基于 Spatial-DCTHash 动态参数网络的视觉问答算法

孟祥申, 江爱文*, 刘长红, 叶继华, 王明文

江西师范大学计算机与信息工程学院, 南昌 330022

* 通信作者. E-mail: jiangaiwen@jxnu.edu.cn

收稿日期: 2016–12–16; 接受日期: 2017–03–06; 网络出版日期: 2017–06–26

国家自然科学基金 (批准号: 61365002, 61462045, 61462042, 61662030) 和江西省教育厅科技项目 (批准号: GJJ150350) 资助

摘要 随着深度学习和多模态融合的深入研究, 问答系统从传统的纯文本方式扩展到结合图片的视觉问答, 成为计算机视觉与自然语言理解的交叉研究热点. Hyeonwoo 等最新提出的动态参数预测模型 (DPPnet) 能够简单、有效地实现问句和图片信息的融合. 但是该模型在进行网络参数动态 Hash 分配时位置随机, 缺乏图像内容空间分布的考虑. 针对此不足, 本文提出一种新的空间离散余弦 Hash 动态参数网络. 算法采用全卷积方式提取保留图像内容空间分布信息的 conv7 特征. 在此基础上, 利用问题信息生成具有一定空间结构性的卷积核, 实现问题与图像信息融合预测视觉答案的过程. 在 COCOqa 和 MSCOCO-VQA 两个公开数据集上与同期主流算法进行了实验对比, 实验结果表明, 本文的网络模型较以往的模型能够更为精确地预测视觉问答答案.

关键词 视觉问答, 动态参数预测, 空间离散余弦 Hash, 深度网络, 多模态融合

1 引言

基于文本的问答系统 (question answering) 在传统自然语言处理领域得到了广泛的研究和应用^[1]. 近年来, 人们对问答系统的适用性有了更高的要求. 问答系统的答案不仅仅局限于从文本内容中获取, 也可以来源于图像视觉内容, 如图 1 所示. 这类形式的问答被称为视觉领域的问答 (即视觉问答, visual question answering, 以下简称 VQA). 视觉问答是计算机视觉与自然语言处理的交叉研究方向, 自 2014 年 Malinowski 等^[2]的工作之后, 迅速成为近年来的研究热点, 受到广泛的关注.

VQA 涉及到图片和文本两种模态数据, 需要有效结合这两种信息, 才能得到正确的答案. 因此, VQA 算法模型既需要利用计算机视觉算法对图片进行信息特征的提取, 又需要应用自然语言理解的方法对问题句子进行分析. 在图像特征提取方面, 主流方法是采用深度卷积神经网络 (convolutional neural network, CNN). 代表性的模型有 AlexNet^[3], VGGNet^[4], GoogLeNet^[5] 和 ResNet^[6] 等. 在问句

引用格式: 孟祥申, 江爱文, 刘长红, 等. 基于 Spatial-DCTHash 动态参数网络的视觉问答算法. 中国科学: 信息科学, 2017, 47: 1008–1022, doi: 10.1360/N112016-00288
Meng X S, Jiang A W, Liu C H, et al. Visual question answering based on spatial DCTHash dynamic parameter network (in Chinese). Sci Sin Inform, 2017, 47: 1008–1022, doi: 10.1360/N112016-00288

图 1 (网络版彩图) 视觉问答样例示意图

Figure 1 (Color online) Illusional samples for visual question answering

分析方面, 深度递归神经网络 (recurrent neural network) 得到广泛运用. 长短期记忆网络 (long-short term memory, LSTM^[7]) 和门限循环单元 (gated recurrent unit, GRU^[8]) 均取得了非常不错的性能. 在 VQA 的初期研究中, 主流思路是借鉴图片描述生成 (image captioning) 方向的多模态信息处理策略. 例如 Malinowski 等^[9] 工作中处理 VQA 的方式与 Donahue 等^[10] 有关生成图片描述的研究基本相似. 他们的方法都是利用 CNN 先提取出图像的特征, 用 LSTM 处理问题序列, 区别在于 VQA 是将问题和图像同时输入到 LSTM 产生答案, 而图片描述生成则只是将图片输入到 LSTM 产生图片的描述语句. 他们的处理方式只利用了图像的全局信息, 缺点是没有充分考虑图像内容的空间分布与问题的关联. 随着研究的深入, 部分学者认为在图像中的某些局部区域能够显式地定位问题答案所在, 于是开始引入选择注意 (attention) 机制^[11~13] 来处理 VQA. 他们共同的做法是通过非线性方法 (如多层感知器 MLP), 利用问题相关特征学习图像区域的局部权重系数, 对图像区域特征进行加权, 最终完成 VQA 的答案预测. 这类算法总体来说在性能上较之前的模型有了较大的提升, 但是模型也越趋复杂, 加大了深度网络的训练难度.

Noh 等^[14] 提出的动态参数预测模型 (dynamic parameters prediction, DPPnet), 能够简单、有效地实现问句和图片特征的融合. DPPnet 通过在 CNN 网络中引入问题相关的动态参数全连接层, 利用 Chen 等^[15] 的 HashedNet 压缩策略, 自适应提取问题相关的图像特征, 预测 VQA 问题的答案. 他们在公开的 MSCOCO-VQA 数据集上取得了非常优秀的结果.

虽然 DPPnet 整个网络清晰简单、易于训练, 但是此模型选择在图像空域网络的全连接层上进行动态参数 Hash 分配, 对应的权重系数位置随机, 缺乏对图像空间信息的利用. 针对 DPPnet 的不足, 本文着重研究了其改进策略. 一方面, 以 VGGnet 为具体 CNN 对象, 对网络的全连接层进行了改造, 获取保留了空间信息的图像特征; 另一方面, 基于离散余弦 Hash, 提出采用 Spatial-DCTHash 卷积层来融合图像与问题信息.

在卷积神经网络训练过程中, 虽然一次卷积操作和一次全连接操作在内积计算上是数学等价的,

图 2 问题句信息处理网络与图像信息处理网络

Figure 2 Deep information extraction networks for question and image, respectively. (a) RNN for question; (b) VGGNet for image

但是 CNN 的全连接层系数是用来联系输入/输出两个一维特征的, 因此没有空间分布的概念. 而图像区域平面内容具有空间分布性, 空间卷积过程是卷积核在图像特征映射平面上进行滑动窗扫的过程. 因此, 对于无空间分布概念的全连接层而言, 权值随机化不会产生多少影响, 但是对于空域卷积核而言, 参数随机化会影响到卷积核的空间结构性. 这在 Chen 等^[16]的 FreshNet 论文中有比较. DCThash 是一种比较巧的、简便的修改权值随机化的方式. 本文对 DPPnet 的改进, 最大的价值在于验证了在空域卷积核上保持结构信息的同时能实现与 DPPnet 同样有效的融合过程. 理论上, DPPnet 是本文网络模型的特殊形式, Spatial-DCTHash 的融合方式更为通用, 可适用范围更大.

在目前公认的两个公开数据集 COCOqa 和 MSCOCO-VQA 上与同期主流算法进行了对比实验, 实验结果表明该算法是有效的, 在性能上相比 DPPnet 有较好的提升, VQA 性能达到了模型提出时的领先水平.

余下部分组织如下: 第 2 节介绍 DPPnet 的网络模型及相关工作, 第 3 节介绍所提出的模型, 第 4 节给出本文的实验细节和实验结果, 最后第 5 节进行总结.

2 DPPnet 模型简介

动态参数预测模型 DPPnet 相比同期的其他 VQA 方法, 具有网络结构清晰、简单和性能优越等优点. 主要体现在两方面: (1) 问题句的处理; (2) 问题特征和图片特征的融合处理.

2.1 问题句和图像的处理

One-hot 向量常被用来表示单词文本, 其维数为数据集问句单词的词汇表长度, 因此, 维度通常比较大. DPPnet 首先对这些单词进行 word2vec^[17] 词嵌入降维, 得到每个单词的词嵌入特征, 然后将问句单词依次输入循环神经网络 GRU. GRU 的最后一个时刻状态输出作为问句的特征表示, 如图 2(a) 所示. 为了更有效地抽取文本特征, DPPnet 的 GRU 采用基于预训练好的 Skip-thought^[18] 向量模型参数作为初始参数.

DPPnet 采用主流预训练好的 VGGNet-16 作为图像特征抽取网络, 取倒数第 2 个全连接层 (图 2(b) 中的 “fc7” 层) 输出的 4096 维特征作为图像的信息表示. “fc7” 特征是图像的全局特征, 具有较强的判别力.

图 3 (网络版彩图) Hash 权值矩阵生成示意图

Figure 3 (Color online) Illustration on hashing weight matrix's generation

2.2 问题与图像的信息融合

DPPnet 采用 HashedNet^[15] 作为问题信息与图像信息的融合策略.

由于神经网络中参数存在着大量冗余性, HashedNet 模型最初提出的动机便是为了压缩卷积神经网络需要学习的参数规模. HashedNet 提出采用权重 Hash 的方式减少全连接的网络参数, 具体实现如下: (1) 首先生成一维的初始化权值 (称为真实权重). 真实权重的元素个数远远小于全连接层的二维权重矩阵的元素个数; (2) 利用 Hash 函数将真实权重的每个元素随机分配到网络全连接层的二维权重矩阵, 如图 3 所示. 虽然在二维矩阵权重分配过程中, 即使真实权重中的同一元素可能存在多次出现或者不同元素在 Hash 时发生碰撞等情况, HashedNet 证明这些情况并不会严重影响全连接网络的性能. HashedNet 实现了用少量的权值参数替代原先庞大的权值矩阵参数, 在网络训练时只需要对应更新原先的真实权重即可, 从而精简了网络.

基于 HashedNet, DPPnet 巧妙地提出利用动态参数全连接层实现图像、问题的特征融合, 并进行答案的预测. 问题特征经过非线性映射生成 HashedNet 的真实权值, 然后通过 HashedNet 的 Hash 处理方式将该真实权值参数用于构建全连接层的网络权值矩阵, 从而问题信息被动态嵌入到全连接层权值系数中. DPPnet 将经过 VGGnet 抽取得到的图片特征输入到该全连接层, 最终输出问题答案的预测.

具体地, 问题句经过 GRU 后的特征输出经过线性映射 $q = \text{GRU}([w_{e1}, w_{e2}, \dots, w_{eT}])W + b$ 得到的 $q \in \mathbb{R}^N$ 用来表示问题特征相关的权值系数, 其中 w_{et} 表示第 t 个单词的词嵌入特征, T 表示文本句单词总个数, W 和 b 分别为权重和偏置. 图像的 “fc7” 层 4096 维特征经过一个全连接层进行降维得到一个 d 维的图像特征 X , 随后输入到 DPPnet 设计的动态参数全连接层 $S = W^q X + b$, 其中 X 为降维后的图像特征, S 为动态参数层的输出, b 为偏置, W^q 代表动态参数层的权重参数, 由问题特征 q 经 HashedNet 动态生成. 输出 S 作为结合了图像和文本信息的综合特征, 用于预测问题的答案.

动态参数 W^q 的生成过程中使用了位置分配函数 $W_{u,v}^q = \xi(u, v) \cdot q_{g(u,v)}$ 将问题相关的权值 q 分配到相应的 Hash 位置, 其中 $W_{u,v}^q$ 为动态参数层的权重矩阵 W^q 在 (u, v) 位置上的权值. $g(u, v) \in \{1, 2, \dots, N\}$, $\xi(u, v) \in \{\pm 1\}$ 是两个独立的 Hash 函数, 可通过调用 xxhash¹⁾ 接口函数生成, N 是问题相关权值 q 的维数.

1) <https://code.google.com/p/xxhash/>.

图 4 (网络版彩图) 空间离散余弦 Hash 动态参数网络模型
Figure 4 (Color online) Spatial-DCTHash based dynamic parameters network

3 基于空间离散余弦 Hash 的动态参数网络

DPPnet 是在网络全连接层进行多模态信息 Hash 融合, 所采用的图像特征是全局特征. DPPnet 没有考虑图像内容空间分布的重要性, Hash 权值的分布也丢失了图像的空间结构信息. 针对 DPPnet 存在的不足, 提出基于空间离散余弦 Hash (Spatial-DCTHash) 动态参数网络模型, 整体框架如图 4 所示.

图像经过改造后的 VGGNet 网络后, 选取网络的 “conv7” 层输出作为保留了空间分布信息的图像特征. 问题句信息 q 经过 Spatial-DCTHash 卷积层与图像信息进行融合. 融合后的输出进行池化, 最终得到的多模态特征用于答案的预测.

3.1 图像卷积网络构建

同期基于注意 (attention) 机制的 VQA 的工作表明, 大部分问题的答案与图像的局部区域相关. 因此, 保留一定的图像空间局部信息对 VQA 的答案生成是有益处的.

DPPnet 采用的 “fc7” 全局特征, 具有非常强的鉴别力. 因此, 采用类似全卷积网络 (fully convolutional network) [19] 的方式, 对传统的 CNN 网络结构进行适当的改造, 提取鉴别力与 “fc7” 相当, 但保留了一定的空间分布信息的 “conv7” 特征. 改造后的 CNN 网络如图 5 所示.

以 VGGNet-16 网络为例, 去掉整个网络的后 3 层全连接层, 在最后一个卷积层 “conv5” 后面加上第 6 层卷积层 “conv6” 和第 7 层卷积层 “conv7”.

卷积层 “conv6” 的卷积核采用 4096 个 $512 \times 7 \times 7$ 滤波器. 本文继续沿用原 “fc6” 的权值作为这些卷积核的权值参数.

卷积层 “conv7” 的卷积核采用的是 4096 个 $4096 \times 1 \times 1$ 滤波器. 同样该卷积核参数也继续沿用 “fc7” 的权值参数. 经过对网络进行简单的改造后, 大小为 $512 \times 14 \times 14$ 的 “conv5” 特征经过 “conv6” 和 “conv7” 两层的卷积操作后得到的 “conv7” 层输出大小为 $4096 \times 8 \times 8$. “conv7” 特征保留了图像的空间分布信息. 每个特征点近似对应到原图像的 28×28 局部区域中心范围.

conv7 中每个特征点的形成过程与 fc7 形成机制总体是一致的, 因此 conv7 的 4096 维通道特征的鉴别力与 “fc7” 特征相当. 在实验过程中有过经验性验证, 即对 conv7 进行 mean pooling 和 max

图 5 (网络版彩图) 图像 conv7 特征提取的卷积网络

Figure 5 (Color online) Convolutional network for feature "conv7"

pooling 操作后, 分类性能均和 fc7 的效果相差不大.

3.2 空间离散余弦 Hash (Spatial-DCTHash) 卷积层构建

DPPnet 生成的是动态全连接层, 采用的是传统的、在空域权重矩阵 (即全连接层的二维权重矩阵) 上进行位置 Hash. CNN 的卷积核在空间上具有一定的结构性, 并不完全随机. 因此, 如果采用类似 HashedNet 的方式直接在空域的权重矩阵上进行 Hash, 网络层系数的分布势必会呈现随机化, 丢失了必要的图像结构性.

Chen 等^[16]提出了 FreshHash. 该算法借鉴图像压缩中的余弦变换方法, 在经过 Fourier、离散余弦等变换得到的频谱图频域内先进行真实权值的 Hash 分配, 然后将频域矩阵通过离散余弦逆变换得到图像的空域卷积核. 虽然在频域上的 Hash 分配依然比较随机, 但是再通过逆变换到空域上的结果却保留一定的结构性, 有效避免了直接 Hash 的完全随机形式. 这种方法得到的 CNN 空域卷积核不仅保留了原始 CNN 卷积核的结构性, 同时又保留 HashedNet 的空间压缩的优点. 相比传统 HashedNet, FreshNet 可以取得更好的性能. FreshNet 的卷积核空间结构保留的优越性在 Chen 等^[16]的工作中也得到有力的验证, 请参考 Chen 等^[16]论文中关于 HashedNet 与 FreshNet 的直观对比效果.

对于一个空域的权值矩阵, 经过离散余弦变换 (DCT) 得到对应在频域上的频谱矩阵. 相应的 DCT 运算定义为

$$\mathcal{W}(u, v) = r_v r_u \sum_{x=0}^{d-1} \sum_{y=0}^{d-1} W(x, y) \cos \frac{(2x+1)u\pi}{2d} \cos \frac{(2y+1)v\pi}{2d}, \quad (1)$$

其中 $u, v, x, y = 0, 1, 2, \dots, d-1$; $r_v, r_u = \begin{cases} \sqrt{1/d}, & u, v = 0 \\ \sqrt{2/d}, & u, v = 1, 2, 3, \dots, d-1 \end{cases}$.

为了方便记忆与后面的书写, 这里用符号 F_{dct} 来表示离散余弦变换. 因此式 (1) 可以简写为 $\mathcal{W} = F_{\text{dct}}(W)$. 频谱矩阵 $\mathcal{W}$ 转换为空域上的矩阵 $W \in \mathbb{R}^{d \times d}$, 需要经过离散余弦逆变换, 如式 (2) 所示:

$$W(x, y) = \sum_{u=0}^{d-1} \sum_{v=0}^{d-1} r_v r_u \mathcal{W}(u, v) \cos \frac{(2x+1)u\pi}{2d} \cos \frac{(2y+1)v\pi}{2d}. \quad (2)$$

同样地, 用符号 F_{dct}^{-1} 表示离散余弦逆变换, 故式 (2) 可以记作 $W = F_{\text{dct}}^{-1}(\mathcal{W})$.

图 6 (网络版彩图) 空间余弦变换 Hash 动态参数卷积核生成过程

Figure 6 (Color online) The generation flowchart of spatial-DCTHash based convolutional kernel

如果将 CNN 空域卷积核看成一幅“图像”的话,那么整个深度网络是要学习这种图像的“最优”表示. DPPnet 所采用的 HashedNet 方式,是一种直接在滤波器空域上以 Hash 采样的手段近似逼近原始滤波器权值.由于 Hash 位置随机,造成 Hash 出来的空域滤波器呈现随机形式.

离散余弦变换压缩是数字图像压缩中常用的方法.在图像的 DCT 频谱分布中,一般频谱图的左上角是低频,右下角是高频.在 DCT 压缩的时候,即使去掉位于高频部分的能量,只要保留了大部分低频能量,图像的信息不会有太多损失,仍可以实现对图像的有效逼近.

本文类似于 DPPnet 对 HashedNet 的修改运用,也相应地对 FreshNet 进行了修改,从而实现了基于空间离散余弦 Hash 卷积的问题与图像信息融合方式,使之更加满足 VQA 具体任务.

以具体某一层卷积层为例,假设输入通道数为 I ,输出通道数为 O ,感受野区域大小为 $k \times k$,那么总共有 $I \times O$ 个 $k \times k$ 大小的卷积核.在传统的卷积网络,卷积核权值可以表示为一个四维的矩阵 $W \in \mathbb{R}^{I \times O \times k \times k}$,需要学习的卷积参数数目为 $I \times O \times k \times k$ 个.记第 i 输入通道和第 o 输出通道对应的卷积核为 $W^{io} \in \mathbb{R}^{k \times k}$, W_{xy}^{io} 表示第 i 输入通道和第 o 输出通道对应的卷积核上 (x, y) 位置的权值,而 W_{uv}^{io} 则是该卷积核对应的频域上 (u, v) 位置的权值.

采用频谱分布相关的分配原则,将根据问题特征动态生成的实际权值参数经过 Hash 位置映射到 CNN 卷积核的频域矩阵.具体地,在进行空间余弦 Hash 映射时,让频域左上角低频部分的分配的 Hash 值更多,低频区域的权值不会有太多冲突,从而参数的表示更加多样化.让右下角的高频部分 Hash 值分配较少些.因为高频信息对图像的整体内容结构分布的重要性没有低频那么大,因此在 Hash 权值分配时,即使发生冲突也不会产生太大影响.相比一般的均匀 Hash,这样处理的好处是,可以尽量保证频域的低频有效部分信息的丰富性,有利于信息的保留.将生成的频域权值矩阵经过逆离散余弦变换,便可以得到对图像做卷积操作所需的空域卷积核.整个过程如图 6 所示.

当执行模型训练时, Spatial-DCTHash 卷积层的具体前向 - 后向过程如 3.2.1 和 3.2.2 小节.

3.2.1 前向传播过程

首先,通过外部传入所需要嵌入的权值 $q \in \mathbb{R}^N$. N 是问题相关的权值向量 q 的长度,一般情况下 $N \ll I \times O \times k \times k$. 频域卷积核矩阵 W_{uv}^{io} 中的元素和权值 q 的关系可以用式 (3) 表示:

$$W_{uv}^{io} = \xi(i, o, u, v) \cdot q_{g(i, o, u, v)}, \quad (3)$$

其中, $g(i, o, u, v) \in \{1, 2, \dots, N\}$, $\xi(i, o, u, v) \in \{\pm 1\}$ 是两个相互独立的 Hash 函数, $\cdot$ 表示内积.

其次,频域卷积核矩阵 W^{io} 通过 DCT 逆变换到空域上的卷积核矩阵 W^{io} . 用这个嵌入了问题信

息的卷积核对有空间分布信息的图像特征进行卷积, 从而实现问题和图片特征的有效融合.

$$W_{uv}^{io} = F_{\text{dct}}^{-1}(\mathcal{W}_{uv}^{io}). \quad (4)$$

3.2.2 后向传播梯度计算过程

根据 Spatial-DCTHash 空域卷积核的生成过程, 很容易计算在网络进行反向传播时, W 对 $\mathcal{W}$ 的偏导:

$$\frac{\partial W_{xy}^{io}}{\partial \mathcal{W}_{uv}^{io}} = r_v r_u \cos \frac{(2x+1)u\pi}{2d} \cos \frac{(2y+1)v\pi}{2d}. \quad (5)$$

假设整个网络模型的损失函数为 Γ , 根据式 (2) 和 (5), 很容易计算得到损失函数对频域权值 $\mathcal{W}_{uv}^{io}$ 的偏导为

$$\begin{aligned} \frac{\partial \Gamma}{\partial \mathcal{W}_{uv}^{io}} &= \frac{\partial \Gamma}{\partial W_{xy}^{io}} \frac{\partial W_{xy}^{io}}{\partial \mathcal{W}_{uv}^{io}} \\ &= \sum_{x=1}^{d-1} \sum_{y=1}^{d-1} r_v r_u \cos \frac{(2x+1)u\pi}{2d} \cos \frac{(2y+1)v\pi}{2d} \frac{\partial \Gamma}{\partial W_{xy}^{io}} \\ &= r_v r_u \sum_{x=1}^{d-1} \sum_{y=1}^{d-1} \cos \frac{(2x+1)u\pi}{2d} \cos \frac{(2y+1)v\pi}{2d} \frac{\partial \Gamma}{\partial W_{xy}^{io}}. \end{aligned} \quad (6)$$

对比式 (1), 式 (6) 可以简写成式 (7) 的形式:

$$\frac{\partial \Gamma}{\partial \mathcal{W}^{io}} = F_{\text{dct}} \left(\frac{\partial \Gamma}{\partial W^{io}} \right). \quad (7)$$

式 (7) 明确地表达了卷积核频域权值梯度和空域卷积核梯度之间的关系.

由于频域权值矩阵参数由问题信息 Hash 得到, 因此, 在反向传播时, 问题特征的权值梯度可以相应地根据式 (8) 计算得到:

$$\begin{aligned} \frac{\partial \Gamma}{\partial q_g} &= \sum_{i=0}^{I-1} \sum_{o=0}^{O-1} \sum_{x=0}^{d-1} \sum_{y=0}^{d-1} \frac{\partial \Gamma}{\partial \mathcal{W}_{uv}^{io}} \frac{\partial \mathcal{W}_{uv}^{io}}{\partial q_g} \\ &= \sum_{i=0}^{I-1} \sum_{o=0}^{O-1} \sum_{x=0}^{d-1} \sum_{y=0}^{d-1} \left[F_{\text{dct}} \left(\frac{\partial \Gamma}{\partial W^{io}} \right) \right]_{(u,v)} \frac{\partial \mathcal{W}_{uv}^{io}}{\partial q_g}. \end{aligned} \quad (8)$$

在计算频域权值矩阵对问题特征的梯度时, 需要知道频域上的权值矩阵和问题特征生成的权值参数的对应关系, 即损失函数对问题特征 q 的梯度以及 q 在频域权值矩阵的 Hash 位置. 因此, 结合式 (3), 式 (8) 可以进一步改写为

$$\frac{\partial \Gamma}{\partial q_g} = \sum_{i,o,u,v} \sum_{g=g(i,o,u,v)} \xi(i,o,u,v) \left[F_{\text{dct}} \left(\frac{\partial \Gamma}{\partial W^{io}} \right) \right]_{(u,v)}. \quad (9)$$

至此, 完整地解决了损失函数对问题特征的梯度求解过程.

Spatial-DCTHash 方式和 Attention 方式有一定的共通之处, 即均考虑到了图像的空间分布重要性. 区别在于, 在 Attention 的方式中, 区域权重系数由问题信息通过映射生成, 所形成的图像“期望特征”根据图像特征的概率加权计算决定, 融合过程相对间接. 基于 Spatial-DCTHash 的空间信息利用机制采取了 CNN 的卷积思想, 由问题信息生成空间卷积核, 卷积的过程便是问题信息与图像局部内容信息的融合过程, 因此图像和问题信息的融合更为直接, 嵌入 CNN 网络更简便.

4 实验设置与分析

4.1 实验数据集与实验设置

为了验证所提出的网络模型的有效性, 本文在两个已公开的主流视觉问答数据集上进行了对比实验.

COCOqa 数据集由多伦多大学 Ren 等^[20] 发布, 包含了 123287 张图片、78736 条训练问题和 38948 条测试问题. 该数据集的图片源自微软 MSCOCO 彩色图片数据集. 问题句最长包含 55 个单词, 平均包含 9.65 个单词.

MSCOCO-VQA 是 Antol 等^[21] 2015 年发布的数据集. 该数据集将 MSCOCO 数据集中所有的彩色图片作为问答图片. 训练集问题约 248349 条, 验证集问题约 121512 条, 测试问题集约 244302 条. 该数据集是目前规模最大的, 也是使用最广的视觉问答数据集.

为了实验的相对公平, 在所有的实验中, Spatial-DCTHash 的网络参数设置与 DPPnet 基本保持一致, Hash 的随机种子和参数也均采用与 DPPnet 同样的设定. 在所提出的网络模型中, 需要训练的网络参数规模大致包括: 在问题特征抽取网络中的 GRU 单元部分, 输入的 skip-thought 词嵌入特征维度为 620 维, 隐含层节点数为 2400. DCThash 卷积核通道数设置为 1000, 卷积核大小在本实验中取的是 3×3 . Spatial DCTHash 的压缩比为 1 : 100. 答案预测网络是 1000×1000 的全连接层. 图像特征和问题特征维度均降维 1000 之后再进行 spatial-DCThash 方式融合. 网络参数初始化为 $[-0.005, 0.005]$ 的均匀分布, 学习率初始值设为 0.001, 批处理的 mini-batch 大小设为 48. 训练过程采用 Adam 算法进行优化.

为了能够清晰地表现利用空间离散余弦变换 Hash 卷积在多模态信息融合方面的作用, 在整个实验的训练过程中, 保持 CNN 网络参数固定不变, 不进行微调. 因为我们认为微调是一个极具技巧性的工作, 虽然能够帮助模型在数据集上取得更好的结果, 但是与本文的主要贡献无关. “DPPnet-[CNN-FIXED]” 代表 CNN 参数固定的情况下 DPPnet 模型. “Spatial-DCTHash-[CNN-FIXED]” 代表本文提出的基于空间离散余弦 Hash 的动态参数网络模型.

实验程序基于 Torch 实现, 运行环境为 Ubuntu 14.04, GPU 为 GTX970.

4.2 实验结果与分析

在 COCOqa 数据集上, 采用答案准确率 (Acc) 和刻画 Wu-Palmer 相似性^[22] 的 WUPS^[2] 得分作为算法性能的评价标准.

WUPS 的计算标准为

$$\text{WUPS} = \frac{1}{M} \sum_{i=1}^M \min \left\{ \prod_{a \in A^i} \max_{t \in T^i} \mu(a, t), \prod_{t \in T^i} \max_{a \in A^i} \mu(a, t) \right\}, \quad (10)$$

其中 M 表示测试样例总数, A^i 和 T^i 分别表示第 i 个测试样例的预测答案和真实答案, $\mu(\cdot, \cdot)$ 表示预测答案和真实答案的 Wu-Palmer 相似性. 这里根据通用做法, 设定两个阈值 0.9 和 0.0 来进行性能比较, 分别用 “WUPS@0.9” 和 “WUPS@0.0” 表示. “Acc” 表示准确率.

实验所采用的性能评价程序源自 Ren 等^[20] 发布的代码. 具体的性能如表 1 所示. “IMG+BOW”, “2VIS+BLSTM” 与 “Ensemble” 是 Ren 等^[20] 在发布 COCOqa 数据集时提出的基准方法.

由表 1 的实验结果, 可以看出所提出的 Spatial-DCTHash 模型无论是在准确率还是 WUPS 得分上均比其他模型要好. 该模型在 COCOqa 数据集上能取得更高的性能.

表 1 COCOqa 数据集的性能比较

Table 1 Performance comparisons on COCOqa dataset

	WUPS@0.9	WUPS@0.0	Acc
IMG+BOW ^[20]	66.78	88.99	55.92
2VIS+BLSTM ^[20]	65.34	88.64	55.09
Ensemble ^[20]	67.90	89.52	57.84
ConvQA ^[23]	65.36	88.58	54.95
DPPnet-[CNN-FIXED] ^[14]	69.61	90.38	59.52
Spatial-DCTHash-[CNN-FIXED]	69.95	90.47	60.01

表 2 MSCOCO-VQA 数据集 open-ended 任务 val 集的性能

Table 2 Performance comparisons on open-ended task on MSCOCO-VQA val dataset

Val set	Yes/no	Number	Other	All
DPPnet-[CNN-FIXED] ^[14]	81.05	33.49	41.04	55.07
Spatial-DCTHash-[CNN-FIXED]	80.76	33.85	41.46	55.21

表 3 MSCOCO-VQA 数据集 open-ended 任务 test-dev 集的性能

Table 3 Performance comparisons on open-ended task on MSCOCO-VQA test-dev dataset

Test-dev set	Yes/no	Number	Other	All
Question ^[21]	75.66	36.7	27.14	48.09
Image ^[21]	64.01	0.42	3.77	28.13
Q+I ^[21]	75.55	33.67	37.37	52.64
LSTM Q ^[21]	78.2	35.68	26.59	48.76
LSTM Q+I ^[21]	78.94	35.24	36.42	53.74
DPPnet-[CNN-FIXED] ^[14]	80.48	37.20	40.90	56.74
DPPnet ^[14]	80.71	37.24	41.69	57.22
Spatial-DCTHash-[CNN-FIXED]	80.54	36.81	42.52	57.51

表 4 MSCOCO-VQA 数据集 open-ended 任务 test-standard 集的性能

Table 4 Performance comparisons on open-ended task on MSCOCO-VQA test-standard dataset

Test-standard set	Yes/no	Number	Other	All
Human ^[21]	95.77	83.39	72.67	83.3
LSTM Q+I ^[21]	—	—	—	54.06
DPPnet ^[14]	80.28	36.92	42.24	57.36
Spatial-DCTHash-[CNN-FIXED]	80.20	35.29	42.94	57.50

MSCOCO-VQA 的准确率并非简单计算答对测试样例的比例, 其计算公式为

$$\text{Acc}_{\text{VQA}} = \frac{1}{M} \sum_{i=1}^M \min \left\{ \frac{\# \text{ humans that provided that answer}}{3}, 1 \right\}, \quad (11)$$

其中, M 表示测试样例总数, “# humans that provided that answer” 表示预测答案与 MSCOCO-VQA 人工回答的答案相同的个数.

表 5 MSCOCO-VQA val 集上每个问题类型准确率 (Acc) 对比
Table 5 Comparisons of accuracy on different question types on MSCOCO-VQA val dataset

Question type	Spatial-DCTHash	DPPnet	Question type	Spatial-DCTHash	DPPnet
are there	82.29	83.15	what is the man	50.37	49.5
what brand	36.2	34.73	which	41.65	38.3
what room is	82.22	84.11	are these	76.29	77.07
what color is	51.58	48.73	what are	47.03	47.59
is	79.96	80.41	what is the	36.77	36.67
are they	78.15	78.05	where are the	29.82	29.57
what number is	3.33	2.03	is this a	79.34	79.67
what sport is	83.91	84.78	can you	76.8	76.37
are	75.52	75.68	what time	19.32	20.43
is the	76.21	76.41	what are the	37.37	38.31
what is the person	51.04	49.14	are there any	74.9	73.73
how many	39.74	39.59	what color are the	50.63	50.14
does this	78.73	79.83	why	16.1	15.48
is there a	89.27	89.34	what is this	51.35	51.02
is he	80.08	80.59	how many people are in	35.61	34.54
what	36.7	36.07	do you	81.09	82.72
does the	78.32	79.6	is this	77.82	78.63
is the person	76.59	75.89	why is the	16.92	19.19
where is the	25.69	26.83	what is the color of the	62.51	61.41
what animal is	62.25	60.95	what is	29.1	29.25
how	23.07	22.75	could	90.21	90.53
what is the woman	42.56	41.45	is that a	74.54	73.03
none of the above	54	54.02	what is in the	35.46	33.78
who is	24.38	25.61	what does the	22.2	20.73
is the woman	78.86	77.64	what kind of	45.78	45.87
are the	75.69	76.03	is it	81.47	83.27
how many people are	39.92	38.55	is the man	79.61	79.15
what is on the	34.16	33.51	what is the name	7.07	7.21
has	78.39	79.45	is there	83.16	84.41
was	82.82	82.67	what color is the	53.55	51.44
what type of	44.71	45.56	what color	39.23	36.47
is this an	80.18	80	is this person	75.25	75.21
do	75.7	74.62			

根据 CVPR2016 VQA Challenge 发布的评价工具进行计算准确率. 当计算在 test-dev 和 test-standard 数据集上的性能时, 将测试结果上传到 VQA Challenge 服务器上, 由服务器返回计算的绩效. 在 test-dev 和 test-standard 上, 还引用了 “DPPnet” 的测试结果作为参照.

我们注意到 MSCOCO-VQA 数据集在线测试结果反馈会有延迟, 并且网站结果提交次数也有限制.

图 7 (网络版彩图) 部分预测结果样例展示

Figure 7 (Color online) Typical samples on VQA predicted results

同时还注意到 MSCOCO-VQA 的验证集 val 数据规模相对较小, 可以根据 CVPR2016 VQA Challenge 官方发布的性能计算程序进行线下测试. 因此, 在 MSCOCO-VQA 数据集上, 利用训练集 train 数据来训练所提出的模型, 验证集 val 数据进行验证, 将结果模型在 test-dev 和 test-standard 数据上进行测试. 具体的性能如表 2 ~ 4 所示. 其中 “yes/no”, “number” 和 “other” 分别表示该数据集包含的 3 种主要的子问题类型. “All” 表示整个数据集.

表 3 中 “Question” (只使用问题来预测答案)、“Image” (只使用图像来预测答案)、“Q+I” (只采用问题和图像的简单结合来预测答案)、“LSTM Q” (用问题通过 LSTM 得到的特征来预测答案) 和 “LSTM Q+I” (将问题和图像通过 LSTM 预测答案) 是 Antol 等^[21] 发布 MSCOCO-VQA 数据集时提出的基准方法. “DPPnet” 代表对 DPPnet 网络 (包括 CNN) 整体微调后的网络模型. 表 4 中 “Human” 是在 MSCOCO-VQA 数据集上人类测试的结果.

从表 3 中可以看出, 所提出的网络模型在同样处于 CNN 不做微调的情况下较 DPPnet 高出近 1 个百分点, 甚至比 DPPnet 进行微调的结果还高.

test-standard 数据集较 test-dev 数据集的测试样例数目要多得多, 数据量更大. 即便如此, 从表 4 显示的性能比较结果来看, 本文不做 CNN 微调的效果最终依然比 DPPnet 进行微调的结果还要高. 这显示了所提出的模型具备有效的处理 VQA 问题的能力.

还注意到, 在 MSCOCO-VQA 数据集的 3 类子问题上, 所提出的网络模型在 “other” 类型上的提升相对更加明显. 整理发现, 在 “other” 类型问句中, 有较多的针对图像具体局部细节内容的提问, 而非像 “yes/no” 和 “number” 类型问句那么笼统. 这在一定程度上显示了保留空间内容分布信息对于 VQA 的有效性. 具体更细的划分可以从表 5 中窥探一二.

在 VQA 数据集的 ground truth annotation 文件中包含了每条问题的 question type 属性. 为了

更加深入地了解所提出的网络模型在不同类型的问题上的准确率, 在表 5 中列出了在验证集 val 上的各个子问题类型的实验结果, 供参考. Spatial-DCTHash 与 DPPnet 均为 CNN-FIXED 情况下的模型. 问题子类型的划分采用 VQA 发布的标准工具包计算并根据该 question type 属性统计而来. 不难发现, 对于大部分 “what” 问句之类的、针对局部区域内容的问题, 本文的算法均表现比较好, 从而显示了空间结构信息保留的重要性. 部分实验样例结果比较如图 7 所示.

综上, 本文提出的 Spatial-DCTHash 模型在性能上能较同期主流算法有较好的提升, 实现更精确的答案预测, 表现出良好的优越性.

5 总结

本文提出了一种有效的、基于空间离散余弦变换 Hash 动态参数预测网络模型的视觉问答算法. 算法从问题与图像特征融合的 Hash 过程、图像内容空间分布表示的网络结构两方面对最新的 DPPnet 视觉问答算法进行改进, 弥补了 DPPnet 在空间结构信息表示方面的缺失. 在保留 DPPnet 网络简洁、易训练的优点的同时, 实现了更为有效的、问题与图像多模态信息融合方式. 在 VQA 公开数据集上与同期主流算法进行了实验比较, 实验结果表明, Spatial-DCTHash 具有良好的算法优越性, 能实现较好的性能提升. 目前选择在 CNN-FIXED 情况下对网络模型的有效性进行验证. 在进一步工作中, 将研究采用参数微调等迁移学习方式进行网络优化, 以实现更好的视觉答案预测.

参考文献

- 1 Yates A, Cafarella M, Banko M, et al. Textrunner: open information extraction on the web. In: Proceedings of Human Language Technologies: the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Rochester, 2007. 25–26
- 2 Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input. In: Proceedings of the Annual Conference on Neural Information Processing Systems, Quebec, 2014. 1682–1690
- 3 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of the Annual Conference on Neural Information Processing Systems, California, 2012. 1097–1105
- 4 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations, San Diego, 2015. 1–14
- 5 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 1–9
- 6 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
- 7 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
- 8 Chung J Y, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of the Annual Conference on Neural Information Processing Systems Deep Learning Workshop, Quebec, 2014. 1–9
- 9 Malinowski M, Rohrbach M, Fritz M. Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1–9
- 10 Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 2625–2634
- 11 Shih K J, Singh S, Hoiem D. Where to look: focus regions for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4613–4621
- 12 Jiang A, Wang F, Porikli F, et al. Compositional memory for visual question answering. arXiv: 1511.05676, 2015

- 13 Xu K, Ba J L, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning, Nashville, 2015. 2048–2057
- 14 Noh H, Seo P H, Han B. Image question answering using convolutional neural network with dynamic parameter prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 30–38
- 15 Chen W L, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick. In: Proceedings of the International Conference on Machine Learning, Lille, 2015. 2285–2294
- 16 Chen W L, Wilson J T, Tyree S. Compressing convolutional neural networks. arXiv: 1506.04449, 2015
- 17 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations, Arizona, 2013.
- 18 Kiros R, Zhu Y K, Salakhutdinov R, et al. Skip-thought vectors. In: Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, 2015. 3294–3302
- 19 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3431–3440
- 20 Ren M, Kiros R, Zemel R. Exploring models and data for image question answering. In: Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, 2015. 2953–2961
- 21 Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, New York, 2015. 2425–2433
- 22 Wu Z B, Palmer M. Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Las Cruces, 1994. 133–138
- 23 Ma L, Lu Z D, Li H. Learning to answer questions from image using convolutional neural network. In: Proceedings of the Association for the Advance of Artificial Intelligence, Phoenix, 2016. 3567–3573

Visual question answering based on spatial DCTHash dynamic parameter network

Xiangshen MENG, Aiwen JIANG*, Changhong LIU, Jihua YE & Mingwen WANG

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China

* Corresponding author. E-mail: jiangaiwen@jxnu.edu.cn

Abstract As research on deep learning and multimodal fusion continues to develop, question-answering systems have evolved from purely textual style to incorporating visual information. Visual question answering is becoming one of the inter-discipline topics between natural language processing and computer vision, and is receiving much attention. The dynamic parameter prediction network proposed by Hyeonwoo et al. is capable of effectively combining questions and visual information. However, when the network did weights hashing, the locations of the hashing codes were random, ignoring the image contents' spatial distributions. To overcome this shortcoming, in this paper, a new spatial DCTHash-based dynamic parameter prediction network on multimodal fusion is proposed for predicting visual question's answers. A conv7 feature mapping is extracted to retain the spatial visual information in a fully convolutional style. Then, question-related and visual structure-preserved convolution kernels are generated to perform the visual answer prediction. The proposed model has been compared with current commonly used algorithms on two public datasets: COCOqa and MSCOCO-VQA. The experimental results demonstrate that the proposed model has competitive advantages and can achieve relatively better performance.

Keywords visual question answering, dynamic parameter prediction, spatial DCTHash, deep networks, multimodal fusion

Xiangshen MENG was born in 1992. He received a bachelor's degree in computer science from Jiangxi Normal University, Nanchang, in 2012. Currently, he is a master's degree candidate at Jiangxi Normal University.

Aiwen JIANG was born in 1984. He received his Ph.D. in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2010. Currently, he is an associate professor at Jiangxi Normal University. His research interests include visual question answering and cross-modal retrieval.

Mingwen WANG was born in 1965. He received his Ph.D. in computer science from Shanghai Jiao Tong University, Shanghai, in 2000. Currently, he is a full-time professor at Jiangxi Normal University. His research interests include natural language processing, information retrieval, and machine learning.