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信息工程学院，江西省南昌市 邮编 330022）

摘要：近年来，随着深度学习的应用和多模态的深入研究，问答系统从传统的文本问答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之一。Hyeonwoo Noh 等人在 CVPR2016 中提出一种简单、有效的动态参数预测模型（Dynamic Parameter Prediction Network，DPPnet），但是此模型仅在空域滤波器上进行 Hash，得到权重位置是随机的，没有考虑利用图像的空间信息。对于如何利用图像的空间信息以提高模型性能，本文采用类似 Fully Convolutional Network 的方式改造传统的 VGGnet 卷积神经网络，提取具有空间信息的图像特征，在此基础上，提出一种新的空间离散余弦哈希动态参数网络来结合问题特征和图像特征预测视觉答案。本文在 COCOqa 和 MSCOCO-VQA 数据集上与已有的方法进行了对比实验，实验结果表明本文的算法在性能上有较大提高。

关键字：视觉问答；离散余弦变换；Hash；卷积神经网络

中图分类号：TP391　　**文献标识码：**A

Visual Question Answering based on Spatial-DCTHash Dynamic Parameter Network

Meng Xiangshen, Jiang Aiwen, Liu Changhong, Ye Jihua, Wang Mingwen

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

Abstract: In recent years, with the use of deep learning and the intensive research of the multi-modality, the Question Answering (QA) system has been extended from the traditional text QA into visual QA combining pictures and it has become a hot research topic in computer vision and natural language understanding. In CVPR 2016, Hyeonwoo Noh's team put forward a simple and effective model of Dynamic Parameter Prediction Network (DPPnet). However, this model only hashes in spatial filter and the position weights are obtained randomly without considering the pictures' spatial information. Therefore, centering around the improvement of the DPPnet model by the pictures' spatial information, first, we employ a method similar to Fully Convolutional Network to transform the traditional VGGnet, and then extract image features with spatial information. On this basis, we finally put forward a new model using Spatial Discrete Cosine Transform Hash Network which combines the question and image features to predict the answer. The experimental results on COCOqa and MSCOCO-VQA datasets show that the algorithm used is greatly improved.

Key words: Visual Question Answering; Discrete Cosine Transform (DCT); Hash; Convolutional Neural Network（CNN）

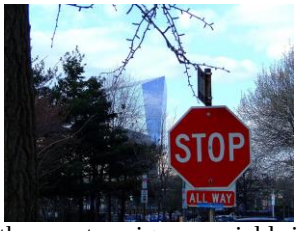
1.引言

基于文本的问答系统[1]（Question Answering）已有很好的应用。近年来，人们对问答系统的答案不满足于仅从文本中提取，也可从图片提取，如图 1 示例所示，这种问答形式称为视觉领域的问答（也叫视觉问答，Visual Question Answering,以下简称 VQA）。视觉问答问题的解决涉及计算机视觉与自然语言理解等领域，在 2014 年 M. Malinowski[2]等人开始研究后，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信息模态，需要有效结合这两种信息，才能得到正确的答案。因此，VQA 算法模型既需要计算机视觉的方法对图片进行信息特征的提取，同时还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，主要采用卷积神经网络（Convolutional Neural Network, CNN），目前效果较好的 CNN 网络模型有 AlexNet[3]，VGGNet[4]、GoogLeNet[5]等。而在问句分析方面，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[6]）和门限循环单元（Gated Recurrent Unit , GRU[7]）均取得了很好的性能。在 VQA 的初期研究中，借鉴多模态信息处理研究中的图片描述生成（Image Captioning）模型，例如 Mateusz Malinowski 等[8]工作中处理 VQA 的方式与 J. Donahue 等[9]有关生成图片描述的研究。他们的方法都是利用 CNN 先提取出图像的特征，用 LSTM 处理序列问题，区别在于 VQA 是将问题输入到 LSTM 产生答案，而图片描述生成则是将图片输入到 LSTM 产生图片的描述语句。这种处理方式存在的缺点是没有考虑图像的局部信息，只用图像的全局信息去处理问题。于是有部分学者开始引入选择注意(Attention)机制[10]来处理 VQA。这类算法的前提假设是，在图像中

*基金项目：国家自然科学基金(批准号: 61365002, 61272212, 61462045, 61462042)；江西省自然科学基金(批准号: 20142BAB217010)；江西省教育厅科技项目(批准号: GJJ150350)

的某个区域能够显式地找到问题答案所在。他们相同的做法是通过非线性方法（一般采用多层感知器 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计算，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。



Is there a stop sign or a yield sign in the image?

What is the little boy holding in his right hand?

What color is the elephant?

Are there any apples on the table?

图 1 视觉问答样例

Hyeonwoo Noh 等[11] 在 CVPR 2016 提出动态参数预测模型（Dynamic Parameters Prediction，DPPnet），利用 Wenlin Chen 等人[12]的 HashedNet 压缩策略，简单、有效地实现了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，自适应提取问题相关的图像特征，用于解决 VQA 答案的预测,并且在公开的 VQA 数据集上取得了非常优秀的性能结果。虽然 DPPnet 整个网络清晰简单、易于训练，但是此模型是在空域滤波器上进行 Hash，得到权重位置是随机的，没有利用图像的空间信息。受 Wenlin Chen 等[13]的 FreshNets 中 DCTHash 算法的启发，本文先使用我们改造的 VGGnet 对图像进行特征提取，获取保留了空间信息的图像特征，然后采用基于离散余弦哈希（DCTHash）的 Spatial-DCTHash 卷积层来结合图像特征与问题特征。本文在 COCOqa 和 MSCOCO-VQA 数据集上进行了实验，实验结果表明本文的算法在性能上有较大提高。

本文的组织如下：第 2 节介绍 DPPnet 的网络模型及相关工作，第 3 节介绍本文所提出的模型，然后在第 4 节给出本文的实验细节和实验结果，第 5 节进行总结。

2．DPPnet 模型

DPPnet 是 Hyeonwoo Noh 等人在 CVPR2016 中提出的动态参数预测模型。该模型相比同期的方法网络结构简单、性能优越，在 VQA 处理取得了较好的结果。DPPnet 主要亮点在于两个方面：(1) 对问题句的处理；（2）对问题特征和图片特征的融合处理。

- 问题处理：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[14]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 GRU，取 GRU 最后一个时刻输出作为问句的特征表示，如图 2 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[15]向量模型参数作为初始参数。

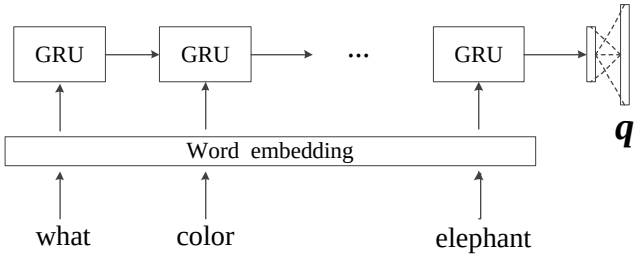


图 2 问题特征提取

➢ 图像处理：DPPnet 采用主流预训练好的 VGGnet-16 作为图像特征抽取网络。首先，将图片归一化成 224*224 大小的 RGB 图像作为网络输入，取倒数第二个全连接层（图 3 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。由于“fc7”维数较高，因此需要进行适当降维。

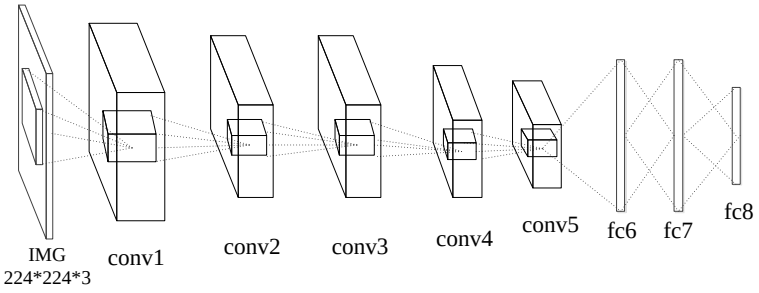


图 3 VGGnet 网络结构

➢ 问题信息与图像信息融合：

在介绍 DPPnet 问题信息与图像信息的融合方式之前，先介绍一下 HashedNet[12]，HashedNet 模型是为了解决神经网络中全连接层的参数过多的问题，参数过多使得神经网络不易于移植到手持终端设备以及嵌入式设备中，同时网络中有些参数对任务的完成并不是起到很大的作用。于是 HashedNet 提出采用权重 hash 的方式减少全连接的网络参数，实现如下：1、首先生成一维的初始化权值（称为真实权重）；2、而全连接层的权重是一个二维的矩阵（二维权重矩阵的元素个数远远大于真实权重的元素个数），采用 hash 函数来处理真实权重的每个元素随机分配到二维权重矩阵，在网络训练时只需要对应更新原先的真实权重即可，如图 4 所示。虽然得到的二维权重矩阵中，真实权重中的同一元素可能多次出现或者不同元素在 hash 时存在碰撞的情况，然而在 HashedNet 文章中证明事实上这些情况并不会严重影响全连接网络的性能。至此 HashedNet 完成了用少量的权值参数替代原先参数庞大的权值矩阵，从而精简了网络。

基于 HashedNet，DPPnet 采用一个动态参数全连接层用来融合图像、问题的特征，并进行答案的预测。其特点在于由问题特征动态生成网络权值参数来代替 HashedNet 的一维真实权值，并通过 HashedNet 的处理方式将一维的网络权值参数将构建成全连接层的网络权值矩阵。得到动态嵌入问题特征为权重的全连接层后，将图片特征输入到该全连接层，得到的输出用于问题答案的预测。

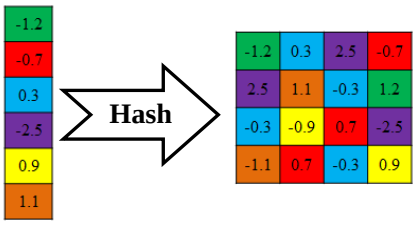


图 4 Hash 权值矩阵生成示意图

DPPnet 在前面的问题处理过程中 GRU 最后一个时刻输出的问题特征经过再经过线性映射 $q = GRU([w_{e1}, w_{e2}, ..., w_{eT}])W + b$，其中 w_{et} 表示第 t 个单词的词嵌入特征，T 表示文体句单词总个数，W、b 分别为权重和偏置，得到的 $q \in \mathbb{R}^N$ 用来表示问题特征。DPPnet 将图像通过卷积层以及全连接层后提取图 3 中的“fc7”层 4096 维特征，再将其在经过一个全连接层得到一个 d 维的图像特征 X。再将图片特征经过 DPPnet 设计的一个动态参数全连接层 $S = \boldsymbol{W}^q X + b$，其中 X 为上述的图像特征，S 为动态参数层的输出，b 为偏置，而 $\boldsymbol{W}^q$ 则为动态参数层的权重参数。DPPnet 将上述得到的问题特征 q 用来生成动态参数作为动态参数层所需的权重 $\boldsymbol{W}^q$。生成动态参数 $\boldsymbol{W}^q$ 过程如下：使用 Hash 函数将所生成的问题特征 q 分配到相应的 hash 位置，具体的分配方法为：$\boldsymbol{W}^q_{u,v} = \xi(u,v) \bullet q_{g(u,v)}$，其中 $\boldsymbol{W}^q_{u,v}$ 为动态参数层的权重矩阵 $\boldsymbol{W}^q$ 在 (u,v) 位置上的权值，$g(u,v) \in \{1,2,...,N\}$，$\xi(u,v) \in \{\pm 1\}$，g, ξ 是

两个独立的 Hash 函数（调用 xxhash[1]接口的两个独立 hash 函数），q 表示问题特征，N 是问题特征 q 的维数。到的输出即是结合了图像和文本信息的综合信息，最后便可用得到的综合信息来预测问题的答案。

3. 空间离散余弦哈希动态参数网络构建

本文旨在提出空间离散余弦哈希动态参数网络模型。以求达到更高的 VQA 回答准确率，Spatial-DCTHash 动态参数网络模型整体框架如图 5 所示。

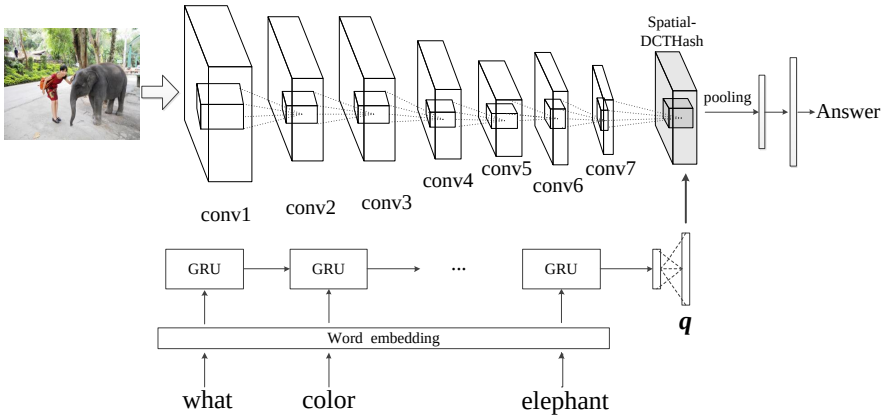


图 5 空间离散余弦哈希动态参数网络模型

首先图像先大小归一化后，输入到本文改造的类似 Fully Convolutional Network 方式的 VGGnet 网络中取“conv7”的输出作为图像特征，这样的图像特征保留了空间分布信息。然后对于问题的处理我们借鉴了 DPPnet 的工作，通过 GRU 后得到问题句子的特征。最后将得到的问题特征输入到本文提出的 Spatial-DCTHash 卷积层作为卷积权重对图像特征进行卷积，再对 Spatial-DCTHash 卷积后的输出做一次最大池化后进行答案预测。

3.1 图像特征提取

DPPnet 采用的是“fc7”全局特征。虽然该特征具有非常强的鉴别力，但是损失了图像的空间信息。根据同期基于 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将对 VQA 的答案生成有益处。因此，本文对传统的 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 6 所示。

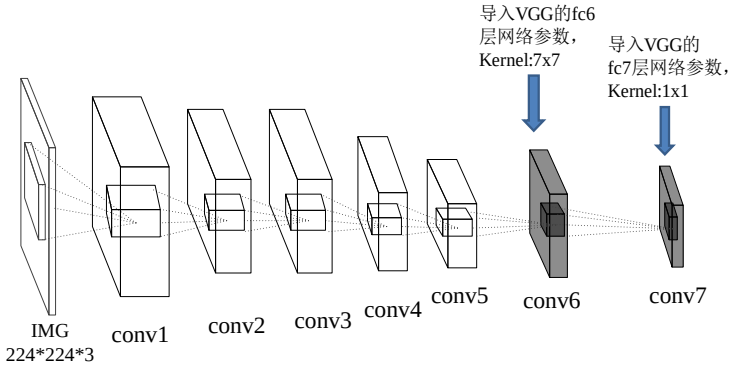


图 6 图像 Conv7 特征的卷积网络

以 VGGnet-16 为例，本文去掉整个网络的后三层全连接层，在最后一个卷积层“conv5”后面加了第 6 层卷积层“conv6”和第七层卷积层“conv7”。

对于“conv6”层，卷积核采用 4096* 512 个 7*7 的滤波器，本文沿用原“fc6”的权值作为这些卷积核的权值参数。对于“conv7”层，卷积核采用 4096 个*4096 个 1*1 的滤波器。同样该卷积核参数沿用“fc7”的权值参数。经过对网络进行如此改造后，便可实现将原先的“fc6”，“fc7”全连接层的权值参数无损转移到“conv6”，“conv7”卷积层中，而且对于输出大小为 512*14*14 的“conv5”特征，经过 “conv6”和 “conv7”两层的卷积操作后得到“conv7”层输出大小为 4096*8*8。“conv7”的特征保留了图像 8*8 的空间分布信息，

[1] https://code.google.com/p/xxhash/

并且每个点的 4096 维特征与原先“fc7”特征的鉴别力相当。

结合空间离散余弦哈希卷积层（Spatial-DCTHash）和图像“conv7”特征的构建，便可以实现在空间上对图像信息和问题信息的多模态融合，实现 VQA 的答案预测。

3.2 空间离散余弦哈希卷积层（Spatial-DCTHash）构建

DPPnet 在生成动态全连接层的过程中，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）上进行位置哈希。由于 CNN 卷积滤波器（即卷积核）在空间上具有一定的结构性并不是完全随机的，如果直接在采用类似 HashedNet 的方式直接在空域的权重矩阵上进行哈希，将会使得滤波器的分布呈现随机化，缺乏必要的图像结构性。因此 Wenlin Chen 等[13]提出 FreshHash，借鉴图像压缩中的余弦变换方法，首先将一维的真实权值在频域（本文中的频域一般指图像经过傅里叶变换、离散余弦变换等等得到的频谱图）先进行 hash 分配，然后将得到的频域矩阵通过离散余弦逆变换得到图像的空域滤波器（即卷积核）。虽然得到频域上的 hash 结果比较随机，但是再通过变换得到空域上的结果就存在一定的机构性，避免了直接 hash 的完全随机。这种方法得到的 CNN 空域滤波器不仅保留了原始 CNN 滤波器的结构性，同时又保留 Hash 的空间压缩的优点。

相比传统 HashNet，FreshNet 可以取得更好的性能，这个在 Wenlin Chen 等[13]的工作中得到有力的验证，如图 7 所示。

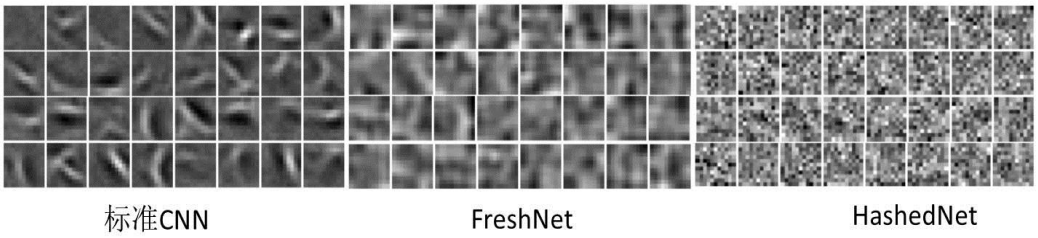


图 7 有关滤波器组可视化效果对比（源自 Wenlin Chen 等[13]）

对于一个空域的权值矩阵 $\boldsymbol{W} \in \mathbb{R}^{d\times d}$，经过离散余弦变换（DCT）得到对应在频域上的输出矩阵 $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{d\times d}$。对应的 DCT 运算定义如下式（1）：

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

其中：

$$u,v = 0,1,2,...,d-1;\ \ x,y = 0,1,2,...,d-1;$$

$$r_v, r_u = \begin{Bmatrix} \sqrt{1/d},\ u,v=0 \\ \sqrt{2/d},\ u,v=1,2,3...,d-1 \end{Bmatrix}$$

为了方便记忆与后面的书写，这里用符号 F_{dct} 来表示 DCT 运算。因此方程式（1）可以写为：$\boldsymbol{\mathcal{W}} = F_{dct}(\boldsymbol{W})$ 。

相应地，如果将频域矩阵 $\boldsymbol{\mathcal{W}}$ 转换为空域上的矩阵，那么需要使用 DCT 的逆变换，如公式（2）所示：

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

同样用符号 F_{dct}^{-1} 表示 DCT 的逆变换，故而公式（2）可以记做：$\boldsymbol{W} = F_{dct}^{-1}(\boldsymbol{\mathcal{W}})$。

如果我们将 CNN 空域滤波器看成一幅图像的话，那么整个深度网络是要学习这幅图像的“最优”表示。DPPnet 所采用的 HashedNet 的方式，在空域上的 hash 位置上“逼近”该表示。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，去掉位于高频的噪音，保留了大部分信息能量的低频部分，可以实现对图像的有效逼近。HashedNet 和 FreshedNet 的显式效果对比如图 7 所示。

图像在 DCT 的频谱分布，左上角是低频，右下角是高频。所以在 DCT 压缩的时候，一般保留左上角绝大部的低频部分，去掉右下角的高频，图像的信息不会损失太多。

类似于 DPPnet 对 HashNet 的修改，我们也相应地对 FreshNet 进行了修改，从而实现了基于空间离散余弦哈希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。

一般对于 CNN 传统的卷积层来说，我们假设输入通道数为I，输出通道数为O，卷积核的大小为$k\times k$，那么总共有$I\times O$个$k\times k$大小的卷积核，且对于这个传统的卷积层它的权值可以表示为一个 4 维的矩阵：$\boldsymbol{W}\in\mathbb{R}^{I\times O\times k\times k}$，这个卷积层总共有$I\times O\times k\times k$个参数。记第 i 输入通道和第 o 输出通道对应的卷积核为$\boldsymbol{W}^{io}\in\mathbb{R}^{k\times k}$，$\boldsymbol{W}_{xy}^{io}$表示第 i 输入通道和第 o 输出通道对应的卷积核上 (x,y)位置上的权值，而$\boldsymbol{\mathcal{W}}_{uv}^{io}$则是该卷积核对应的频域上(u,v)位置上的权值。

本文利用问题特征动态生成权值参数，然后经过 Hash 位置映射到 CNN 卷积核频域矩阵。这样得到的空间 DCTHash 映射的特点是，让频域左上角低频部分的分配的 hash 值更多，不会有太多冲突，从而可学习的不同参数更丰富。让右下角的高频部分 hash 值分配较少些，发生冲突也不会产生太大影响。相比传统的均匀 hash，这样的处理可以尽量保证频域的低频有效部分。生成为频域权值矩阵经过逆离散余弦变换，从而可以得到所需的空域卷积核，实现对图像的卷积操作。整个过程如图 8 所示。

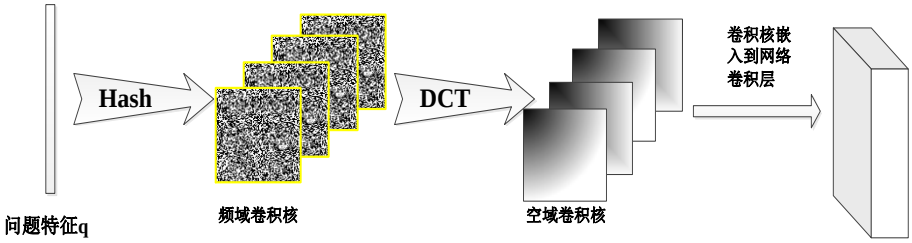


图 8 空间余弦变换哈希动态参数卷积核生成过程

当本文在进行模型训练时，Spatial-DCTHash 卷积层的具体前向后向过程如下：

- 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是问题特征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$\boldsymbol{\mathcal{W}}_{uv}^{io}$中的元素和问题特征$q$的关系可以用下面的公式（3）表示：

$$\boldsymbol{\mathcal{W}}_{uv}^{io}=\xi(i,o,u,v)\bullet q_{g(i,o,u,v)} \tag{3}$$

其中$g(i,o,u,v)\in\{1,2,...,N\}$，$\xi(i,o,u,v)\in\{\pm1\}$，$\bullet$ 表示内积，g,ξ 是两个 Hash 函数且相互独立。然后把我们这个频域上的卷积核矩阵通过 DCT 逆变换转换到空域上的卷积核矩阵：

$$\boldsymbol{W}_{uv}^{io}=F_{dct}^{-1}(\boldsymbol{\mathcal{W}}_{uv}^{io}) \tag{4}$$

用这个嵌入问题特征的卷积核对空间分布的图像特征进行卷积处理，从而实现问题和图片特征的有效融合。

- 反向传播梯度计算过程。

Spatial-DCTHash 生成的卷积核，其权值是用问题特征通过 hash 得到频域权值矩阵$\boldsymbol{\mathcal{W}}$，然后再通过 DCT 逆变换转到空域$\boldsymbol{W}$上的权值。在整个网络反向传播时，我们很容易计算$\boldsymbol{W}$对$\boldsymbol{\mathcal{W}}$偏导，如公式（5）所示：

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

假设整个网络模型的损失函数为Γ，损失函数频域权值$\boldsymbol{\mathcal{W}}_{uv}^{io}$的偏导为：$\frac{\partial\Gamma}{\partial\boldsymbol{\mathcal{W}}_{uv}^{io}}$，根据公式（2）和公式（5）可得：

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

对比上面的公式（1），公式（6）可以改写为：

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

公示（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特征 hash 得到，因此，反向传播时问题特征的权值梯度我们可以相应地如公式（8）所示：

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

在公式（3）中我们可以看到，在计算频域权值矩阵对问题特征的梯度时，则必须知道频域上的权值矩阵和问题特征生成参数的对应关系，即损失函数对问题特征 q 的梯度与 q 在频域权值矩阵的 hash 位置相关。结合公式（3），公式（8）可写为：

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

至此，可以实现损失函数对问题特征的梯度求解。

4. 实验设置与分析

4.1．实验数据集与实验设置

为了验证本文所提出模型的有效性，本文在两个已公开的主流视觉问答数据集（COCOqa 和 MSCOCO-VQA）上进行了对比实验。

COCOqa 数据集由多伦多大学 M. Ren 等人[16]发布。COCOqa 数据集的图片采用微软 MSCOCO 彩色图片数据集，包含了 123287 张图片，其中训练问题集 78736 条，测试问题集 38948 条。问题句最长包含 55 个单词，平均包含 9.65 个单词。

MSCOCO-VQA 是 Stanislaw Antol 等[17]人 2015 年发布的数据集。该数据集采用微软 MSCOCO 数据集中所有的彩色图片作为问答图片，是目前规模最大的 VQA 数据集。其中训练集问题约 248349 条，验证集问题约 121512 条，测试问题集约 244302 条，也是目前使用最广的视觉问答数据集。

为了与 DPPnet 对比，本文将 DPPnet 和 Spatial-DCTHash 网络基本参数设置保持一致。本文所有的实验，均在相同的 hash 压缩率（1：100）下面进行，而且 hash 随机种子和参数均采用与 DPPnet 一样的设定。学习率初始值设为 0.001，批处理的 mini-batch 大小设为 48。优化算法采用 Adam 算法，网络参数初始化为[-0.005,0.005]的均匀分布。

Spatial-DCTHash 卷积层和 DPPnet 的普通 hash 全连接层的输入输出维度均设置为 1000。在整个实验训练过程中，图像特征抽取部分，CNN 的参数保持固定不变，不进行微调，因为微调是一个极具技巧性的工作，能够帮助模型在数据集上取的更好的结果但是与模型本身的优劣无关。本文采用固定的 CNN 是为了避免由于微调不当带来的实验结果的偏差，固定 CNN 有利于我们明确我们所做的模型改进的有效性。

实验程序的运行环境为：Ubuntu 14.04，GPU 为 GTX970。

4.2．实验结果与分析

在 COCOqa 数据集上，本文采用答案准确率（Acc）和刻画 Wu-Palmer 相似性[18]的 WUPS[2]得分作为算法性能的评价标准。WUPS 的计算标准为公式（10）所示：

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

其中 N 表示总共测试样例个数，A^i 和 T^i 分表表示第 i 个测试例子的预测答案和真实答案，$\mu(\bullet,\bullet)$ 表示预测答案和真实答案的 Wu-Palmer 相似性的阈值。这里本文采用常用的两个阈值 0.9 和 0.0 来进行性能比较，在表 1 中分别用“WUPS@0.9”和“WUPS@0.0” 表示,“Acc”表示准确率。在本实验中，所采用的性能评价程序源自 M. Ren 等[16]发布的代码。具体的性能如表 1 所示。

表 1 COCOqa 数据集的性能

	wups@0.9	wups@0.0	Acc
IMG+BOW[16]	66.78	88.99	55.92
2VIS+BLSTM[16]	65.34	88.64	55.09
Ensemble[16]	67.90	89.52	57.84
ConvQA[19]	65.36	88.58	54.95
DPPnet-[CNN-FIXED]	69.61	90.38	59.52
Spatial-DCTHash-[CNN-FIXED]	**69.95**	**90.47**	**60.01**

说明：表中“IMG+BOW”、“2VIS+BLSTM”、“Ensemble”是 M. Ren 等人[16]发布 COCOqa 数据集时提出 baseline 方法的结果。“ConvQA”是 Lin Ma 等人[19]的实验结果，“DPPnet-[CNN-FIXED]”是 DPPnet 在本文中的实验参数设置下 CNN 固定的情况中得到结果，而“Spatial-DCTHash-[CNN-FIXED]”表示本文模型的结果 。

通过表 1 可以看出本文所提出的 Spatial-DCTHash 模型无论是在准确率还是 WUPS 得分上均比其他模型和 COCOqa 数据集的 baseline 都高。这说明本文的模型在 COCOqa 数据集上的确能取得更高的性能。

在 MSCOCO-VQA 数据集上，我们采用 train 数据集进行训练，val 数据集进行验证，分别在 test-dev 和 test-standard 数据集上进行了测试。MSCOCO-VQA 的准确率计算并非简单计算答对测试样例的比例，MSCOCO-VQA 的准确率计算公式如公式（11）所示：

$$\mathrm{Acc}_{\mathrm{VQA}}=\frac{1}{N}\sum_{i=1}^{N}\min\{\frac{\#\text{ humans that provided that answer}}{3},1\} \qquad (11)$$

其中N表示总共测试样例个数，“$\#_{\text{humans that provided that answer}}$”表示预测答案与MSCOCO-VQA人工回答的答案相同的个数。

本文采用在CVPR2016 VQA Challenge 公布的MSCOCO-VQA评价工具进行计算。当计算test-dev和test-standard数据集上的性能时，我们将测试结果上传到VQA Challenge服务器上，由服务器返回计算的性能。具体的性能如表2、表3、表4所示。其中“other”、“number”、“Yes/no”分别表示MSCOCO-VQA这个数据集三种主要的问题类型上的正确率，其中“All”表示的整个数据集上的准确率得分情况。

表 2 MSCOCO-VQA 数据集 Open-ended 任务 Val 集的性能

Val set	Yes/no	number	other	All
DPPnet –[CNN-FIXED]	81.05	33.49	41.04	55.07
Spatial-DCThash-[CNN-FIXED]	80.76	33.85	41.46	**55.21**

说明：为了验证我们的模型相比 DPPnet 更有效，我们首先在 Val 集上进行验证试验，“DPPnet-[CNN-FIXED]”与“Spatial-DCTHash-[CNN-FIXED]”分别表示 DPPnet 和本文模型在相同实验参数设置，固定 CNN 网络不做微调的结果。

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

Test-dev set	Yes/no	number	other	All
Question[17]	75.66	36.7	27.14	48.09
Image[17]	64.01	0.42	3.77	28.13
Q+I[17]	75.55	33.67	37.37	52.64
LSTM Q[17]	78.2	35.68	26.59	48.76
LSTM Q+I[17]	78.94	35.24	36.42	53.74
DPPnet –[CNN-FIXED][11]	80.48	37.20	40.90	56.74

DPPnet[11]	**80.71**	**37.24**	41.69	57.22
Spatial-DCThash-[CNN-FIXED]	80.54	36.81	**42.52**	**57.51**

说明：表中“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 得到特征来预测答案）、“LSTM Q+I”（将问题和图像通过 LSTM 预测答案）是 Stanislaw Antol 等[17]发布 MSCOCO-VQA 数据集时提出 baseline 方法的结果。这里的“DPPnet-[CNN-FIXED]”是 Hyeonwoo Noh 等[11]的 DPPnet 论文中 CNN 固定时的结果，“DPPnet”则是 DPPnet 论文中对整个网络进行了微调的结果，而“Spatial-DCTHash-[CNN-FIXED]”表示本文模型的结果 。

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

Test-standard set	Yes/no	number	other	All
Human[17]	**95.77**	**83.39**	**72.67**	**83.3**
LSTMQ+I[17]	-	-	-	54.06
DPPnet[11]	80.28	36.92	42.24	57.36
Spatial-DCThash-[CNN-FIXED]	80.20	35.29	42.94	**57.50**

说明：表中 “LSTM Q+I” baseline 方法的结果，而“Human”是 MSCOCO-VQA 这个数据集人类回答的结果。“DPPnet”则是 DPPnet 论文中对整个网络进行了微调的结果，而“Spatial-DCTHash-[CNN-FIXED]”表示本文模型的结果 。

从表 2、表 3、表 4 可以看出，本文首先在验证集上进行验证试验，验证集数据规模相对较小，而且可以离线测试（鉴于 MSCOCO-VQA 数据集在线测试有时反馈结果比较迟缓而且 MSCOCO-VQA 测试网站提交次数也有限制，为了验证本文模型的有效性先在训练集上进行训练，在验证集上进行性能验证）。然后对于 MSCOCO-VQA 在线测试数据集 test-dev 和 test-standard 也只使用在训练集上训练得到的模型进行测试将结果提交 MSCOCO-VQA 网站进行性能评测。

在表 3、表 4 中还罗列 MSCOCO-VQA 数据集的 baseline 方法以及人类回答的准确率，这是该数据集的基本参考数据，表 3 中可以看出本文的方法在同样处于 CNN 不做微调的情况下较 DPPnet 高出近 1 个百分点，而且也比 DPPnet 做微调的结果高。在表 4 中对于 test-standard 集进行性能比较，test-standard 较 test-dev 集的测试样例要多很多，但是本文的效果最终还是比 DPPnet 进行微调的结果要高。另外值得一提的是本文在 MSCOCO-VQA 数据集上的训练全部只使用训练集训练，而表 3、表 4 中 DPPnet 的结果是在训练集和验证集上进行训练得到模型的测试结果，总体的训练数据较本文的多。综上显示本文提出的模型能够在性能上超过 DPPnet，能够实现更精确的答案预测。在 test-standard 上，本文不做微调的 Spatial-DCTHash 模型得到的性能也比 DPPnet 经过 CNN 网络微调的结果好，这更能说明本文的模型确实具备更强的处理 VQA 能力。

为了能够比较深入地了解算法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。从中我们可以稍微窥探到，对于大部分“what”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，从而可以显示保留空间结构信息的重要性。

表 5 MSCOCO-VQA 在验证集上每个问题类型准确率 Acc 对比

问题类型	**Spatial-DCTHash-[CNN-FIXED]**	**DPPnet-[CNN-FIXED]**	**问题类型**	**Spatial-DCTHash-[CNN-FIXED]**	**DPPnet-[CNN-FIXED]**
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			

4.3．结果样例展示

本文挑选了几张 VQA 结果的例子进行展示，如图 9 所示，每张图片下面对应的是问题、真实答案（GT）、我们模型 Spatial-DCTHash 的回答结果以及 DPPnet 的回答结果。

Q:what sleeps on the blanket and mattress?
GT: dog
DCThash : dog
DPPnet : cat

Q:what are there sitting on the shelf?
GT: bananas
DCThash : bananas
DPPnet : bananas

Q:what provides all the major food groups?
GT: lunch
DCThash : tray
DPPnet : containers

Q:what is the person with a cowboy hat riding trying to get a cow?
GT: horse
DCThash : horse
DPPnet : horse

Q:how many teddy bear are the doll holding in the living room?
GT: two
DCThash : two
DPPnet : one

Q:what provide the wondrous sight to see?
GT: mountains
DCThash : mountains
DPPnet : mountain

图 9 部分结果样例展示

5. 总结

本文提出了一种有效的、基于空间余弦变换哈希的动态参数预测网络模型的视觉问答算法。算法针对 DPPnet 进行了深入分析，从问题与图像特征融合的 hash 处理方式、图像的空间特征表示网络两方面提出新的处理方式，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。本文在 COCOqa 和 MSCOCO-VQA 数据集上与 DPPnet 进行了实验比较。实验表明，在同等网络参数设置下,本文 Spatial-DCTHash 的结果性能均比 DPPnet 的结果要高。即使在 CNN 模型固定不做微调的劣势情况下，在 MSCOCO-VQA 的 test-dev 和 test-standard 集上也比 DPPnet 做过网络整体微调的准确率高。因此，本文所提出的模型能够达到更高的 VQA 性能，具有较好的优越性。

进一步的工作，结合关注机制来分析图像的空间信息的区域权重，建立问题特征与图像区域的关联关系的计算模型，以更好的预测视觉答案，提高模型的性能。

参考文献

[1]. Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., and Soderland, S. Textrunner: Open information extraction on the web. In HLT-NAACL (Demonstrations), 2007.

[2]. M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In NIPS, 2014.

[3]. Krizhevsky, A., Sutskever, I. and Hinton, G. E.:Imagenet classification with deep convolutional neural networks.NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada

[4]. K. Simonyan, A. Zisserman:Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015 (oral)

[5]. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[6]. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[7]. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS Deep Learning Workshop, 2014. 4, 5, 7

[8]. Mateusz Malinowski, Marcus Rohrbach, Mario Fritz :Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. ICCV 2015

[9]. J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR 2015

[10]. Shih K J, Singh S, Hoiem D. Where To Look: Focus Regions for Visual Question Answering[J]. arXiv preprint arXiv:1511.07394, 2015.

[11]. Hyeonwoo Noh,Paul Hongsuck Seo and Bohyung Han: Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. CVPR 2016

[12]. W. Chen, J. Wilson, S. Tyree, K. Weinberger and Y. Chen, Compressing Neural Networks with the Hashing Trick, Proc. International Conference on Machine Learning (ICML-15).

[13]. Chen W, Wilson J T, Tyree S, et al. Compressing Convolutional Neural Networks[J]. arXiv preprint arXiv:1506.04449, 2015.

[14]. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[15]. Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C]//Advances in Neural Information Processing Systems. 2015: 3276-3284.

[16]. M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In NIPS, 2015.

[17]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh：VQA: Visual Question Answering. ICCV 2015.

[18]. Z.Wu and M. Palmer. Verbs semantics and lexical selection.In ACL, pages 133–138, 1994.

[19]. Lin Ma, Zhengdong Lu, and Hang Li,: Learning to Answer Questions From Image Using Convolutional Neural Network.AAAI2016.

[20]. Aiwen Jiang, Fang Wang, Fatih Porikli, Yi Li：Compositional Memory for Visual Question Answering . arXiv :1511.05676 , November 2011