
代码链接: https://rowanzellers.com/merlotreserve/
论文机构：华盛顿大学
支持模态：视频、图像(理论上支持)、文本、声音
创新类型：训练创新



作者使用了视频信息（帧）、文本信息（描述视频信息，即字幕）、语音信息（视频的音频部分），通过对模型的训练，能够实现给出视频的条件下，模型给出视频对应的文本以及语音信息的功能。将三种数据各自独立地编码后输入到联合编码器中，来将所有模态的数据融合在一起，并根据数据序列进行预测，恢复遮挡住的文本或者音频信息。

3.3 VATT（视频、文本、音频）

论文名：VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text
论文链接：https://arxiv.org/abs/2104.11178
代码链接: https://github.com/google-research/google-research/tree/master/vatt
论文机构：谷歌
支持模态：视频、文本、声音
创新类型：训练创新



VATT 直接对原始信号进行处理，视觉输入是 3 通道 RGB 像素视频帧，音频输入是初始波形，文本输入是单词序列。 对于视频片段，方法是将一整个 T*H*W 的视频片段划分为 [T/t]*[H/h]*[W/w]的小 Patch 序列，其中每个 patch 包含 t*h*w*3 个立体像素，然后对每个 patch 的所有像素应用一个线性投影，这个线性投影由一个可学习的权重执行。模型整体是通过 Transformer 提取不同模态的表示然后映射到公共空间，并通过对比损失来比较各个模态进行对齐。

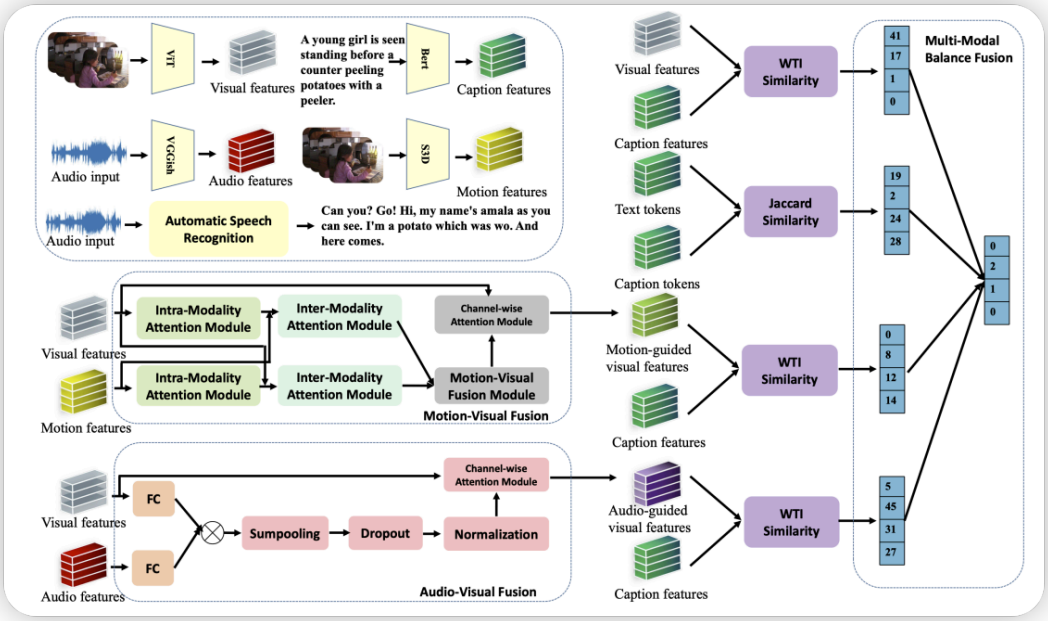
3.4 M2HF（视频、文本、音频）

M2HF: MULTI-LEVEL MULTI-MODAL HYBRID FUSION FOR TEXT-VIDEO RETRIEVAL

Shuo Liu*[1], Weize Quan[1], Ming Zhou[2], Sihong Chen†[3], Jian Kang[3], Zhe Zhao[3], CHEN CHEN[3] and Dong-Ming Yan‡[1]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Donghua University, Shanghai, China
[3]Tencent TEG AI, Shenzhen, China

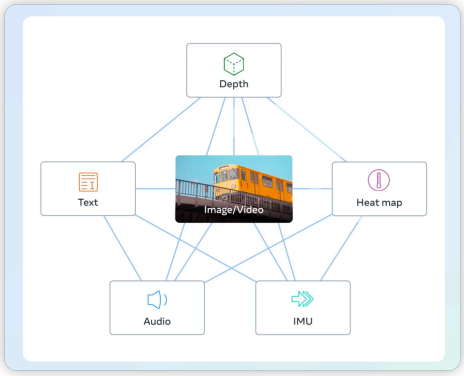
论文名：M2HF: Multi-level Multi-modal Hybrid Fusion for Text-Video Retrieval
论文链接：https://arxiv.org/abs/2208.07664
代码链接: 暂无
论文机构：中科院
支持模态：视频、文本、声音

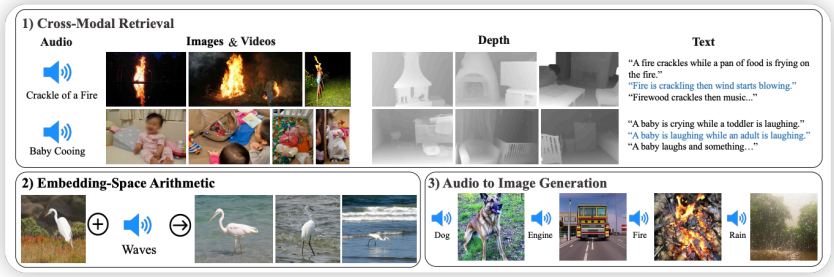


过建立语言对与从视频中提取的图像、音频、运动和文本之间的关系，设计了多层次的框架. 除此之外，作者还设计了一种后期多模态平衡融合方法，通过在各层次中选择最优排序结果进行融合来得到最终的排序结果。

3.5 ImageBind（图像、文本、音频、深度、热像和 IMU 数据）

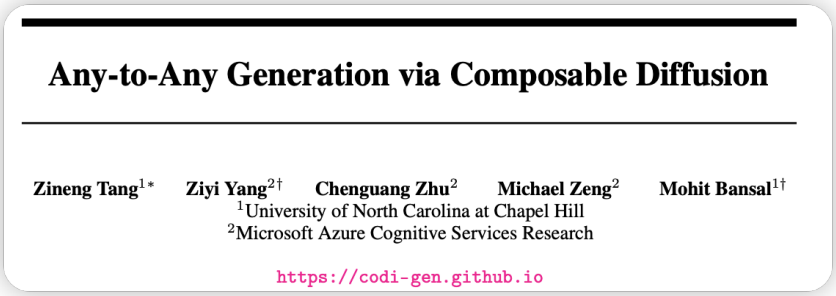
论文名：IMAGEBIND: One Embedding Space To Bind Them All
论文链接：https://arxiv.org/pdf/2305.05665.pdf
代码链接: https://github.com/facebookresearch/ImageBind
论文机构：facebook
支持模态：图像、文本、音频、深度、热像和 IMU 数据





Meta AI 提出了 ImageBind，它通过利用多种类型的图像配对数据来学习单个共享表示空间。该研究不需要所有模态相互同时出现的数据集，相反利用到了图像的绑定属性，只要将每个模态的嵌入与图像嵌入对齐，就会实现所有模态的迅速对齐。具体而言，ImageBind 利用网络规模（图像、文本）匹配数据，并将其与自然存在的配对数据（视频、音频、图像、深度）相结合，以学习单个联合嵌入空间。这样做使得 ImageBind 隐式地将文本嵌入与其他模态（如音频、深度等）对齐，从而在没有显式语义或文本配对的情况下，能在这些模态上实现零样本识别功能。通过将六种模态的嵌入对齐到一个公共空间，ImageBind 可以跨模态检索未同时观察到的不同类型的内容，添加不同模态的嵌入以自然地对它们的语义进行组合。将 ImageBind 与 LLM 结合就能很容易的构建一个多模态对话系统，目前已有开源的 ImageBind-LLM 系统（https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/imagebind_LLM）。

3.6 CoDi（视频、图像、文本、音频四模态到四模态生成模型）

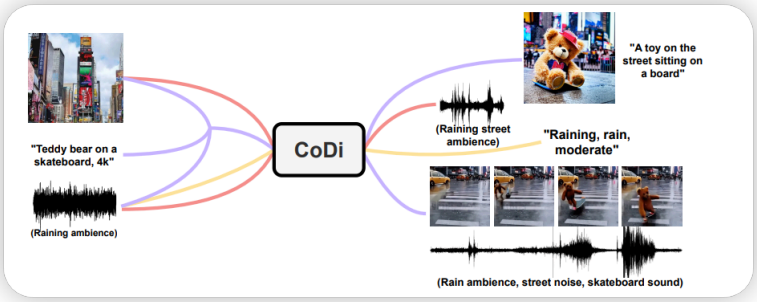


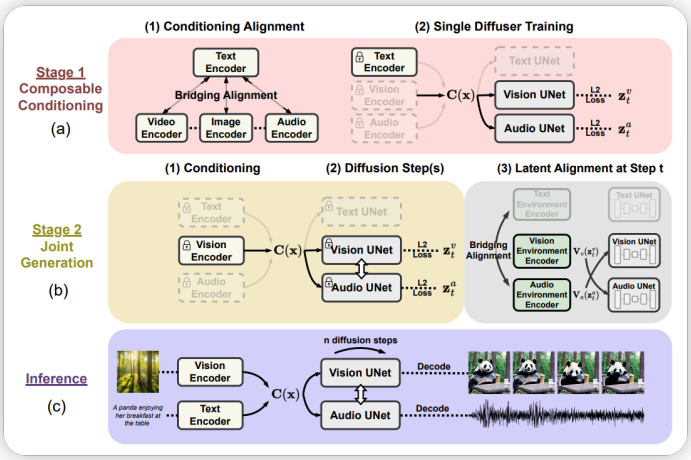
Any-to-Any Generation via Composable Diffusion

Zineng Tang[1*] **Ziyi Yang**[2†] **Chenguang Zhu**[2] **Michael Zeng**[2] **Mohit Bansal**[1†]
[1]University of North Carolina at Chapel Hill
[2]Microsoft Azure Cognitive Services Research

https://codi-gen.github.io

论文名：CoDi: Any-to-Any Generation via Composable Diffusion
论文链接：https://arxiv.org/abs/2305.11846
代码链接: https://github.com/microsoft/i-Code/tree/main/i-Code-V3
论文机构：微软
支持模态：图像、文本、音频到图像、文本、音频的生任务





这是一个多模态生成模型，输入视频，图片、语音、文本可以生成视频，图片、语音、文本。这个模型具有可以结合 LLM 构建一个强大的多模态对话系统的巨大潜力。