ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 5,Sep. 2018,pp. 801－808
DOI:10. 16337/j. 1004-9037. 2018. 05. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于 Attention 的弱监督多标号图像分类

张　文　谭晓阳

(南京航空航天大学计算机科学与技术学院,南京,211106)

摘　要:深度学习依赖于大数据在很多的任务中取得巨大成功,但目前大部分方法都依赖于严格标注的数据,或者假定仅含一个物体大致位于图片近中心位置且背景较少。而现实场景中背景复杂,出现的物体多样,增加了分类的难度,而且标注的代价很大。本文关注于弱监督场景下的分类任务,提出了基于注意力机制(Attention)结合递归神经网络的深度模型,利用图片级的标注进行多标号学习,利用损失函数进行梯度下降训练自动调整关注区域,使模型每次关注图片的局域区域,并在数据集 PASCAL VOC 2007/2012 上验证算法的有效性,与其他方法相比具有更强的可解释性。

关键词:弱监督;多标号;注意力;深度学习

中图分类号:TP391. 4　　**文献标志码**:A

Weakly-Supervised Multi-label-Classification-Based Attention Mechanism

Zhang Wen, Tan Xiaoyang

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China)

Abstract: Deep learning has become new state-of-the-art framework in many task in big data circumstance. Most of methods need full annotated data or assume only an object in the image with simple background. However, complex background, more than one object in the image and expensive full annotation in the reality, object recognition becomes more challenging. Here, we propose a deep-model-based attention mechanism and recurrent neural network. It trains the network end-to-end on multi-label data with image-level label. The glimpses change along with stochastic gradient descent and focus on different local region in every step. Finally, the effectiveness of the proposed algorithm is verified on the PASCAL VOC 2007 and 2012 datasets. Results show that the network is easily interpretable than other methods.

Key words: weakly supervised; multi-label; attention; deep learning

引　　言

对象识别计算机视觉中的一个重要任务,主要是判定给定图片中是否包含某些特定的对象。近几年,随着网络的发展,数据的获取变得更为容易,卷积神经网络(Convolutional neural network, CNN)[1-3]已成为对象识别中最为流行的框架。它们取得巨大成功主要是因为网络能够从大量的标注数据中学习对象的特征表示,更为吸引人的是学习到的特征能够迁移到别的任务上。特别的,在 ImageNet 上预训

基金项目:中央高校基本科研业务费专项资金(NP2017108)资助项目。
收稿日期:2017-06-12;**修订日期**:2018-01-23

练的模型在其他任务上也取得了不错的效果，例如，物体检测[4]、语义分割[5]、姿态估计[6]等。

弱监督学习(Weakly supervised learning, WSL)[7]是指仅使用图片级(Image-level)的标注，不利用任何关于位置的信息。对于某个给定物体(例如，狗)，只有 Image-level 的标签，如果某张图片标注为正标签，则图片中某个位置有只狗，但是不知道物体的具体位置；如果标注为负，则图片中没该物体。需要注意到在一张图片中可能包含多个物体，弱监督就是利用这样弱标注的数据进行分类器的学习。而传统的多类分类任务，所使用的数据主要物体大致位于图片的中心，背景较少。但是随着数据量的增加，这样要求显得略微苛刻，在实际的场景中，物体可能出现在图片中的任意位置，物体大小未知，背景也是多变的。因此，利用弱监督学习能够大大减少数据标注的代价，对图像分割、图像检测等任务都有很大的意义。

现存的方法大致分成两类，一类是将 WSL 当作多示例(Multiple instance learning, MIL)[8]问题，在该框架下，图片可以看成区域的集合(Bag)，如果图片标注为正包(Positive bag)，则某个区域包含了对应的物体；反之，标注为负(Negative bag)，则没有区域包含物体。多示例主要是在正包中挑选对应物体的区域和学习物体模型之间不断切换，MIL 框架导致优化上的非凸性，最终解的依赖于模型初始化，容易陷入局部最优解。很多新方法尝试克服这些困难，例如寻找更好的初始化或者其他的优化策略。Kumar 等[9]提出逐步将困难样本将入到初始化的训练集合中的自步学习策略，Cinbis 等[10]提出将训练数据分成多份来跳出局部解的方法。Song 等[11]将 Nesterov[12]的光滑方法加入到隐 SVM[13]中从而使得算法对初始化更加鲁棒。和多示例的方法不同，Cabral 等[7]利用柱状图表示(例如 Bag of words)的可加性可将整张图片表示为不同子部分的加权，基于这种性质，可以将一张图片分解成多个类别表示的加权再加上对于背景的损失，然后将该问题当作矩阵补全通过低秩来优化。另一类方法是利用提取候选区(Proposal)[14]或者多尺度[15]，进行端到端卷积神经网络模型学习。深度学习尤其是 CNN 通过大数据的特征学习，该方法在很多不同任务上都取得了很大成功，例如图像分类[2]，物体检测[4]等。但大部分方法都需要详细的图片标注，例如物体矩形框的标注对在错综复杂场景下的图片分类具有明显的效果。最近也有不少利用 CNN 进行弱监督学习的模型，Oquab 等[16]利用预训练 CNN 来计算 PASCAL VOC 图片的中层表示，在其另外工作[15]中，Oquab 设计了一个 CNN 结构利用多尺度训练的方法在预测标号的同时粗略地对物体进行定位。在 Bilen 等[14]的工作中，用提取 Proposal 的方法结合 SPP 网络设计了一个同时进行分类和检测的多任务框架。Durand 等[17]受多示例启发，通过挑选得分高的几个区域表示作为最终表示，并针对分类设计新的排序损失。以上方法取得了一定的效果，但是这些方法也存在着一定的缺点：(1)MIL 方法在优化时经常是个 NP 难问题，对初始化特别敏感；(2)由于没有位置信息，选择候选区比较困难，往往选择的候选区中包含很多的噪声，尤其是几千个候选区中可能只有几个候选区包含真正的对象；(3)模型训练的难度较大，目前训练 CNN 一般是两种方法，一种是提取大量 Proposal 来表示一张图片，另一种是多尺度训练，这两种方法极大影响了模型训练速度。

近几年，人们研究发现人类的认知过程并不是一次性将注意力(Attention)放在整个场景上，相反，逐步关注场景中不同区域的同时抽取相关的信息。基于注意力机制的模型在很多具有挑战性的任务上取得了很好的效果，例如机器翻译[18]、问答系统[19]。文献[20]将其用在生成图片描述上，在文献[21]中利用 RNN 结合 Attention 进行视频动作的识别。虽然注意力机制在不同的任务上都有所涉及，但据了解，目前没有将注意力机制使用在弱监督多标号的分类任务上。

本文提出了一种新的弱监督多标号分类算法，将注意力机制应用到了弱监督学习中，利用 CNN 的卷积特征作为递归神经网络(Recurrent neural network, RNN)的输入，RNN 的每一步只关注于图像的局部区域，下个步骤中自动调整关注的区域，如此进行多步，最终结合所有的预测结果作为最终的预测。相比于其他方法，本文利用注意力机制代替提取 Proposal 方法，避免引入过多的噪声，而且该方法是端到端的，能够依据损失函数自动调整关注区域，网络的训练更为简单。

1　模型和 Attention 机制

本节主要描述整个算法的框架(图 1),包括网络的特征选择、Attention 机制、如何将 RNN 应用到多标号问题、损失函数,以及网络训练方法。

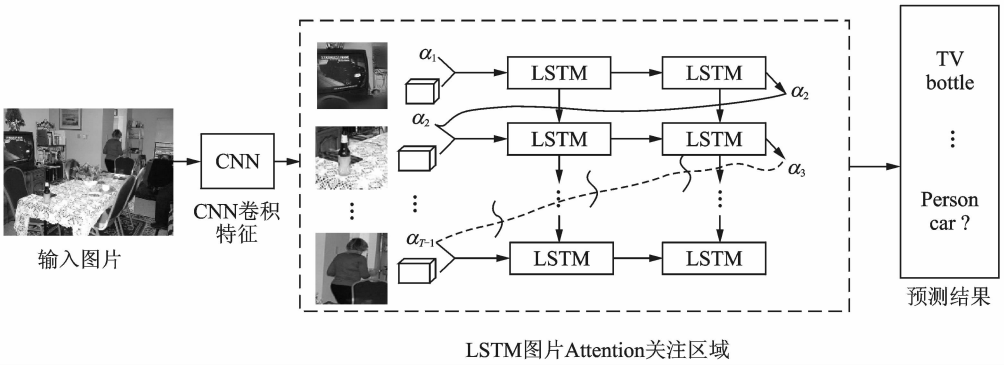

图 1　模型框架

Fig. 1　Framework of model

1.1　卷积特征

本文模型的输入是原始图片经过在 ImageNet 上预训练的 CNN 模型提取的特征,与之前的方法所不同的是,本文提取的是 CNN 卷积特征,假设卷积层的特征图大小为 $D\times K\times K$,可以看作 K^2 个长度为 D 的向量,表示为

$$\boldsymbol{X}=\{x_1,x_2\cdots,x_{k^2}\}\quad x_i\in\mathbf{R}^D \tag{1}$$

上述向量对应了原始图片中的高层语义信息,或者不同的向量对应原始图片中不同的区域,这些区域中包含不同的对象,本文模型的目标就是从 K^2 中找出与任务相关的信息或者区域提取特征学习分类器。

1.2　LSTM 和 Attention

为了能够逐步改变观察到的区域,本文使用了长短期记忆(Long short-term memory, LSTM)[22]网络,该网络具有记忆功能,能够根据之前所记忆的状态影响下一步的更新动作,在序列学习中广泛应用。本文采用的是文献[23]中提出的实现方式,形式化表达如下

$$\begin{pmatrix}i_t\\f_t\\o_t\\g_t\end{pmatrix}=\begin{pmatrix}\sigma\\\sigma\\\sigma\\\tanh\end{pmatrix}\boldsymbol{M}\begin{pmatrix}h_{t-1}\\z_t\end{pmatrix} \tag{2}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t \tag{3}$$

$$h_t=o_t\odot\tanh(c_t) \tag{4}$$

式中:i_t 是输入门,f_t 是记忆门,o_t 是输出门,g_t 由等式计算,c_t 是 Cell 的状态,h_t 是隐状态,x_t 是 LSTM 在 t 时刻的输入,$\boldsymbol{M}:\mathbf{R}^a\rightarrow\mathbf{R}^b$ 是 $a=d+D$ 到 $b=4d$ 的一个仿射变换,其中 d 是 i_t,f_t,o_t,g_t,c_t, 和 h_t 的维度,σ 是 sigmoid 激活函数,$\odot$ 是元素乘法。z_t 是在 t 时刻图像的动态表示,对每个 $x_i,i=1,\cdots,L$ 对应图片中不同位置的特征,每个位置有一个对应权重 α_i 表示第 i 个位置含有对应物体的概率,或者可以表示该位置的重要性,α_i 可以根据前一时刻的隐状态 h_{t-1} 作为多层感知机 f_{att} 的输入来计算。

$$e_{ti}=f_{\text{att}}(h_{t-1}) \tag{5}$$

$$\alpha_{ti}=\frac{\exp(e_{ti})}{\sum_{j=1}^{K^2}\exp(e_{tj})} \tag{6}$$

LSTM 的 Cell 状态和隐状态初始状态可以用卷积特征向量 $\boldsymbol{X}$ 的平均作为输入，经过多层感知机进行初始化。

$$c_0=f_{\mathrm{init},c}\left(\frac{1}{K^2}\sum_{i}^{K^2}x_i\right) \tag{7}$$

$$h_0=f_{\mathrm{init},h}\left(\frac{1}{K^2}\sum_{i}^{K^2}x_i\right) \tag{8}$$

本文所采用的是 Bahdanau 等[18]提出的软注意力(Soft attention)机制，该方法是可导的，可以进行端到端的梯度下降训练，计算式如下

$$E(z_t)=\sum_{i=1}^{K^2}\alpha_{ti}x_i \tag{9}$$

式中：x_i 表示第 i 个长度为 D 的卷积特征，α_{ti} 是在 t 时刻的第 i 个卷积特征的权重，z_t 表示在 t 时刻图片的表示。式(6)中进行的 Softmax 操作可以认为是对高层的卷积特征进行选择的过程，α_{ti} 即是该特征的得分，对应原始图片中该区域是物体的可能性，Softmax 进行了归一化处理，而式(9)是对特征进行加权得到整张图片的表示，该操作十分重要，如何进行特征选择，调整注意力关注的区域尤为重要，结合 LSTM 的状态在每一步进行得分 α_{ti} 的调整，得分变大就代表算法的注意力集中于该区域，该区域是物体的可能性大，反之相反。

1.3 目标函数

本文采用加上惩罚项的逻辑回归(Logistic regression)损失，并对 Attention 的权重加以 $\sum_{t=1}^{T}\alpha_{t,i}\approx 1$ 约束，使得模型整个过程中能够关注于不同的区域，保持区域之间的判别性。

$$E(w)=\sum_{t=1}^{T}\sum_{i=1}^{C}\log\left(1+\mathrm{e}^{-y_i\phi_{t,i}^{y}(x|\boldsymbol{w})}\right)+\gamma\sum_{i=1}^{K^2}\left(1-\sum_{t=1}^{T}\alpha_{t,i}\right)^2+\frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{10}$$

式中：$\phi_{t,i}^{y}(x|\boldsymbol{w})$ 表示 t 时刻对类别 i 的预测，$\boldsymbol{w}$ 表示该模型中相关参数，$y\in\{-1,1\}^C$ 表示图片的标注，C 为类别数量，T 为 LSTM 的长度，γ 是约束的惩罚项，λ 是正则化项。

1.4 训练方法

本文使用在 ImageNet 上预训练的 VGG[1]网络，将所有的图片全部归一化到 $224\times224\times3$ 后，输入到 CNN 中，提取网络最后的卷积特征作为表示，大小为 $14\times14\times512$，即 196×512($K^2\times D$)，经过注意力机制对特征进行选择后，作为 LSTM 输入进行判别，同时对下一个 Attention 进行调整。在实验中，使用训练集和验证集进行训练，在测试集上汇报性能。在网络实现中，使用两个维度都为 256 的 LSTM，为了研究 Step 对实验性能的影响，设置不同长度的 Step 进行对比，并且为了防止过拟合，使用了 Drop-out 层，模型正则化参数为10^{-5}。整个模型采用 Adam[24]优化算法进行训练，实验代码实现使用的是开源的深度学习框架 Tensorflow，实验环境为 Ubuntu 14.04，硬件配置为 Geforce GTX Titan 显存 12 GB，CPU 为 Intel Xeon E5-2630，内存 32 GB。

2 实　　验

2.1 实验数据集

本文在标准的多标号图片库 Pascal VOC 2007 和 VOC 2012 上进行评价，两者是弱监督分类中广泛使用的标准库，图片来自于生活场景，包含了生活中常见的物体，例如飞机、鸟、椅子等，由于图片来自于

无约束的自然场景中,图片中并不是仅包含一个物体,且物体的位置和大小都不确定,关照条件不一致。其中 VOC 2007 数据集中包含了图片训练集 2 501 张,验证集 2 510 张,测试集 5 011 张,以及 20 个类别的标注;VOC 2012 数据集中含有训练集 4 998 张,验证集 5 105 张,测试集 9 637 张,同样是 20 个类别。需要注意的是 VOC 2007 测试集已经有标注信息,而 VOC 2012 测试集的标注无法获得,必须在线提交预测结果并评价。本文在这两个数据集上汇报图像分类任务的性能,评价标准采用的是 Mean average precision(mAP),先计算每个类别的 Average precision(AP),然后计算平均值得到 mAP。

2.2　实验结果及分析

实验中主要使用两个在 ImageNet 上预训练的 CNN 提取卷积特征,即 VGG-CNN-S[1] 和 VGG-CNN-F[1]。在 ImageNet 上 top-5 的精度分别为 18.8%和 13.1%,本文去掉了网络最后面的全连接层,提取卷积层的特征,网络结构如表 1 所示,其中 St 表示卷积层的步长(Stride),Pad 表示扩充像素,卷积核大小表示为 num×size×size,LRN 表示局部归一化(Local response normalisation),以及 Pool 表示 Pooling 层的参数。文献[1]在数据集 Pascal VOC 2007 对该模型微调后汇报性能,两个模型的特征学习能力不同,为了比较的公平性,分别进行卷积特征抽取,特征经过 LSTM 学习 Attention,最后进行分类器学习。训练方法如 1.4 节所述,在 VOC 2007 和 VOC 2012 上,实验中设置 LSTM 的步长为 10,Attention 的惩罚因子为 0.01。

表 1　CNN 结构

Tab. 1　Architectures of CNN

Arch	Conv1	Conv2	Conv3	Conv4	Conv5
VGG-CNN-F	16×11×11 St. 4, Pad 0 LRN, Pool ×2	256×5×5 St. 1, Pad 2 LRN, Pool ×2	256×3×3 St. 1, Pad 1 —	256×3×3 St. 1, Pad 1 —	256×3×3 St. 1, Pad 1 Pool ×2
VGG-CNN-S	97×7×7 St. 2, Pad 0 LRN, Pool ×3	256×5×5 St. 1, Pad 1 Pool ×2	512×3×3 St. 1, Pad 1 —	512×3×3 St. 1, Pad 1 —	512×3×3 St. 1, Pad 1 Pool ×3

表 2 是模型在数据集 VOC 2012 上对比结果,由于对比方法没有给出每个类别的 AP,本文只比较 mAP,mAP 越大表示方法分类性能越高。从表 1 可以看出,本文方法在原有 CNN 的基础上性能都有所提高,尤其是在模型 VGG-F 上,效果提升明显,该模型网络参数较少,特征表示能力有限,在弱监督多标号的问题上学习的特征不能很好表示整张图片,这是因为 VGG 原始模型是利用整张图片进行训练,关注的是全局的信息,在多类分类任务中,主要的物体在图片近中心位置,全局训练方式在该类任务中效果明显,但图片中包含多个物体时,全局的表示方法不能很好表示那些体积较小的物体,导致分类器偏向于体积大的物体,同时混乱的背景对物体的特征学习存在很大影响。本文的模型是局部方法,通过注意力机制对局部信息进行加权后进行分类器学习,与全局训练方法相比,更适合弱监督多标号场景,表 2 结果表明加入注意力机制后对整体分类性能有明显提升。

表 2　VOC 2012 测试集分类结果

Tab. 2　VOC 2012 test classification results

方法	mAP
VGG-S+ attention(本文方法)	84.6
VGG-F+ attention(本文方法)	82.5
VGG-S[1]	83.2
VGG-F[1]	79.9
VGG-M[1]	82.5

数据集 VOC 2007 上的性能对比见表 3，表中各个类别物体分类精度，其中飞机(Aero)、马(Horse)、火车(Train)等，这些类别的体积一般比较大，在图片中所占的位置较大，对比方法都是采用全局的训练方法，能很好的表示这类图片，所以传统方法在这些类别上的精度较高。本文方法在类别鸟(Bird)、狗(Dog)、瓶子(Bottle)、猫(Cat)等上有明显的提升，如图 2 所示，这些类别在图像中一般比较小，提升原因有三：首先，因为采用了 Softmax 操作加上软注意力机制对卷积特征进行选择，使网络能够关注到这些比较小的物体，而不仅仅是关注全局；其次，LSTM 的序列学习能力能利用过去的知识改变所关注的区域，在每步时只关注图片中的局部，学习到的信息存储在自己的状态中；最后，利用设计的损失函数进行端到端的训练，根据 LSTM 的隐状态调整关注区域。传统基于全局信息的方法只是关注单一的全局信息，容易丢失图片中的很多细节信息，本文提出的弱监督多标号分类算法，经过 3 个步骤比传统的方法更适用于多标号问题，通过每个类别的对比，注意力机制的加入对多个物体的分类学习任务有明显帮助。和原始 CNN 模型 VGG[1] 相比，整体 mAP 有 1～2 倍的提升，进一步证明注意力机制方法的有效性，尤其是在体积较小的物体上分类性能明显提升。

表 3 VOC 2007 测试集分类结果

Tab. 3 VOC 2007 test classification results

方法	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
Oquab et al[3]	88.5	81.5	87.9	82.0	47.5	75.5	90.1	87.2	61.6	75.7
VGG_F[1]	88.7	83.9	87.0	84.7	46.9	77.5	86.3	85.4	58.6	71.0
VGG-M-1024[1]	91.4	86.9	89.3	85.8	53.3	79.8	87.8	88.6	59.0	77.2
VGG-S[1]	95.3	90.4	92.5	89.6	54.4	81.9	91.5	91.9	64.1	76.3
VGG-F + attention(本文方法)	90.2	89.0	93.1	88.3	48.2	79.3	93.8	86.4	60.3	72.5
VGG-S +attention(本文方法)	95.6	92.5	93.4	90.2	55.2	82.1	93.5	92.5	65.4	76.1

table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
67.3	85.5	83.5	80.0	95.6	60.8	76.8	58.0	90.4	77.9	77.7
72.6	82.0	87.9	80.7	91.8	58.5	77.4	66.3	89.1	71.3	77.4
73.1	85.9	88.3	83.5	91.8	59.9	81.4	68.3	93.0	74.1	79.9
74.9	89.7	92.2	86.9	95.2	60.7	82.9	68.0	95.5	74.4	82.4
71.5	83.2	90.1	81.2	90.6	59.2	78.4	65.3	91.5	73.6	79.3
75.6	90.2	91.6	87.0	96.2	61.2	82.5	69.3	94.7	76.3	83.1

LSTM 的 Step 的选取对最终的性能有着很大的影响，为了观察该因素对性能的影响，本文实验中设置不同 Step，整体训练并计算 mAP，从图 2 可以看出，在 Step 较小时，观察的区域较少，注意力机制不能很好的学习到关注的区域，特征表示能力弱，整体效果较差。但随着 Step 的增加，模型能够观察到更多的区域，分类精度在上升，在 9，10，11 时，效果达到最好。随后增加 Step 没有明显的提升效果，反而会导致性能下降，这是由于在弱监督多标号中，每个 Step 使用的是相同的监督信息，在 Step 增多的过程中，容易导致每个 Step 趋于相同，最终所关注的区域相同，Attention 机制失效，这个问题可以通过增加 Step 之间的判别性来缓解。

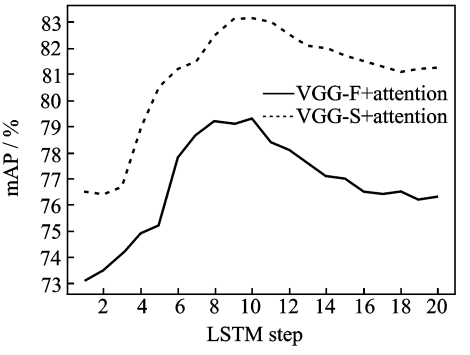

图 2 LSTM step 对 mAP 的影响

Fig. 2 Impact of LSTM step

2.3 可视化

为了能更直观地观察 LSTM 学习的 Attention 关注图像中的哪块区域，本文将学习到的 Attention 权重系数映射回到原始图片进行可视化。在 19 层的 VGG 中只有 Max-pooling 层改变特征图的大小，大小为 224×224 的

原始图片经过 4 个 Max-pooling,到达最后的卷积层,输出大小为 14×14,为了可视化 Attention 的权重,将权重以 16 的比例向上采样,然后使用高斯 Filter 进行可视化。

图 3 是部分图片的最后一个步骤的 Attention 可视化,上一行是原始图片,下面是对应 Attention 的可视化,较亮的地方代表了模型当前所关注的区域。可以很明显地看到,在物体比较小的时候,例如鸟(Bird)和船(Boat),该网络结构依旧能够将注意力集中于对应的区域,这主要是式(5)(6)(9)的效果,使网络有自动特征选择的能力,除此之外本文方法有个明显的优点,图片中有多个物体时,该模型依然能准确注意到,这是采用 LSTM 后的效果,通过自身的记忆功能不断调整关注区域,但不至于全部忘掉之前的状态,能够综合学习到的信息,而不是仅仅关注于一片区域,或者仅关注于全局信息,该可视化方法对卷积神经网络的学习机制理解有一定帮助,更加直观。

图 3　PASCAL VOC 2007 部分图片 Attention 可视化

Fig. 3　Visualization of attention scores on some sample image on PASCAL VOC 2007

3　结束语

本文首次将 Attention 机制结合 LSTM 应用在弱监督多标号图像分类任务中,在不使用任何位置信息前提下,与传统弱监督方法不同,本文利用序列学习的方式,使得模型逐步关注于图像的不同区域,结合 LSTM 的记忆功能,端到端训练分类器,通过实验证明方法在原有基础上性能有所提升,经过可视化 Attention 的权重可以直观观察注意力机制学习的过程,本文方法的可解释性更强,对于理解深度学习的机制有一定帮助。同时本文提出的基于注意力机制的端到端框架在其他任务上同样适用,例如弱监督物体检测、图像分割等。当然,该方法还存在着一定的改进空间,例如,如何在该方法中加入更强的判别性,使注意力机制更容易找到相关对象,以及在更复杂或类别更多的图片中使用注意力机制时,如何解决出现频率低的物体容易被忽略的问题。

参考文献：

[1] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets [EB/OL]. (2014-05-14)[2017-01-12]. http://arXiv.org/abs/1409.2329, 2014.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. Lake, Tahoe, Nevada, USA: Curran Associates Inc, 2012: 1097-1105.

[3] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE Computer Society, 2014: 1717-1724.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 580-587.

[5] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation[J]. Lecture Notes in Computer Science, 2014, 8695: 297-312.

[6] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks[C]//Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 1653-1660.

[7] Cabral R, De L T F, Costeira J P, et al. Matrix completion for weakly-supervised multi-label image classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(1): 121-135.

[8] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1997, 89(1/2): 31-71.

[9] Kumar M P, Packer B, Koller D. Self-paced learning for latent variable models[C]//International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc, 2010: 1189-1197.

[10] Cinbis R G, Verbeek J, Schmid C. Weakly supervisedobject localization with multi-fold multiple instance learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 189-203.

[11] Song H O, Girshick R, Jegelka S, et al. On learning to localize objects with minimal supervision[J]. Eprint Arxiv, 2014, 22(12): 1611-1619.

[12] Nesterov Y. Smooth minimization of non-smooth functions [J]. Mathematical Programming, 2005, 103(1): 127-152.

[13] Felzenszwalb P, Girshick R, Mcallester D, et al. Object detection with discriminatively trained part-based models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[14] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]//Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 2846-2854.

[15] Oquab M, Bottou L, Laptev I, et al. Is object localization for free? —Weakly-supervised learning with convolutional neural networks[C]//Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 685-694.

[16] Oquab M, Bottou L, Laptev I, et al. Weakly supervised object recognition with convolutional neural networks[C]//Computer Vision and Pattern Recognition Columbus. OH, USA: IEEE, 2014: 141-151.

[17] Durand T, Thome N, Cord M. WELDON: Weakly supervised learning of deep convolutional neuralnetworks[C]//Computer Vision and Pattern Recognition Columbus. Las Vegas, NV, USA: IEEE, 2016: 4743-4753.

[18] Bahdanau D, Cho D, Bengio Y. Neural machine translationby jointly learning to align and translate[EB/OL]. (2014-09-01)[2017-01-12]. http://arXiv.org/abs/1409.0473, 2014.

[19] Shih J K, Singh S, Hoiem D. Where to look: Focus regions for visual question answering[C]//Computer Vision and Pattern Recognition Columbus. Las Vegas, NV, USA: IEEE, 2016: 4613-4621.

[20] Xu K, Lei B J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proceedings of International Conference on Machine Learning. Lille, France: JMLR org, 2015: 2048-2057.

[21] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention[EB/OL]. (2015-11-12)[2016-11-25]. https://arxiv.org/pdf/1511.04119, 2015.

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

[23] Zaremba W, Sutskever I, Vinyals O. Reccurent neural network regularization [EB/OL]. (2014-09-08)[2017-01-12]. http://arXiv.org/abs/1409.2329, 2014.

[24] Kingma D P, Ba J. A method for stochastic optimization[EB/OL]. (2014-12-22)[2017-01-12]. http://arXiv.org/abs/1412.6980, 2014.

作者简介：

张文(1992-)，男，硕士，研究方向：计算机视觉、深度学习，E-mail：w.zhang@nuaa.edu.cn。

谭晓阳(1971-)，男，教授，博士生导师，研究方向：人脸识别、机器学习、模式识别、计算机视觉。

（编辑：张　彤）