周报 2022.5.27

一、 完成工作

这周主要看了两篇论文，一篇是阿里2020ACM SIGIR会议上的文章，FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval，也是一篇跨模态检索的文章，是以Bert为backbone，以多任务进行训练的文章，主要是用于时尚领域的图片文本匹配。另一篇文章是Activity Image-to-Video Retrieval by Disentangling Appearance and Motion是上交和美团发表在AAAI上的跨模态检索文章，主要是用于非对称（自己提出的概念）跨模态检索（图片和视频）的文章。

1. FashionBERT:

主要是用于时尚领域的图片文本匹配，我看这篇文章主要是比较好奇这种用在一些特殊领域的是什么样的创新。

文章解决问题：文中说，用于时尚领域的多模态匹配和别的领域不同，需要更多的关注图像和文本中的细粒度信息，以前的方法都是在图像中检测 ROI，然后使用 ROI 嵌入作为图像表示，但是 ROI 倾向于“物体级”信息，而时尚文本倾向于描述更详细的信息，例如样式，属性，所以使用 ROI 不是很合适。

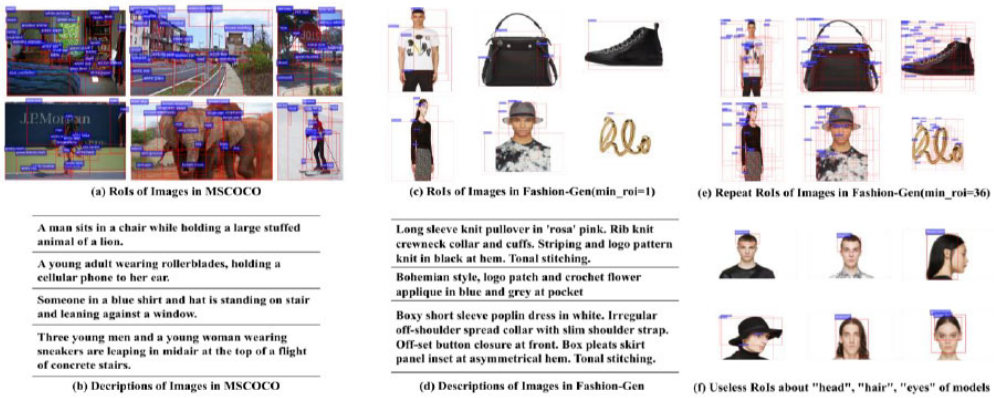


Figure 1: Comparison of text and image in the general and fashion domains. (a) and (b) are the RoIs and descriptions of MSCOCO Images from the general domain. (c) and (b) are the relatively-rare RoIs and fine-grained descriptions of Fashion-Gen Images from the fashion domain. (e) and (f) are large amount of the repeated and useless RoIs detected from fashion images.

问题 1：ROI 区域数量过多且有重复
问题 2：一些 ROI 区域对文本图像匹配没用（例如模特身体部位）
问题 3：ROI 可以知识主要物体但是无法区分描述中的细粒度属性和样式。

文章提出将图片分割成相同大小的 patch，这样可以解决 ROI 区域重复的问题，并且还是有序的，非常适合作为 BERT 模型的序列输入。除此之外，FashionBERT 训练过程是一个多任务学习过程，即 Masked Language Modeling、Masked Patch Modeling 和 Text&Image

Alignment)，文章提出了一种自适应算法来平衡每个任务的学习。

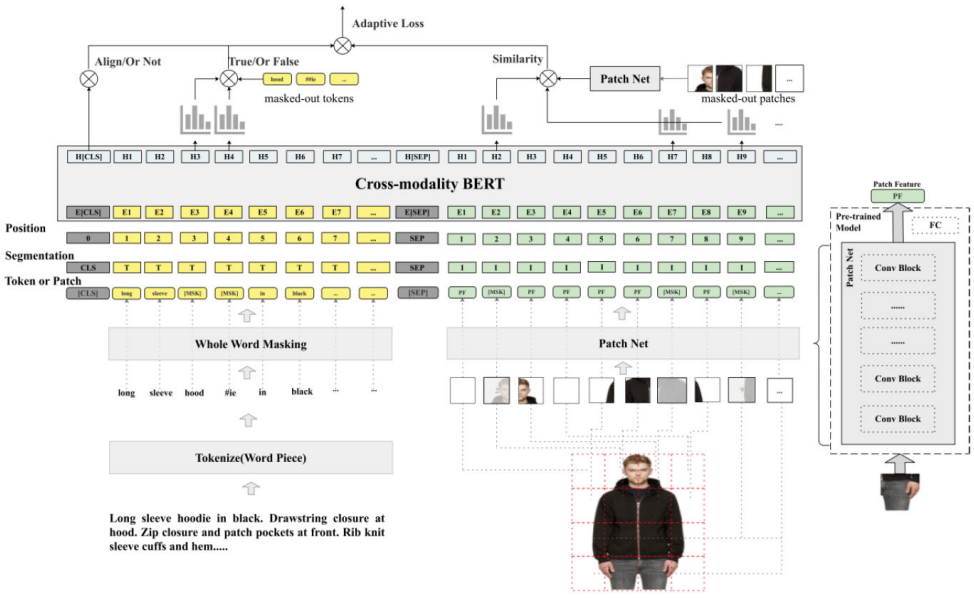
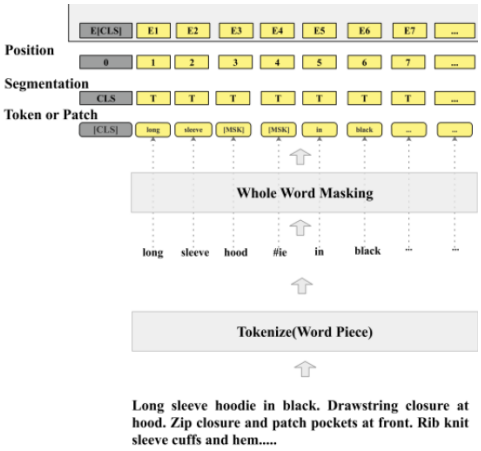
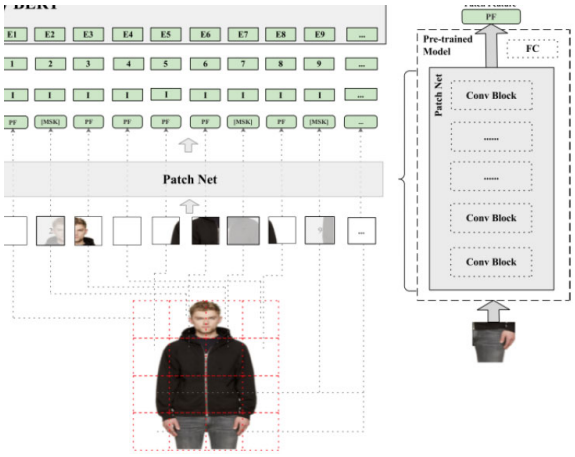


Figure 2: our FashionBERT framework for text and image matching. We cut each fashion image into patches and treat these patches as "image tokens". After the interaction of text tokens and image patches in BERT, three tasks with adaptive loss weights are proposed to train the entire FashionBERT model.

模型分为四个部分，主要是文本表示，图像表示,Matching Backbone 和 Adaptive Loss。其中文本表示可以看出来也就是标准的 bert，先进行 Word Piece ，然后对句子中的词做（全词）Masking，然后就和 bert 一样，生成位置 分割和 token。



文本表示



图片表示

图片表示则是将图片分割成 patch 来满足 bert 的序列输入，然后其他都和标准 bert 一样。

之后就是自适应 loss：

多任务 Adaptive Loss：

1.MLM 任务：也就是 mask 掉句子中的某个词，然后给定其他词来预测该词。

$$l_{MLM}(\theta) = -E_{t\sim D}\log P(t_i|t_{\setminus i},\theta)$$

2.MPM 任务：也就是 mask 掉 patch sequence 中的某个 patch，然后给定其他 patch 来预测该 patch。

$$l_{MPM}(\theta) = E_{KL_{p\sim D}}(Distr.(p_i|p_{\setminus i},\theta)|Distr.(p_i))$$

3.TIA 任务：也就是文本和图像数据是否匹配

$$l_{TIA}(\theta) = -E_{<t,p>\sim D}[y * \log P(\hat{y}| < t,p >,\theta) +$$

$$(1-y) * \log(1 - P(\hat{y}| < t,p >,\theta))]$$

而证明也相对简单，就是转化为优化问题，做一个拉格朗日问题出来，平衡多任务间重要性：

Adaptive:

$$\mathcal{L}(\theta) = \sum_{i=1}^{L} \omega_i l_i(\theta)$$

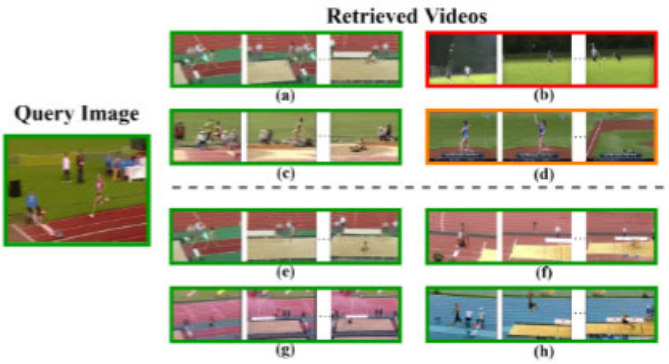
转化为优化问题:

$$\begin{aligned} &\operatorname{argmin} -\frac{1}{2}\sum_{i=1}^{L} \|\omega_i \nabla l_i\|^2 + \frac{1}{2}\sum_{i,j=1}^{L} \|\omega_i - \omega_j\|^2 \\ &s.t. \sum_{i-1}^{L} \omega_i = 1 \; and \; \exists \omega_i \geq 0 \end{aligned} \tag{5}$$

求得:

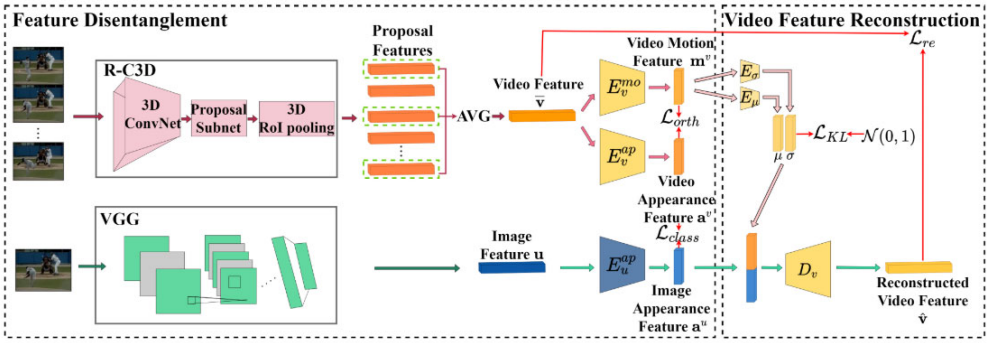
$$\omega_i^* = \frac{(L - \nabla {l_i}^2)^{-1}}{\sum_{i=1}^{L}(L - \nabla {l_i}^2)^{-1}}$$

2. Activity Image-to-Video Retrieval by Disentangling Appearance and Motion：这是美团和上交在 2021 年 AAAI 上发表的一篇非对称跨模态检索的图像视频跨模态检索文章，主要是来寻找与 query 图像“运动”相似的视频。相对于仅仅是在视频中查询跟图片物体相似的工作来说，寻找和图像运动相似的视频更困难一些。



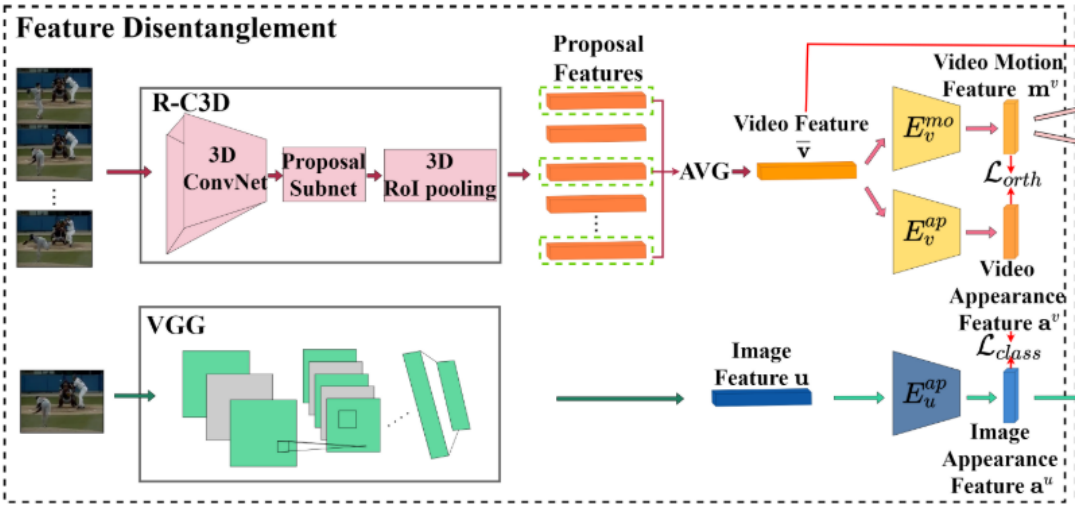
论文贡献:

1. 论文考虑到图像模态和视频模态的不对称关系，借助视频运动特征来促进 AIVR 任务。
2. 论文设计了一种基于运动辅助活动的图像到视频检索（MAP-IVR）方法来捕获图像和视频之间的活动相关性。



模型主要模块分为特征解缠模块和视频特征重建模块。

1. 特征解缠模块（也就是将视频特征分为 appearance feature 和 motion feature）



也就是模型前半部分，可以看到在特征提取部分，图片方面使用 VGG-16 来得到图片特征 u，然后视频方面，采用预训练的 R-C3D 来得到一些候选 proposal（可能包含活动）。然后 R-C3D 可以预测每个 proposal 对于所有活动的分数，然后使用所有活动类别中最大的置信度分数作为每个提案的置信度分数，并选择置信度分数最大的前 k 个 proposal。然后平均一下，得到视频特征。

然后就到了把视频特征分解为 appearance 特征和 motion 特征的时候了，除此之外，还需要将图片和视频的 appearance 在共享的 appearance 空间中对齐。所以，首先使用 emo 和 eap 将视频特征分解。然后再使用另一个 eap 将图像特征投影到外观特征空间，这样就可以将 au 和 av 在公共空间对齐。

因为需要分割开运动特征，所以使用正交约束来加强分割。

$$\mathcal{L}_{orth} = \cos(\mathbf{m}^v, \mathbf{a}^v),$$

也就是让他俩相似度越低越好，除此之外，还需要将公共空间中不同的分开

$$\mathcal{L}_{class} = -\log(p(\mathbf{a}^v)_y) - \log(p(\mathbf{a}^u)_y),$$

这里直接使用交叉熵。

2. Video Feature Reconstruction 重构模块

为了确保视频特征被准确的分为 motion 和 appearance，所以还需要能够重构回去。利用视频的运动表征和图像的外观表征，重建同种类视频的视频表征。为了支持测

试阶段的随机采样，类似于变分自编码器(VAE)的做法，我们把视频的运动表征映射到服从单位高斯分布的隐空间，叫做运动不确定性编码 (motion uncertainty code)。运动不确定性编码填补了基于图像表征重建视频表征所缺失的信息，消除了从图像表征转换到视频表征过程中的不确定性。这样在测试阶段，就能从单位高斯分布采样随机向量，和图像的外观表征结合，生成和图像同种类的视频表征。

$$\mathcal{L}_{KL} = KL(q_\phi(\mathbf{z}|\mathbf{m}^v)||p_\theta(\mathbf{z})),$$

然后在测试阶段：

给定一个图像 I 和一个视频 V，我们可以得到它们的外观表征，在外观表征空间计算相似度。另外，我们可以从单位高斯分布随机采样运动不确定性编码，和图像 I 的外观表征结合后生成视频表征，和视频 V 的视频表征计算相似度。由于采样的随机性，采样多个运动不确定性编码得到的视频表征各不相同，代表了从图像表征转换到视频表征的多种可能性。为了结果的合理性，采样多次，使用和 V 最接近的视频表征计算相似度。

$$S_A = 1 - \cos(\mathbf{a}^u, \mathbf{a}^v).$$

ure Space Given an image with featur
s appearance feature $\mathbf{a}^u$, we sample m
de from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ for h times, and co
n uncertainty code with $\mathbf{a}^u$, leading to
features $\{\hat{\mathbf{v}}_i | i = 1 \dots h\}$ containing
rmation. For the comparison between t
with feature $\bar{\mathbf{v}}$, we need to find out
feature $\hat{\mathbf{v}}_i$ which is closest to $\bar{\mathbf{v}}$. For
e distance as

$$S_V = \min_{i=1}^{h}(1 - \cos(\bar{\mathbf{v}}, \hat{\mathbf{v}}_i)).$$

ll advantage of two spaces, we perform
e weighted average of S_A and S_V:

$$S_{all} = (1 - \lambda_v)S_A + \lambda_v S_V,$$

这篇论文有一点问题就是视频运动特征一定就是与类别无关的吗，还有就是如何确定分割出来的一定是运动特征呢。。因为他只是可以重构回去，只能代表 appearance 和另一个特征可以重构回，但是不代表另一个一定是运动把。