基于注意力机制的OCR超分蒸馏

1.目的意义

本文的教师网络是OCR识别网络，学生网络是超分网络，本文使用从网络中间进行知识蒸馏的方法将教师网络的文字识别特征转移到学生超分网络，从而让学生网络更好的学习教师网络的文字特征，以得到更好的超分效果。

2.网络模型

2.1教师网络

本文的教师网络是基于OCR的CRNN识别网络，是参考文献[1]的网络模型。如下图所示是CRNN的网络结构图。

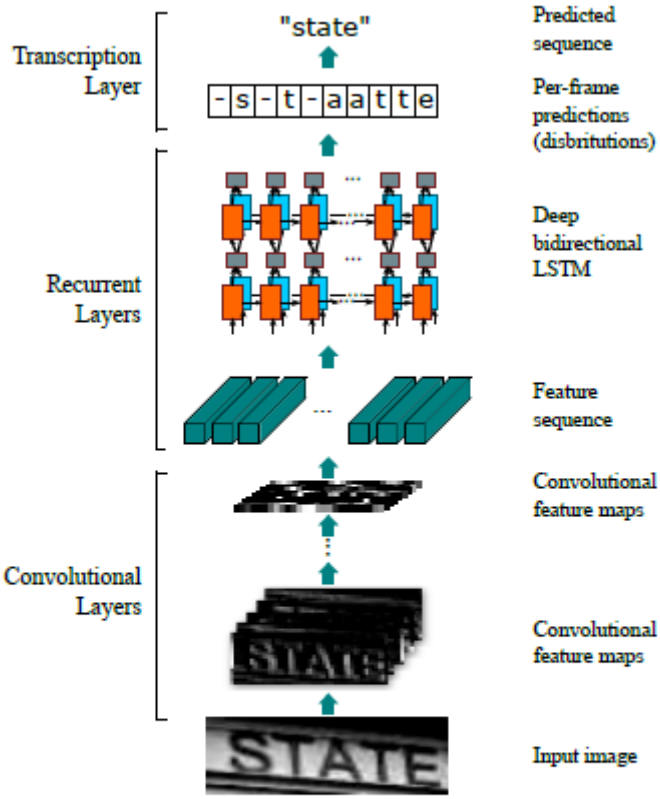

Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

图2.1教师网路CRNN网络架构

如图2.1所示，是CRNN的网络架构，从下往上看主要包括三部分：a)卷积层，从输入图像中提取特征序列；b)循环层，预测每一帧的标签；c)转录层，将每一帧的预测变为最终的标签序列。

如表2.1所示是CRNN网络的具体细节，其中#maps代表每一个卷积的输出通道，k代表卷积核的大小，s代表卷积核每次移动的步长，p代表图像边沿填充的方式，Convolution是由卷积和激活组成。

表2.1网络配置总结

Type	Configurations
Transcription	-
Bidirectional-LSTM	#hidden units:256
Bidirectional-LSTM	#hidden units:256
Map-to-Sequence	-
Convolution	#maps:512, k:2 × 2, s:1, p:0
MaxPooling	Window:1 × 2, s:2
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
MaxPooling	Window:1 × 2, s:2
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:128, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:64, k:3 × 3, s:1, p:1
Input	W × 32 gray-scale image

2.2学生网络

本文所用的学生超分网络是FSRCNN网络，是参考文献[2]的网络模型。如图2.2所示是FSRCNN的网络结构图。

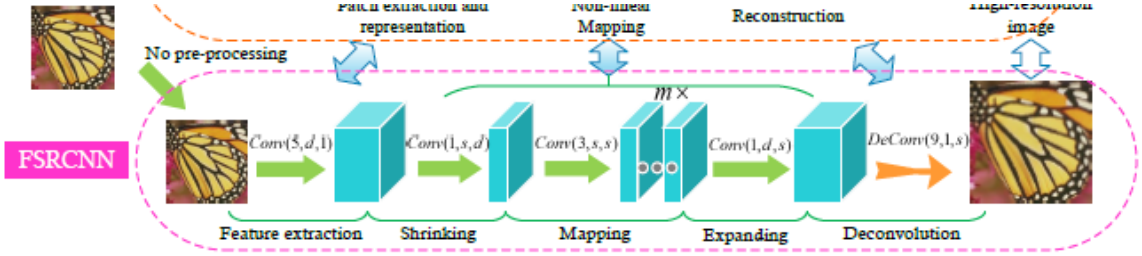

图2.2学生网络FSRCNN网络结构图

本节定义一个卷积$Conv(f_i, n_i . c_i)$,其中f_i是卷积核的大小，n_i是代表输出通道，c_i代表输入通道，其中FSRCNN的网络结构为Conv(5,d,1)-PReLU-Conv(1,s,d)-PReLU-mXConv(3,s,s)-PReLU-Conv(1,d,s)-PReLU-DeConv(9,1,d)。

3.特征提取

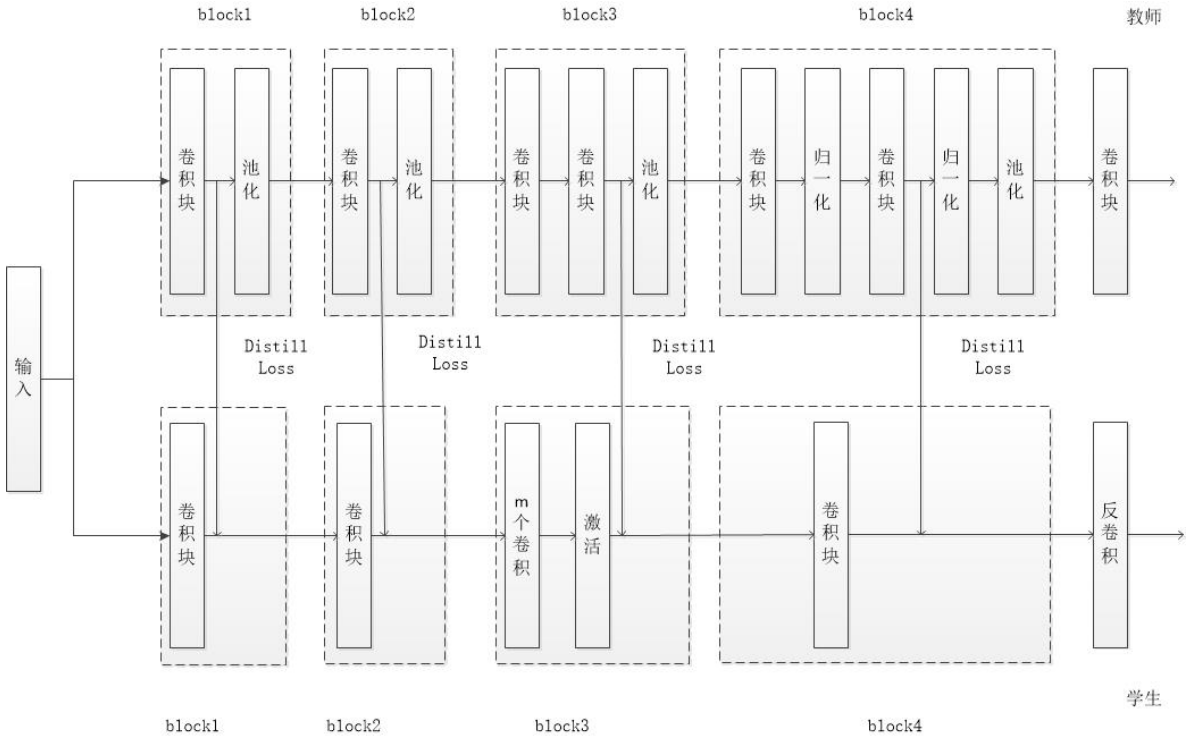

图3.1知识蒸馏网络

如图3.1所示，是教师-学生的知识蒸馏网络。教师网络是采用OCR的CRNN网络，学生网络是采用反卷积的FSRCNN网络。其中教师网络的卷积块是由conv-relu组成，学生网络的卷积块是由conv-prelu组成。

3.1图片大小的计算

a）卷积后图片大小计算

当输入图片的高度为H,宽度为W，卷积后图片的高度H1和宽度W1计算为

$$H1 = (H - k + 2 \times p)/s + 1 \tag{1}$$

$$W1 = (W - k + 2 \times p)/s + 1 \tag{2}$$

其中k是卷积核的大小，p是填充数量，s是卷积核每次移动的步长。

b)池化后图片大小计算

当输入图片的高度为H,宽度为W，池化后图片的高度H1和宽度W1的计算为

$$H1 = (H - k)/s + 1 \tag{3}$$

$$W1 = (W - k)/s + 1 \tag{4}$$

其中k是池化核的大小，s池化核每次移动的步长。

c)反卷积计算公式

$$output = (intput - 1) \times stride + outputpadding - 2 \times padding + kernelsize \tag{5}$$

其中，input是输入的高度\宽度，stride是卷积的步长，outputpadding输出边补充0的层数，高宽都增加padding，padding是输入的每一条边补充0的层数，高宽都增加2*padding，kernelsize是卷积核的大小。

3.2教师网络特征的计算

本文的教师网络特征是取每一个block层的最后一个relu激活函数输出，如图3.1所示，本文提取了4个教师网络的特征图，其中每一个卷积块是由conv-relu组成。设输入图像的张量维度为[N,C,H,W],其中 N 表示这批图像有几张，H 表示图像在竖直方向有多少像素，W 表示水平方向像素数，C 表示通道数。由2.3.1节可知，第一个特征图的输出张量的维度为[N,64,H,W]，第二个特征图的输出张量的维度是[N,128,H/2,W/2]，第三个特征图的输出张量的维度是[N,256,H/4,W/4]，第四个特征图的输出张量的维度是[N,512,H/8,W/8]。

3.3学生网络特征的计算

本文的学生网络特征是取每一个block层的最后一个PReLU激活函数的输出， 如图3.1所示，本文提取了4个学生网络的特征图，其中每一个卷积块是由conv-prelu组成。设输入图像的张量维度为[N,C,H,W]，由于教师网络的第一个特征图的输出张量维度为[N,64,H,W]，学生网络的第一个到第四个特征图是和教师网络的第一个到第四个特征图特征图的张量维度是对应的。

由3.1节的公式可知

学生网络的每个卷积的输入和输出通道，卷积核的大小，步长和padding为

卷积	输入通道	输出通道	卷积核大小	步长	padding
第一个block	C	64	5	1	2
第二个block	64	128	3	2	1
第三个block	128	256	3	2/1	1
第四个block	256	512	5	1	2
反卷积	512	C	8	8	0

4.方法

4.1基于空间注意力的方法

本文利用空间注意力的方法对OCR进行蒸馏，该方法使用教师-学生网络来改善图像超分辨率效果。教师网络通过生成的注意力特征图来指导学生网络的注意力特征图学习，这样学生网络学习到了这些知识，便能够生成尽可能与教师网络相似的特征图。本文的空间注意力方法是参考文献[3]。

本文的工作如下：

a) 将知识蒸馏方法用在超分辨率问题中。利用知识蒸馏将知识从超分辨率教师网络转移到超分辨率学生网络，在不改变其网络结构的前提下，大幅提高了学生网络的超分辨率重建性能；

b)为了确定从教师网络到学生网络的有效知识传递方法，从教师模型的中间层进行学习，要求学习后的学生模型可以很好的保留教师网络中的空间注意力特征。

考虑一个CNN层和对应的激活张量$A \in R^{C\times H\times W}$，它表示有C个通道，每个维度是$H \times W$。基于激活的映射函数F把这个3维张量作为输入，输出一个2维空间特征图：

$$F : R^{C\times H\times W} \rightarrow R^{H\times W} \tag{6}$$

为了定义这个空间注意力映射函数，我们的一个潜在假设是，隐藏层神经元激活的绝对值可以用于指示这个神经元的重要性，这样我们就可以计算通道维度的统计量。

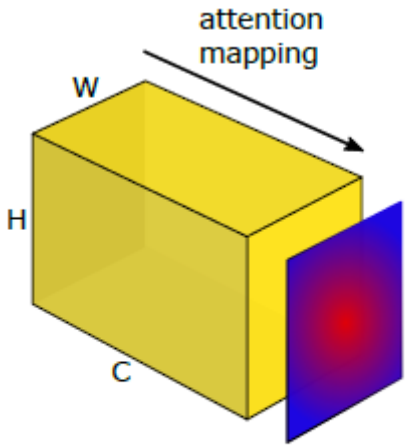

Figure 3: Attention mapping over feature dimension.

具体而言，我们考虑如下的空间注意力图：

a)绝对值求和：

$$F_{sum}(A) = \sum_{i=1}^{C} |A_i| \tag{7}$$

其中，C为输出张量中的通道数。

损失函数

a)

$$L_{MSE}^{SR} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G(I^{LR}))^2 \tag{8}$$

其中，$I_{x,y}^{HR}$是给定的超分图像，$G(I^{LR})$是重建后的超分图像

b)

$$L_{AD}(Q_T, Q_S) = \sum_{j \in I} || \frac{Q_S^j}{||Q_S^j||_2} - \frac{Q_T^j}{||Q_T^j||_2} ||_2 \tag{9}$$

其中，$Q_S^j = vec(F(A_S^j))$表示学生网络的第j个激活图对的向量，$Q_T^j = vec(F(A_T^j))$表示教师网络的第j个激活图对的向量，l为4。

c)

$$L = L_{MSE}^{SR} + \beta L_{AD}(Q_T, Q_S) \tag{10}$$

4.2基于相似性矩阵的方法

本文的相似性矩阵知识蒸馏方法是利用文献[4]中的方法，是利用教师网络的中间层输出特征来指导学生网络的学习。

$$\widetilde{G}_T^{(l)} = Q_T^l \cdot {Q_T^l}^T; G_{T[i,:]}^{(l)} = \widetilde{G}_{T[i,:]}^{(l)} / ||\widetilde{G}_{T[i,:]}^{(l)}||_2 \tag{11}$$

其中，$A_T^l \in R^{b\times c\times h\times w}$是表示教师网络特定层l的激活映射,$Q_T^l \in R^{b\times chw}$是$A_T^l$的重构矩阵，$G_T^{(l)}$是bxb维矩阵是$\widetilde{G}_T^{(l)}$通过L2 normalization得到的，[i,:]表示一个矩阵的第i行，b是一次训练所取的样本数，c是输出的通道数，h和w是空间尺寸，l表示层数。

$$\widetilde{G}_S^{(l)} = Q_S^l \cdot {Q_S^l}^T; G_{S[i,:]}^{(l)} = \widetilde{G}_{S[i,:]}^{(l)} / ||\widetilde{G}_{S[i,:]}^{(l)}||_2 \tag{12}$$

其中，$A_S^{l'} \in R^{b\times c\times h\times w}$是表示学生网络特定层l的激活映射,$Q_S^l \in R^{b\times chw}$是$A_S^{l'}$的重构矩阵，$G_S^{(l)}$是bxb维矩阵是$\widetilde{G}_S^{(l)}$通过L2 normalization得到的，[i,:]表示一个矩阵的第i行，b是一次训练所取的样本数，c是输出的通道数，h和w是空间尺寸，l'表示层数。我们可以定义相似性知识蒸馏损失函数为：

$$L_{SP}(G_T, G_S) = \frac{1}{b^2} \sum_{(l,l')\in I} ||G_T^l - G_S^{l'}||_F^2 \tag{13}$$

其中，L_{SP}是相似性知识蒸馏损失函数，$||\cdot||_F$是Frobenius范数。

Frobenius 范数，简称F-范数，是一种矩阵范数，记为||·||F。矩阵A的Frobenius范数定义为矩阵A各项元素的绝对值平方的总和开根，即$||A||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$，最后用于训练学生网络的总的损失函数为：

$$L = L_{MSE}^{SR} + \gamma L_{SP}(G_T, G_S) \tag{14}$$

其中，L是总的损失函数，去训练学生网络，γ是损失项的权重，便于控制不同项对训练过程的影响程度，L_{MSE}^{SR}是超分损失函数。

4.3知识蒸馏方法的结合

可以将4.1节和4.2节中的知识蒸馏方法结合起来，去更新学生网络的参数，让学生网路从教师网络里面学到更多的知识，则总的损失函数为

$$L = L_{MSE}^{SR} + \beta L_{AD}(Q_T, Q_S) + \gamma L_{SP}(G_T, G_S) \tag{15}$$

其中，β和γ是损失项的权重，便于控制不同项对学生网络训练过程的影响程度。

参考文献

[1]Shi B , Bai X , Yao C . An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(11):2298-2304.

[2]Dong C , Loy C C , Tang X . Accelerating the Super-Resolution Convolutional Neural Network[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[3]Zagoruyko S , Komodakis N . Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer[J]. 2016.

[4]Tung F , Mori G . Similarity-Preserving Knowledge Distillation[C]// International Conference on Computer Vision.