开题报告

莫文韬

2021 年 5 月 20 日

目录

1 开题依据

Visual Question Answering(VQA) 作为一个多模态任务这几年受到了广泛的关注, 一种经典的方法是是使用模态之间的特征融合方法来预测答案, 视觉和语言特征分别地进行特征提取后进行特征融合来得到全局特征. 最近的方法试图利用 Co-Attention 机制 (MFH, MCAN, TRRNet etc.), 寻找并建模图像不同区域和问题不同词组之间的注意力和相关性, 来更好地得到多模态融合的特征 (用于回答问题). 进一步地, Transformer 作为顺序/特征种类无关的注意力机制被用于预训练的视觉-语言多模态特征学习的特征融合方法 (ViLBERT, UNITER, OSCAR etc.), 可以将任意多的特征信息 (除了视觉和文本特征, 还可以包括图像区域的语义标签的特征, 图像语义分割后的区域特征等等) 作为 Transformer 输入序列的 token 进行特征融合. 此外, 作为一种特殊的注意力机制, 若给定场景/对象/词之间的关联信息, GNN 也可以用于特征的融合. Co-Attention 可以视作二分图上的图注意力机制, 而 Transformers 可以看做是完全图上的图注意机制.

另一方面, 为了减少 Data Bias 和不同数据集之间的 Domain Gap, 也为了改进在预训练模型的性能, 一些新的视觉特征提取方法也被提出. 不同于广泛采用的 Faster RCNN 作为图片中对象和对象特征提取器的做法. Visual Commonsense RCNN 提出利用整个数据集中对象之间的因果关系来改进提取到的特征, 减少因为 Data Bias 导致的表示模型学到的对象之间的伪因果关系. VinVL 使用了超大数据集来与训练 Faster RCNN, 来得到一个好的特征预提取器, 防止 Domain Gap 导致

图 1: 一个典型的 VQA 任务特征融合模型的 Pipeline(Duplicated from MFB)

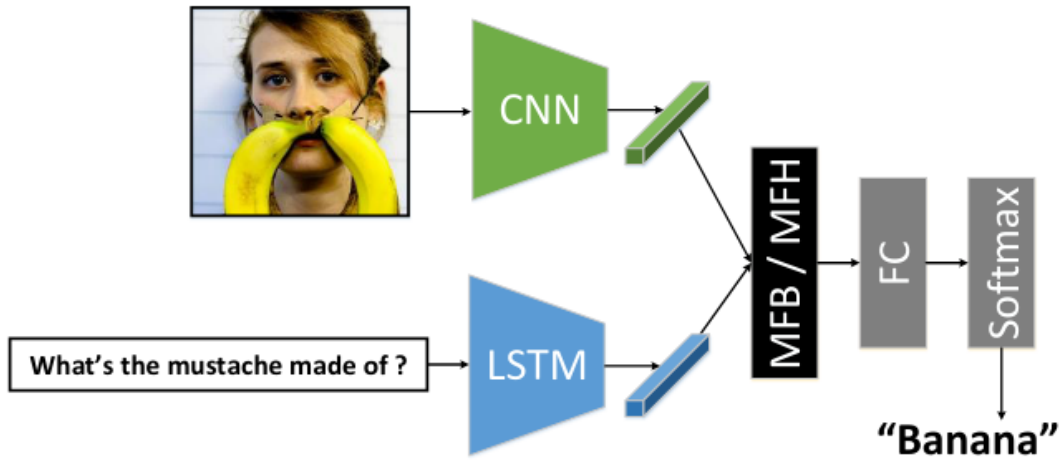


Fig. 4. The baseline network architecture with MFB or MFH and without the attention mechanism for VQA.

图 2: 利用 Transformer 进行特征融合 (Duplicated from OSCAR)

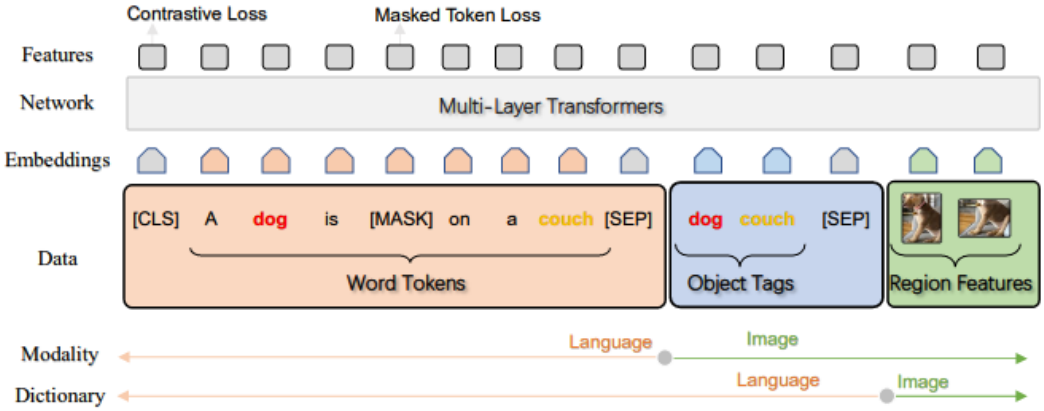
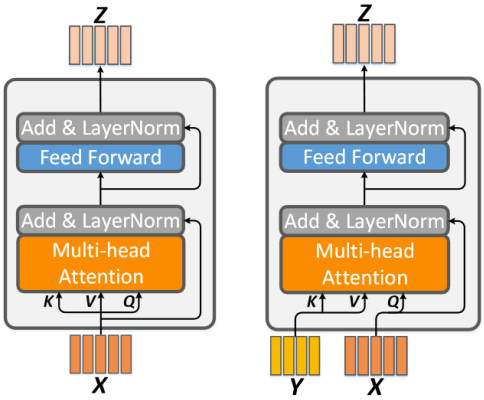
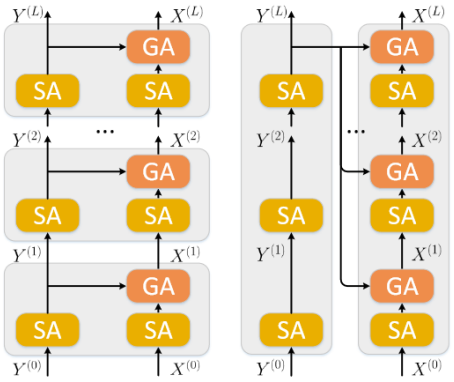


Fig. 3: Illustration of OSCAR. We represent the image-text pair as a triple [word tokens , object tags , region features], where the object tags (*e.g.*, "dog" or "couch") are proposed to align the cross-domain semantics; when removed, OSCAR reduces to previous VLP methods. The input triple can be understood from two perspectives: a *modality* view and a *dictionary* view.



(a) Self-Attention (SA) (b) Guided-Attention (GA)

Figure 2: Two basic attention units with multi-head attention for different types of inputs. SA takes one group of input features X and output the attended features Z for X; GA takes two groups of input features X and Y and output the attended features Z for X guided by Y.



(a) Stacking (b) Encoder-Decoder

Figure 5: Two deep co-attention models based on a cascade of MCA layers (*e.g.*, SA(Y)-SGA(X,Y)).

图 3: 利用 Co-Attention 进行特征融合 (Dubplicated from MCAN)

在和 Target Domain 不同视觉数据集/分布上的性能降低 (无法检测某些未见到的物体中类等等). 问题是, 无法捕捉不同区域之间的不同关系, 只是考虑二分的 Attention 和稠密的完全 Attention!

2 任务内容

图神经网络是一种自然的融合不同图像-问题中不同的区域的特征的方法 (这些区域可以是问题或对象标签的词嵌入, 也可以是图像的 Grids/Regions/Semantic Regions 特征等等), 它能考虑到不同区域之间的互动和关系 (或者说某种注意力). 作为稀疏版的 Transformers, GNN 可以大大降低模型计算复杂度, 并且具有 Transformers 可以具有任意多输入区域, 同时对于输入区域的顺序具有 Permutation Invariance. 作为一种带有额外信息的特征, 图的邻接矩阵如何获得是如何得到一个好的 GNN 融合的特征的关键. 自然的想法使用现成的方法检测场景中物体之间的互动/作用, 作为输入特征送入 GNN 进行融合. 端到端地学习邻接矩阵可能是另一种更优的解决方案. DGCNN 的工作证明了, 在网络计算过程中使用动态图 (邻接矩阵) 是可行 (可优化) 且高效的. 这可能意味着优化一个好的特征映射可以得到好的邻接矩阵, 所以使用端到端的动态图训练可能是一个好的选择.

Sample reweighting? GPNN? minimize L1-norm? k-thresholding

图 4: 在 Object Detector 中加入因果的额外信息的一种框架 (Duplicated from VC RCNN)

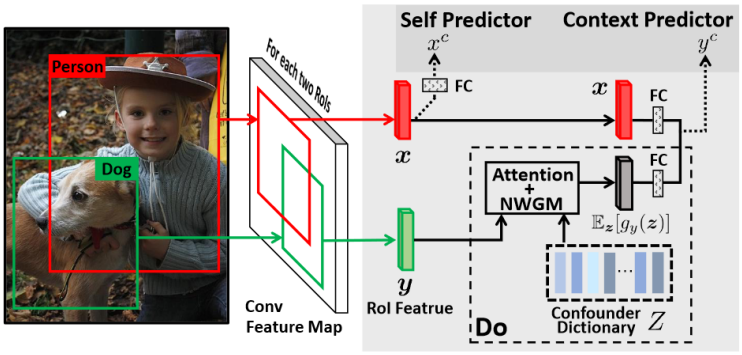


Figure 4. The overview of VC R-CNN. Any R-CNN backbone (*e.g.*, Faster R-CNN [54]) can be used to extract regions of interest (RoI) on the feature map. Each RoI is then fed into two sibling branches: a **Self Predictor** to predict its own class, *e.g.*, x^c, and a **Context Predictor** to predict its context labels, *e.g.*, y^c, with our **Do** calculus. The architecture is trained with a multi-task loss.

3 任务方案及可行性分析

Dynamic Graph Attention 不同于传统 GNN 的输入邻接矩阵是固定的, 我们可以采用动态建立的图来作为邻接矩阵, 具体地, 对于每一层选取特征空间上的 kNN 作为邻接点

$$a_{ij} = \mathbf{1}[d(i,j) \in topK(\{d(i,u)|u \in V\})] \tag{1}$$

其中距离函数 $d(i,j) = d(\boldsymbol{h}_i, \boldsymbol{h}_j)$, 以及每一层的 Forward Pass

$$x_i' = \frac{1}{|N(i)|} \sum_{j \in N(i)} f(s_i, s_j - s_i), s_i = \phi(x_i) \tag{2}$$

或者使用 Multi-head 注意力机制 (GAT)

$$\vec{h}_i' = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j \right) \tag{3}$$

其中注意力 coefficient 为

$$\alpha_{ij} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\sigma\left(\vec{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k\right]\right)\right)} \tag{4}$$

Semantic Label as Fusion Input

根据 OSCAR 模型的研究, 使用输入图像区域的标签作为 VQA 任务的特征融合提高了模型的性能和表达力, 并且可以自然地嵌入到上述的特征融合方法中作为一个图节点的输入. 我们可以选择使用预训练的词嵌入 (Word2Vec, GLoVe 等等) 来得到一个节点特征并且和图像区域特征和问题特征三者一起进行融合.

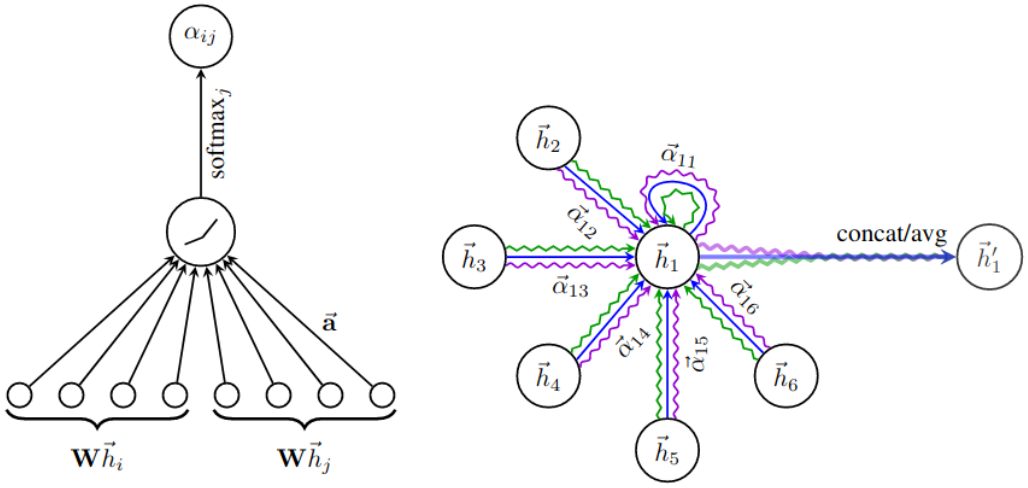


Figure 1: **Left:** The attention mechanism $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ employed by our model, parametrized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$, applying a LeakyReLU activation. **Right:** An illustration of multi-head attention (with $K = 3$ heads) by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain $\vec{h}'_1$.

4 Reference

1. Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." European Conference on Computer Vision. Springer, Cham, 2020.
2. Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." European Conference on Computer Vision. Springer, Cham, 2020.
3. Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
4. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
5. Wang, Yue, et al. "Dynamic graph cnn for learning on point clouds." Acm Transactions On Graphics (tog) 38.5 (2019): 1-12.
6. Veličković, Petar, et al. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).
7. Wang, Tan, et al. "Visual commonsense r-cnn." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
8. Zhang, Pengchuan, et al. "VinVL: Making Visual Representations Matter in Vision-Language Models." arXiv preprint arXiv:2101.00529 (2021).

9. Lu, Jiasen, et al. ”Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.” arXiv preprint arXiv:1908.02265 (2019).

10. Yang, Xiaofeng, et al. ”TRRNet: Tiered Relation Reasoning for Compositional Visual Question Answering.”

11. Yu, Zhou, et al. ”Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering.” IEEE transactions on neural networks and learning systems 29.12 (2018): 5947-5959.