SnailTyan

VGG论文翻译——中英文对照

| 🗋 1095

VGG论文翻译——中英文对照

Deep Learning

翻译论文汇总：https://github.com/SnailTyan/deep-learning-papers-translation

Very Deep Convolutional Networks for Large-Scale Image Recognition

ABSTRACT

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3 × 3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

摘要

在这项工作中，我们研究了卷积网络深度在大规模的图像识别环境下对准确性的影响。我们的主要贡献是使用非常小的（3×3）卷积滤波器架构对网络深度的增加进行了全面评估，这表明通过将深度推到16-19加权层可以实现对现有技术配置的显著改进。这些发现是我们的ImageNet Challenge 2014提交的基础，我们的团队在定位和分类过程中分别获得了第一名和第二名。我们还表明，我们的表示对于其他数据集泛化的很好，在其它数据集上取得了最好的结果。我们使我们的两个性能最好的ConvNet模型可公开获得，以便进一步研究计算机视觉中深度视觉表示的使用。

1 INTRODUCTION

Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition (Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014; Simonyan & Zisserman, 2014) which has become possible due to the large public image repositories, such as ImageNet (Deng et al., 2009), and high-performance computing systems, such as GPUs or large-scale distributed clusters (Dean et al., 2012). In particular, an important role in the advance of deep visual recognition architectures has been played by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014), which has served as a testbed for a few generations of large-scale image classification systems, from high-dimensional shallow feature encodings (Perronnin et al., 2010) (the winner of ILSVRC-2011) to deep ConvNets (Krizhevsky et al., 2012) (the winner of ILSVRC-2012).

1 引言

卷积网络（ConvNets）近来在大规模图像和视频识别方面取得了巨大成功（Krizhevsky等，2012；Zeiler＆Fergus，2013；Sermanet等，2014；Simonyan＆Zisserman，2014）由于大的公开图像存储库，例如ImageNet，以及高性能计算系统的出现，例如GPU或大规模分布式集群（Dean等，2012），使这成为可能。特别是，在深度视觉识别架构的进步中，ImageNet大型视觉识别挑战（ILSVRC）（Russakovsky等，2014）发挥了重要作用，它已经成为几代大规模图像分类系统的测试台，从高维度浅层特征编码（Perronnin等，2010）（ILSVRC-2011的获胜者）到深层ConvNets（Krizhevsky等，2012）（ILSVRC-2012的获奖者）。

With ConvNets becoming more of a commodity in the computer vision field, a number of attempts have been made to improve the original architecture of Krizhevsky et al. (2012) in a bid to achieve better accuracy. For instance, the best-performing submissions to the ILSVRC-2013 (Zeiler & Fergus, 2013; Sermanet et al., 2014) utilised smaller receptive window size and smaller stride of the first convolutional layer. Another line of improvements dealt with training and testing the networks densely over the whole image and over multiple scales (Sermanet et al., 2014; Howard, 2014). In this paper, we address another important aspect of ConvNet architecture design —— its depth. To this end, we fix other parameters of the architecture, and steadily increase the depth of the network by adding more convolutional layers, which is feasible due to the use of very small (3 × 3) convolution filters in all layers.

随着ConvNets在计算机视觉领域越来越商品化，为了达到更好的准确性，已经进行了许多尝试来改进Krizhevsky等人（2012）最初的架构。例如，ILSVRC-2013（Zeiler＆Fergus，2013；Sermanet等，2014）表现最佳的提交使用了更小的感受窗口尺寸和更小的第一卷积层步长。另一条改进措施在整个图像和多个尺度上对网络进行密集地训练和测试（Sermanet等，2014；Howard，2014）。在本文中，我们解决了ConvNet架构设计的另一个重要方面——其深度。为此，我们修正了架构的其它参数，并通过添加更多的卷积层来稳定地增加网络的深度，这是可行的，因为在所有层中使用非常小的（3×3）卷积滤波器。

As a result, we come up with significantly more accurate ConvNet architectures, which not only achieve the state-of-the-art accuracy on ILSVRC classification and localisation tasks, but are also applicable to other image recognition datasets, where they achieve excellent performance even when used as a part of a relatively simple pipelines (e.g. deep features classified by a linear SVM without fine-tuning). We have released our two best-performing models to facilitate further research.

因此，我们提出了更为精确的ConvNet架构，不仅可以在ILSVRC分类和定位任务上取得的最佳的准确性，而且还适用于其它的图像识别数据集，它们可以获得优异的性能，即使使用相对简单流程的一部分（例如，通过线性SVM分类深度特征而不进行微调）。我们发布了两款表现最好的模型1，以便进一步研究。

The rest of the paper is organised as follows. In Sect. 2, we describe our ConvNet configurations. The details of the image classification training and evaluation are then presented in Sect. 3, and the configurations are compared on the ILSVRC classification task in Sect. 4. Sect. 5 concludes the paper. For completeness, we also describe and assess our ILSVRC-2014 object localisation system in Appendix A, and discuss the generalisation of very deep features to other datasets in Appendix B. Finally, Appendix C contains the list of major paper revisions.

本文的其余部分组织如下。在第2节，我们描述了我们的ConvNet配置。图像分类训练和评估的细节在第3节，并在第4节中在ILSVRC分类任务上对配置进行了比较。第5节总结了论文。为了完整起见，我们还将在附录A中描述和评估我们的ILSVRC-2014目标定位系统，并在附录B中讨论了非常深的特征在其它数据集上的泛化。最后，附录C包含了主要的论文修订列表。

2 CONVNET CONFIGURATIONS

To measure the improvement brought by the increased ConvNet depth in a fair setting, all our ConvNet layer configurations are designed using the same principles, inspired by Ciresan et al. (2011); Krizhevsky et al. (2012). In this section, we first describe a generic layout of our ConvNet configurations (Sect. 2.1) and then detail the specific configurations used in the evaluation (Sect. 2.2). Our design choices are then discussed and compared to the prior art in Sect. 2.3.

2. ConvNet配置

为了衡量ConvNet深度在公平环境中所带来的改进，我们所有的ConvNet层配置都使用相同的规则，灵感来自Ciresan等（2011）；Krizhevsky等人（2012年）。在本节中，我们首先描述我们的ConvNet配置的通用设计（第2.1节），然后详细说明评估中使用的具体配置（第2.2节）。最后，我们的设计选择将在2.3节进行讨论并与现有技术进行比较。

2.1 ARCHITECTURE

During training, the input to our ConvNets is a fixed-size 224 × 224 RGB image. The only preprocessing we do is subtracting the mean RGB value, computed on the training set, from each pixel. The image is passed through a stack of convolutional (conv.) layers, where we use filters with a very small receptive field: 3 × 3 (which is the smallest size to capture the notion of left/right, up/down, center). In one of the configurations we also utilise 1 × 1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3 × 3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2 × 2 pixel window, with stride 2.

在训练期间，我们的ConvNet的输入是固定大小的224×224 RGB图像。我们唯一的预处理是从每个像素中减去在训练集上计算的RGB均值。图像通过一堆卷积（conv.）层，我们使用感受野很小的滤波器：3×3（这是捕获左/右，上/下，中心概念的最小尺寸）。在其中一种配置中，我们还使用了1×1卷积滤波器，可以看作输入通道的线性变换（后面是非线性）。卷积步长固定为1个像素；卷积层输入的空间填充要满足卷积之后保留空间分辨率，即3×3卷积层的填充为1个像素。空间池化由五个最大池化层进行，这些层在一些卷积层之后（不是所有的卷积层之后都是最大池化）。在2×2像素窗口上进行最大池化，步长为2。

A stack of convolutional layers (which has a different depth in different architectures) is followed by three Fully-Connected (FC) layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks.

一堆卷积层（在不同架构中具有不同深度）之后是三个全连接（FC）层：前两个每个都有4096个通道，第三个执行1000维ILSVRC分类，因此包含1000个通道（一个通道对应一个类别）。最后一层是soft-max层。所有网络中全连接层的配置是相同的。

All hidden layers are equipped with the rectification (ReLU (Krizhevsky et al., 2012)) non-linearity. We note that none of our networks (except for one) contain Local Response Normalisation (LRN) normalisation (Krizhevsky et al., 2012): as will be shown in Sect. 4, such normalisation does not improve the performance on the ILSVRC dataset, but leads to increased memory consumption and computation time. Where applicable, the parameters for the LRN layer are those of (Krizhevsky et al., 2012).

所有隐藏层都配备了修正（ReLU（Krizhevsky等，2012））非线性。我们注意到，我们的网络（除了一个）都不包含局部响应规范化（LRN）（Krizhevsky等，2012）：将在第4节看到，这种规范化并不能提高在ILSVRC数据集上的性能，但增加了内存消耗和计算时间。在应用的地方，LRN层的参数是（Krizhevsky等，2012）的参数。

2.2 CONFIGURATIONS

The ConvNet configurations, evaluated in this paper, are outlined in Table 1, one per column. In the following we will refer to the nets by their names (A–E). All configurations follow the generic design presented in Sect. 2.1, and differ only in the depth: from 11 weight layers in the network A (8 conv. and 3 FC layers) to 19 weight layers in the network E (16 conv. and 3 FC layers). The width of conv. layers (the number of channels) is rather small, starting from 64 in the first layer and then increasing by a factor of 2 after each max-pooling layer, until it reaches 512.

Table 1: ConvNet configurations (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as “conv⟨receptive field size⟩-⟨number of channels⟩”. The ReLU activation function is not shown for brevity.

□Table 1

2.2 配置

本文中评估的ConvNet配置在表1中列出，每列一个。接下来我们将按网站名称（A-E）来提及网络。所有配置都遵循2.1节提出的通用设计，并且仅是深度不同：从网络A中的11个加权层（8个卷积层和3个FC层）到网络E中的19个加权层（16个卷积层和3个FC层）。卷积层的宽度（通道数）相当小，从第一层中的64开始，然后在每个最大池化层之后增加2倍，直到达到512。

表1：ConvNet配置（以列显示）。随着更多的层被添加，配置的深度从左（A）增加到右（E）（添加的层以粗体显示）。卷积层参数表示为“conv⟨感受野大小⟩-通道数⟩”。为了简洁起见，不显示ReLU激活功能。

□Table 1

In Table 2 we report the number of parameters for each configuration. In spite of a large depth, the number of weights in our nets is not greater than the number of weights in a more shallow net with larger conv. layer widths and receptive fields (144M weights in (Sermanet et al., 2014)).

Table 2: Number of parameters (in millions).

□Table 2

在表2中，我们报告了每个配置的参数数量。尽管深度很大，我们的网络中权重数量并不大于具有更大卷积层宽度和感受野的较浅网络中的权重数量（144M的权重在（Sermanet等人，2014）中）。

表2：参数数量（百万级别）

Table 2

2.3 DISCUSSION

Our ConvNet configurations are quite different from the ones used in the top-performing entries of the ILSVRC-2012 (Krizhevsky et al., 2012) and ILSVRC-2013 competitions (Zeiler & Fergus, 2013; Sermanet et al., 2014). Rather than using relatively large receptive fields in the first conv. layers (e.g. 11 × 11 with stride 4 in (Krizhevsky et al., 2012), or 7 × 7 with stride 2 in (Zeiler & Fergus, 2013; Sermanet et al., 2014)), we use very small 3 × 3 receptive fields throughout the whole net, which are convolved with the input at every pixel (with stride 1). It is easy to see that a stack of two 3 × 3 conv. layers (without spatial pooling in between) has an effective receptive field of 5 × 5; three such layers have a 7 × 7 effective receptive field. So what have we gained by using, for instance, a stack of three 3 × 3 conv. layers instead of a single 7 × 7 layer? First, we incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative. Second, we decrease the number of parameters: assuming that both the input and the output of a three-layer 3 × 3 convolution stack has C channels, the stack is parametrised by $3(3^2C^2) = 27C^2$ weights; at the same time, a single 7 × 7 conv. layer would require $7^2C^2 = 49C^2$ parameters, i.e. 81% more. This can be seen as imposing a regularisation on the 7 × 7 conv. filters, forcing them to have a decomposition through the 3 × 3 filters (with non-linearity injected in between).

2.3 讨论

我们的ConvNet配置与ILSVRC-2012（Krizhevsky等，2012）和ILSVRC-2013比赛（Zeiler＆Fergus，2013；Sermanet等，2014）表现最佳的参赛提交中使用的ConvNet配置有很大不同。不是在第一卷积层中使用相对较大的感受野（例如，在（Krizhevsky等人，2012）中的11×11，步长为4，或在（Zeiler＆Fergus，2013；Sermanet等，2014）中的7×7，步长为2），我们在整个网络使用非常小的3×3感受野，与输入的每个像素（步长为1）进行卷积。很容易看到两个3×3卷积层堆叠（没有空间池化）有5×5的有效感受野；三个这样的层具有7×7的有效感受野。那么我们获得了什么？例如通过使用三个3×3卷积层的堆叠来替换单个7×7层。首先，我们结合了三个非线性修正层，而不是单一的，这使得决策函数更具判别性。其次，我们减少参数的数量：假设三层3×3卷积堆叠的输入和输出有C个通道，堆叠卷积层的参数为$3(3^2C^2) = 27C^2$个权重；同时，单个7×7卷积层将需要$7^2C^2 = 49C^2$个参数，即参数多81%。这可以看作是对7×7卷积滤波器进行正则化，迫使它们通过3×3滤波器（在它们之间注入非线性）进行分解。

The incorporation of 1 × 1 conv. layers (configuration C, Table 1) is a way to increase the non-linearity of the decision function without affecting the receptive fields of the conv. layers. Even though in our case the 1 × 1 convolution is essentially a linear projection onto the space of the same dimensionality (the number of input and output channels is the same), an additional non-linearity is introduced by the rectification function. It should be noted that 1 × 1 conv. layers have recently been utilised in the “Network in Network” architecture of Lin et al. (2014).

结合1×1卷积层（配置C，表1）是增加决策函数非线性而不影响卷积层感受野的一种方式。即使在我们的案例下，1×1卷积基本上是在相同维度空间上的线性投影（输入和输出通道的数量相同），由修正函数引入附加的非线性。应该注意的是1×1卷积层最近在Lin等人(2014)的“Network in Network”架构中已经得到了使用。

Small-size convolution filters have been previously used by Ciresan et al. (2011), but their nets are significantly less deep than ours, and they did not evaluate on the large-scale ILSVRC dataset. Goodfellow et al. (2014) applied deep ConvNets (11 weight layers) to the task of street number recognition, and showed that the increased depth led to better performance. GoogLeNet (Szegedy et al., 2014), a top-performing entry of the ILSVRC-2014 classification task, was developed independently of our work, but is similar in that it is based on very deep ConvNets(22 weight layers) and small convolution filters (apart from 3 × 3, they also use 1 × 1 and 5 × 5 convolutions). Their network topology is, however, more complex than ours, and the spatial resolution of the feature maps is reduced more aggressively in the first layers to decrease the amount of computation. As will be shown in Sect. 4.5, our model is outperforming that of Szegedy et al. (2014) in terms of the single-network classification accuracy.

Ciresan等人（2011）以前使用小尺寸的卷积滤波器，但是他们的网络深度远远低于我们的网络，他们并没有在大规模的ILSVRC数据集上进行评估。Goodfellow等人（2014）在街道号识别任务中采用深层ConvNets（11个权重层），显示出增加的深度导致了更好的性能。GooLeNet（Szegedy等，2014），ILSVRC-2014分类任务的表现最好的项目，是独立于我们工作之外的开发的，但是类似的是它是基于非常深的ConvNets（22个权重层）和小卷积滤波器（除了3×3，它们也使用了1×1和5×5卷积）。然而，它们的网络拓扑结构比我们的更复杂，并且在第一层中特征图的空间分辨率被更积极地减少，以减少计算量。正如将在第4.5节显示的那样，我们的模型在单网络分类精度方面胜过Szegedy等人（2014）。

3 CLASSIFICATION FRAMEWORK

In the previous section we presented the details of our network configurations. In this section, we describe the details of classification ConvNet training and evaluation.

3 分类框架

在上一节中，我们介绍了我们的网络配置的细节。在本节中，我们将介绍分类ConvNet训练和评估的细节。

3.1 TRAINING

The ConvNet training procedure generally follows Krizhevsky et al. (2012) (except for sampling the input crops from multi-scale training images, as explained later). Namely, the training is carried out by optimising the multinomial logistic regression objective using mini-batch gradient descent (based on back-propagation (LeCun et al., 1989)) with momentum. The batch size was set to 256, momentum to 0.9. The training was regularised by weight decay (the L2 penalty multiplier set to $5 \cdot 10^{-4}$) and dropout regularisation for the first two fully-connected layers (dropout ratio set to 0.5). The learning rate was initially set to 10^{-2}, and then decreased by a factor of 10 when the validation set accuracy stopped improving. In total, the learning rate was decreased 3 times, and the learning was stopped after 370K iterations (74 epochs). We conjecture that in spite of the larger number of parameters and the greater depth of our nets compared to (Krizhevsky et al., 2012), the nets required less epochs to converge due to (a) implicit regularisation imposed by greater depth and smaller conv. filter sizes; (b) pre-initialisation of certain layers.

3.1 训练

ConvNet训练过程通常遵循Krizhevsky等人（2012）（除了从多尺度训练图像中对输入裁剪图像进行采样外，如下文所述）。也就是说，通过使用具有动量的小批量梯度下降（基于反向传播（LeCun等人，1989））优化多项式逻辑回归目标函数来进行训练。批量大小设为256，动量为0.9。训练通过权重衰减（L2惩罚乘子设定为$5 \cdot 10^{-4}$）进行正则化，前两个全连接层执行丢弃正则化（丢弃率设定为0.5）。学习率初始设定为10^{-2}，然后当验证集准确率停止改善时，减少10倍。学习率总共降低3次，学习在37万次迭代后停止（74个epochs）。我们推测，尽管与（Krizhevsky等，2012）相比我们的网络参数更多，网络的深度更大，但网络需要更小的epoch就可以收敛，这是由于（a）由更大的深度和更小的卷积滤波器尺寸引起的隐式正则化，（b）某些层的预初始化。

The initialisation of the network weights is important, since bad initialisation can stall learning due to the instability of gradient in deep nets. To circumvent this problem, we began with training the configuration A (Table 1), shallow enough to be trained with random initialisation. Then, when training deeper architectures, we initialised the first four convolutional layers and the last three fully-connected layers with the layers of net A (the intermediate layers were initialised randomly). We did not decrease the learning rate for the pre-initialised layers, allowing them to change during learning. For random initialisation (where applicable), we sampled the weights from a normal distribution with the zero mean and 10^{-2} variance. The biases were initialised with zero. It is worth noting that after the paper submission we found that it is possible to initialise the weights without pre-training by using the random initialisation procedure of Glorot & Bengio (2010).

网络权重的初始化是重要的，因为由于深度网络中梯度的不稳定，不好的初始化可能会阻碍学习。为了规避这个问题，我们开始训练配置A（表1），足够浅以随机初始化进行训练。然后，当训练更深的架构时，我们用网络A的层初始化前四个卷积层和最后三个全连接层（中间层被随机初始化）。我们没有减少预初始化层的学习率，允许他们在学习过程中改变。对于随机初始化（如果应用），我们从均值为0和方差为10^{-2}的正态分布中采样权重。偏置初始化为零。值得注意的是，在提交论文之后，我们发现可以通过使用Glorot＆Bengio（2010）的随机初始化程序来初始化权重而不进行预训练。

To obtain the fixed-size 224×224 ConvNet input images, they were randomly cropped from rescaled training images (one crop per image per SGD iteration). To further augment the training set, the crops underwent random horizontal flipping and random RGB colour shift (Krizhevsky et al., 2012). Training image rescaling is explained below.

为了获得固定大小的224×224 ConvNet输入图像，它们从归一化的训练图像中被随机裁剪（每个图像每次SGD迭代进行一次裁剪）。为了进一步增强训练集，裁剪图像经过了随机水平翻转和随机RGB颜色偏移（Krizhevsky等，2012）。下面解释训练图像归一化。

Training image size. Let S be the smallest side of an isotropically-rescaled training image, from which the ConvNet input is cropped (we also refer to S as the training scale). While the crop size is fixed to 224 × 224, in principle S can take on any value not less than 224: for $S = 224$ the crop will capture whole-image statistics, completely spanning the smallest side of a training image; for $S \gg 224$ the crop will correspond to a small part of the image, containing a small object or an object part.

训练图像大小。令S是等轴归一化的训练图像的最小边，ConvNet输入从S中裁剪（我们也将S称为训练尺度）。虽然裁剪尺寸固定为224×224，但原则上S可以是不小于224的任何值：对于$S = 224$，裁剪图像将捕获整个图像的统计数据，完全扩展训练图像的最小边；对于$S»224$，裁剪图像将对应于图像的一小部分，包含小对象或对象的一部分。

We consider two approaches for setting the training scale S. The first is to fix S, which corresponds to single-scale training (note that image content within the sampled crops can still represent multi-scale image statistics). In our experiments, we evaluated models trained at two fixed scales: $S = 256$ (which has been widely used in the prior art (Krizhevsky et al., 2012; Zeiler & Fergus, 2013; Sermanet et al., 2014)) and $S = 384$. Given a ConvNet configuration, we first trained the network using $S = 256$. To speed-up training of the $S = 384$ network, it was initialised with the weights pre-trained with S = 256, and we used a smaller initial learning rate of 10^{-3}.

我们考虑两种方法来设置训练尺度S。第一种是修正对应单尺度训练的S（注意，采样裁剪图像中的图像内容仍然可以表示多尺度图像统计）。在我们的实验中，我们评估了以两个固定尺度训练的模型：$S = 256$（已经在现有技术中广泛使用（Krizhevsky等人，2012；Zeiler＆Fergus，2013；Sermanet等，2014））和$S = 384$。给定ConvNet配置，我们首先使用$S = 256$来训练网络。为了加速$S = 384$网络的训练，用$S = 256$预训练的权重来进行初始化，我们使用较小的初始学习率10^{-3}。

The second approach to setting S is multi-scale training, where each training image is individually rescaled by randomly sampling S from a certain range $[S_{min}, S_{max}]$ (we used $S_{min} = 256$ and $S_{max} = 512$). Since objects in images can be of different size, it is beneficial to take this into account during training. This can also be seen as training set augmentation by scale jittering, where a single model is trained to recognise objects over a wide range of scales. For speed reasons, we trained multi-scale models by fine-tuning all layers of a single-scale model with the same configuration, pre-trained with fixed $S = 384$.

设置S的第二种方法是多尺度训练，其中每个训练图像通过从一定范围$[S_{min}，S_{max}]$（我们使用$S_{min} = 256$和$S_{max} = 512$）随机采样S来单独进行归一化。由于图像中的目标可能具有不同的大小，因此在训练期间考虑到这一点是有益的。这也可以看作是通过尺度抖动进行训练集增强，其中单个模型被训练在一定尺度范围内识别对象。为了速度的原因，我们通过对具有相同配置的单尺度模型的所有层进行微调，训练了多尺度模型，并用固定的$S = 384$进行预训练。

3.2 TESTING

At test time, given a trained ConvNet and an input image, it is classified in the following way. First, it is isotropically rescaled to a pre-defined smallest image side, denoted as Q (we also refer to it as the test scale). We note that Q is not necessarily equal to the training scale S (as we will show in Sect. 4, using several values of Q for each S leads to improved performance). Then, the network is applied densely over the rescaled test image in a way similar to (Sermanet et al., 2014). Namely, the fully-connected layers are first converted to convolutional layers (the first FC layer to a 7 × 7 conv. layer, the last two FC layers to 1 × 1 conv. layers). The resulting fully-convolutional net is then applied to the whole (uncropped) image. The result is a class score map with the number of channels equal to the number of classes, and a variable spatial resolution, dependent on the input image size. Finally, to obtain a fixed-size vector of class scores for the image, the class score map is spatially averaged (sum-pooled). We also augment the test set by horizontal flipping of the images; the soft-max class posteriors of the original and flipped images are averaged to obtain the final scores for the image.

3.2 测试

在测试时，给出训练的ConvNet和输入图像，它按以下方式分类。首先，将其等轴地归一化到预定义的最小图像边，表示为Q（我们也将其称为测试尺度）。我们注意到，Q不一定等于训练尺度S（正如我们在第4节中所示，每个S使用Q的几个值会导致性能改进）。然后，网络以类似于（Sermanet等人，2014）的方式密集地应用于归一化的测试图像上。即，全连接层首先被转换成卷积层（第一FC层转换到7×7卷积层，最后两个FC层转换到1×1卷积层）。然后将所得到的全卷积网络应用于整个（未裁剪）图像上。结果是类得分图的通道数等于类别的数量，以及取决于输入图像大小的可变空间分辨率。最后，为了获得图像的类别分数的固定大小的向量，类得分图在空间上平均（和池化）。我们还通过水平翻转图像来增强测试集；将原始图像和翻转图像的soft-max类后验进行平均，以获得图像的最终分数。

Since the fully-convolutional network is applied over the whole image, there is no need to sample multiple crops at test time (Krizhevsky et al., 2012), which is less efficient as it requires network re-computation for each crop. At the same time, using a large set of crops, as done by Szegedy et al. (2014), can lead to improved accuracy, as it results in a finer sampling of the input image compared to the fully-convolutional net. Also, multi-crop evaluation is complementary to dense evaluation due to different convolution boundary conditions: when applying a ConvNet to a crop, the convolved feature maps are padded with zeros, while in the case of dense evaluation the padding for the same crop naturally comes from the neighbouring parts of an image (due to both the convolutions and spatial pooling), which substantially increases the overall network receptive field, so more context is captured. While we believe that in practice the increased computation time of multiple crops does not justify the potential gains in accuracy, for reference we also evaluate our networks using 50 crops per scale (5 × 5 regular grid with 2 flips), for a total of 150 crops over 3 scales, which is comparable to 144 crops over 4 scales used by Szegedy et al. (2014).

由于全卷积网络被应用在整个图像上，所以不需要在测试时对采样多个裁剪图像（Krizhevsky等，2012），因为它需要网络重新计算每个裁剪图像，这样效率较低。同时，如Szegedy等人（2014）所做的那样，使用大量的裁剪图像可以提高准确度，因为与全卷积网络相比，它使输入图像的采样更精细。此外，由于不同的卷积边界条件，多裁剪图像评估是密集评估的补充：当将ConvNet应用于裁剪图像时，卷积特征图用零填充，而在密集评估的情况下，相同裁剪图像的填充自然会来自于图像的相邻部分（由于卷积和空间池化），这大大增加了整个网络的感受野，因此捕获了更多的上下文。虽然我们认为在实践中，多裁剪图像的计算时间增加并不足以证明准确性的潜在收益，但作为参考，我们还在每个尺度使用50个裁剪图像（5×5规则网格，2次翻转）评估了我们的网络，在3个尺度上总共150个裁剪图像，与Szegedy等人(2014)在4个尺度上使用的144个裁剪图像。

3.3 IMPLEMENTATION DETAILS

Our implementation is derived from the publicly available C++ Caffe toolbox (Jia, 2013) (branched out in December 2013), but contains a number of significant modifications, allowing us to perform training and evaluation on multiple GPUs installed in a single system, as well as train and evaluate on full-size (uncropped) images at multiple scales (as described above). Multi-GPU training exploits data parallelism, and is carried out by splitting each batch of training images into several GPU batches, processed in parallel on each GPU. After the GPU batch gradients are computed, they are averaged to obtain the gradient of the full batch. Gradient computation is synchronous across the GPUs, so the result is exactly the same as when training on a single GPU.

3.3 实现细节

我们的实现来源于公开的C++ Caffe工具箱（Jia，2013）（2013年12月推出），但包含了一些重大的修改，使我们能够对安装在单个系统中的多个GPU进行训练和评估，也能训练和评估在多个尺度上（如上所述）的全尺寸（未裁剪）图像。多GPU训练利用数据并行性，通过将每批训练图像分成几个GPU批次，每个GPU并行处理。在计算GPU批次梯度之后，将其平均以获得完整批次的梯度。梯度计算在GPU之间是同步的，所以结果与在单个GPU上训练完全一样。

While more sophisticated methods of speeding up ConvNet training have been recently proposed (Krizhevsky, 2014), which employ model and data parallelism for different layers of the net, we have found that our conceptually much simpler scheme already provides a speedup of 3.75 times on an off-the-shelf 4-GPU system, as compared to using a single GPU. On a system equipped with four NVIDIA Titan Black GPUs, training a single net took 2–3 weeks depending on the architecture.

最近提出了更加复杂的加速ConvNet训练的方法（Krizhevsky，2014），它们对网络的不同层之间采用模型和数据并行，我们发现我们概念上更简单的方案与使用单个GPU相比，在现有的4-GPU系统上已经提供了3.75倍的加速。在配备四个NVIDIA Titan Black GPU的系统上，根据架构训练单个网络需要2-3周时间。

4 CLASSIFICATION EXPERIMENTS

Dataset. In this section, we present the image classification results achieved by the described ConvNet architectures on the ILSVRC-2012 dataset (which was used for ILSVRC 2012–2014 challenges). The dataset includes images of 1000 classes, and is split into three sets: training (1.3M images), validation (50K images), and testing (100K images with held-out class labels). The classification performance is evaluated using two measures: the top-1 and top-5 error. The former is a multi-class classification error, i.e. the proportion of incorrectly classified images; the latter is the main evaluation criterion used in ILSVRC, and is computed as the proportion of images such that the ground-truth category is outside the top-5 predicted categories.

4 分类实验

数据集。在本节中，我们介绍了描述的ConvNet架构（用于ILSVRC 2012-2014挑战）在ILSVRC-2012数据集上实现的图像分类结果。数据集包括1000个类别的图像，并分为三组：训练（130万张图像），验证（5万张图像）和测试（留有类标签的10万张图像）。使用两个措施评估分类性能：top-1和top-5错误率。前者是多类分类误差，即不正确分类图像的比例；后者是ILSVRC中使用的主要评估标准，并且计算为图像真实类别在前5个预测类别之外的图像比例。

For the majority of experiments, we used the validation set as the test set. Certain experiments were also carried out on the test set and submitted to the official ILSVRC server as a “VGG” team entry to the ILSVRC-2014 competition (Russakovsky et al., 2014).

对于大多数实验，我们使用验证集作为测试集。在测试集上也进行了一些实验，并将其作为ILSVRC-2014竞赛（Russakovsky等，2014）“VGG”小组的输入提交到了官方的ILSVRC服务器。

4.1 SINGLE SCALE EVALUATION

We begin with evaluating the performance of individual ConvNet models at a single scale with the layer configurations described in Sect. 2.2. The test image size was set as follows: $Q = S$ for fixed S, and $Q = 0.5(S_{min} + S_{max})$ for jittered $S \in [S_{min}, S_{max}]$. The results of are shown in Table 3.

Table 3: ConvNet performance at a single test scale.

Table 3

4.1 单尺度评估

我们首先评估单个ConvNet模型在单尺度上的性能，其层结构配置如2.2节中描述。测试图像大小设置如下：对于固定S的$Q = S$，对于抖动$S \in [S_{min}, S_{max}]$，$Q = 0.5(S_{min} + S_{max})$。结果如表3所示。

表3：在单测试尺度的ConvNet性能

Table 3

First, we note that using local response normalisation (A-LRN network) does not improve on the model A without any normalisation layers. We thus do not employ normalisation in the deeper architectures (B–E).

首先，我们注意到，使用局部响应归一化（A-LRN网络）在没有任何归一化层的情况下，对模型A没有改善。因此，我们在较深的架构（B-E）中不采用归一化。

Second, we observe that the classification error decreases with the increased ConvNet depth: from 11 layers in A to 19 layers in E. Notably, in spite of the same depth, the configuration C (which contains three 1 × 1 conv. layers), performs worse than the configuration D, which uses 3 × 3 conv. layers throughout the network. This indicates that while the additional non-linearity does help (C is better than B), it is also important to capture spatial context by using conv. filters with non-trivial receptive fields (D is better than C). The error rate of our architecture saturates when the depth reaches 19 layers, but even deeper models might be beneficial for larger datasets. We also compared the net B with a shallow net with five 5 × 5 conv. layers, which was derived from B by replacing each pair of 3 × 3 conv. layers with a single 5 × 5 conv. layer (which has the same receptive field as explained in Sect. 2.3). The top-1 error of the shallow net was measured to be 7% higher than that of B (on a center crop), which confirms that a deep net with small filters outperforms a shallow net with larger filters.

第二，我们观察到分类误差随着ConvNet深度的增加而减小：从A中的11层到E中的19层。值得注意的是，尽管深度相同，配置C（包含三个1×1卷积层）比在整个网络层中使用3×3卷积的配置D更差。这表明，虽然额外的非线性确实有帮助（C优于B），但也可以通过使用具有非平凡感受野（D比C好）的卷积滤波器来捕获空间上下文。当深度达到19层时，我们架构的错误率饱和，但更深的模型可能有益于较大的数据集。我们还将网络B与具有5×5卷积层的浅层网络进行了比较，浅层网络可以通过用单个5×5卷积层替换B中每对3×3卷积层得到（其具有相同的感受野如第2.3节所述）。测量的浅层网络top-1错误率比网络B的top-1错误率（在中心裁剪图像上）高7%，这证实了具有小滤波器的深层网络优于具有较大滤波器的浅层网络。

Finally, scale jittering at training time ($S \in [256; 512]$) leads to significantly better results than training on images with fixed smallest side ($S = 256$ or $S = 384$), even though a single scale is used at test time. This confirms that training set augmentation by scale jittering is indeed helpful for capturing multi-scale image statistics.

最后，训练时的尺度抖动（$S \in [256; 512]$）得到了与固定最小边（$S = 256$或$S = 384$）的图像训练相比更好的结果，即使在测试时使用单尺度。这证实了通过尺度抖动进行的训练集增强确实有助于捕获多尺度图像统计。

4.2 MULTI-SCALE EVALUATION

Having evaluated the ConvNet models at a single scale, we now assess the effect of scale jittering at test time. It consists of running a model over several rescaled versions of a test image (corresponding to different values of Q), followed by averaging the resulting class posteriors. Considering that a large discrepancy between training and testing scales leads to a drop in performance, the models trained with fixed S were evaluated over three test image sizes, close to the training one: $Q = S - 32, S, S + 32$. At the same time, scale jittering at training time allows the network to be applied to a wider range of scales at test time, so the model trained with variable $S \in [S_{min}; S_{max}]$ was evaluated over a larger range of sizes Q = {S_{min}, 0.5(S_{min} + S_{max}), S_{max}}.

4.2 多尺度评估

在单尺度上评估ConvNet模型后，我们现在评估测试时尺度抖动的影响。它包括在一张测试图像的几个归一化版本上运行模型（对应于不同的Q值），然后对所得到的类别后验进行平均。考虑到训练和测试尺度之间的巨大差异会导致性能下降，用固定S训练的模型在三个测试图像尺度上进行了评估，接近于训练一次：$Q = S - 32, S, S + 32$。同时，训练时的尺度抖动允许网络在测试时应用于更广的尺度范围，所以用变量$S \in [S_{min}; S_{max}]$训练的模型在更大的尺寸范围Q = {S_{min}, 0.5(S_{min} + S_{max}), S_{max}}上进行评估。

The results, presented in Table 4, indicate that scale jittering at test time leads to better performance (as compared to evaluating the same model at a single scale, shown in Table 3). As before, the deepest configurations (D and E) perform the best, and scale jittering is better than training with a fixed smallest side S. Our best single-network performance on the validation set is `24.8%/7.5% top-1/top-5` error (highlighted in bold in Table 4). On the test set, the configuration E achieves 7.3% top-5 error.

Table 4: ConvNet performance at multiple test scales.

Table 4

表4中给出的结果表明，测试时的尺度抖动导致了更好的性能（与在单一尺度上相同模型的评估相比，如表3所示）。如前所述，最深的配置（D和E）执行最佳，并且尺度抖动优于使用固定最小边S的训练。我们在验证集上的最佳单网络性能为 `24.8%/7.5% top-1/top-5` 的错误率（在表4中用粗体突出显示）。在测试集上，配置E实现了 `7.3% top-5` 的错误率。

表4：在多个测试尺度上的ConvNet性能

Table 4

4.3 MULTI-CROP EVALUATION

In Table 5 we compare dense ConvNet evaluation with mult-crop evaluation (see Sect. 3.2 for details). We also assess the complementarity of the two evaluation techniques by averaging their soft-max outputs. As can be seen, using multiple crops performs slightly better than dense evaluation, and the two approaches are indeed complementary, as their combination outperforms each of them. As noted above, we hypothesize that this is due to a different treatment of convolution boundary conditions.

Table 5: ConvNet evaluation techniques comparison. In all experiments the training scale S was sampled from [256; 512], and three test scales Q were considered: {256, 384, 512}.

Table 5

4.3 多裁剪图像评估

在表5中，我们将稠密ConvNet评估与多裁剪图像评估进行比较（细节参见第3.2节）。我们还通过平均其soft-max输出来评估两种评估技术的互补性。可以看出，使用多裁剪图像表现比密集评估略好，而且这两种方法确实是互补的，因为它们的组合优于其中的每一种。如上所述，我们假设这是由于卷积边界条件的不同处理。

表5：ConvNet评估技术比较。在所有的实验中训练尺度S从[256；512]采样，三个测试适度Q考虑：{256, 384, 512}。

Table 5

4.4 CONVNET FUSION

Up until now, we evaluated the performance of individual ConvNet models. In this part of the experiments, we combine the outputs of several models by averaging their soft-max class posteriors. This improves the performance due to complementarity of the models, and was used in the top ILSVRC submissions in 2012 (Krizhevsky et al., 2012) and 2013 (Zeiler & Fergus, 2013; Sermanet et al., 2014).

4.4 卷积网络融合

到目前为止，我们评估了ConvNet模型的性能。在这部分实验中，我们通过对soft-max类别后验进行平均，结合了几种模型的输出。由于模型的互补性，这提高了性能，并且在了2012年（Krizhevsky等，2012）和2013年（Zeiler＆Fergus，2013；Sermanet等，2014）ILSVRC的顶级提交中使用。

The results are shown in Table 6. By the time of ILSVRC submission we had only trained the single-scale networks, as well as a multi-scale model D (by fine-tuning only the fully-connected layers rather than all layers). The resulting ensemble of 7 networks has 7.3% ILSVRC test error. After the submission, we considered an ensemble of only two best-performing multi-scale models (configurations D and E), which reduced the test error to 7.0% using dense evaluation and 6.8% using combined dense and multi-crop evaluation. For reference, our best-performing single model achieves 7.1% error (model E, Table 5).

Table 6: Multiple ConvNet fusion results.

Table 6

结果如表6所示。在ILSVRC提交的时候，我们只训练了单规模网络，以及一个多尺度模型D（仅在全连接层进行微调而不是所有层）。由此产生的7个网络组合具有7.3%的ILSVRC测试误差。在提交之后，我们考虑了只有两个表现最好的多尺度模型（配置D和E）的组合，它使用密集评估将测试误差降低到7.0%，使用密集评估和多裁剪图像评估将测试误差降低到6.8%。作为参考，我们表现最佳的单模型达到7.1%的误差（模型E，表5）。

表6：多个卷积网络融合结果

Table 6

4.5 COMPARISON WITH THE STATE OF THE ART

Finally, we compare our results with the state of the art in Table 7. In the classification task of ILSVRC-2014 challenge (Russakovsky et al., 2014), our “VGG” team secured the 2nd place with
7.3% test error using an ensemble of 7 models. After the submission, we decreased the error rate to 6.8% using an ensemble of 2 models.

Table 7: Comparison with the state of the art in ILSVRC classification. Our method is denoted as "VGG". Only the results obtained without outside training data are reported.

Table 7

4.5 与最新技术比较

最后，我们在表7中与最新技术比较我们的结果。在ILSVRC-2014挑战的分类任务（Russakovsky等，2014）中，我们的"VGG"团队获得了第二名，
使用7个模型的组合取得了7.3%测试误差。提交后，我们使用2个模型的组合将错误率降低到6.8%。

表7：在ILSVRC分类中与最新技术比较。我们的方法表示为"VGG"。报告的结果没有使用外部数据。

Table 7

As can be seen from Table 7, our very deep ConvNets significantly outperform the previous generation of models, which achieved the best results in the ILSVRC-2012 and ILSVRC-2013 competitions. Our result is also competitive with respect to the classification task winner (GoogLeNet with 6.7% error) and substantially outperforms the ILSVRC-2013 winning submission Clarifai, which achieved 11.2% with outside training data and 11.7% without it. This is remarkable, considering that our best result is achieved by combining just two models —— significantly less than used in most ILSVRC submissions. In terms of the single-net performance, our architecture achieves the best result (7.0% test error), outperforming a single GoogLeNet by 0.9%. Notably, we did not depart from the classical ConvNet architecture of LeCun et al. (1989), but improved it by substantially increasing the depth.

从表7可以看出，我们非常深的ConvNets显著优于前一代模型，在ILSVRC-2012和ILSVRC-2013竞赛中取得了最好的结果。我们的结果对于分类任务获胜者（GoogLeNet具有6.7%的错误率）也具有竞争力，并且大大优于ILSVRC-2013获胜者Clarifai的提交，其使用外部训练数据取得了11.2%的错误率，没有外部数据则为11.7%。这是非常显著的，考虑到我们最好的结果是仅通过组合两个模型实现的——明显少于大多数ILSVRC提交。在单网络性能方面，我们的架构取得了最好节果（7.0%测试误差），超过单个GoogLeNet 0.9%。值得注意的是，我们并没有偏离LeCun（1989）等人经典的ConvNet架构，但通过大幅增加深度改善了它。

5 CONCLUSION

In this work we evaluated very deep convolutional networks (up to 19 weight layers) for large-scale image classification. It was demonstrated that the representation depth is beneficial for the classification accuracy, and that state-of-the-art performance on the ImageNet challenge dataset can be achieved using a conventional ConvNet architecture (LeCun et al., 1989; Krizhevsky et al., 2012) with substantially increased depth. In the appendix, we also show that our models generalise well to a wide range of tasks and datasets, matching or outperforming more complex recognition pipelines built around less deep image representations. Our results yet again confirm the importance of depth in visual representations.

5 结论

在这项工作中，我们评估了非常深的卷积网络（最多19个权重层）用于大规模图像分类。已经证明，表示深度有利于分类精度，并且深度大大增加的传统ConvNet架构（LeCun等，1989；Krizhevsky等，2012）可以实现ImageNet挑战数据集上的最佳性能。在附录中，我们还显示了我们的模型很好地泛化到各种各样的任务和数据集上，可以匹敌或超越更复杂的识别流程，其构建围绕不深的图像表示。我们的结果再次证实了深度在视觉表示中的重要性。

ACKNOWLEDGEMENTS

This work was supported by ERC grant VisRec no. 228180. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

致谢

这项工作得到ERC授权的VisRec编号228180的支持.我们非常感谢NVIDIA公司捐赠GPU为此研究使用。

REFERENCES

Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. CoRR, abs/1412.0623, 2014.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In Proc. BMVC., 2014.

Cimpoi, M., Maji, S., and Vedaldi, A. Deep convolutional filter banks for texture recognition and segmentation. CoRR, abs/1411.6836, 2014.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. Flexible, high performance convolutional neural networks for image classification. In IJCAI, pp. 1237–1242, 2011.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., and Ng, A. Y. Large scale distributed deep networks. In NIPS, pp. 1232–1240, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proc. CVPR, 2009.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. CoRR, abs/1310.1531, 2013.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. The Pascal visual object classes challenge: A retrospective. IJCV, 111(1):98–136, 2015.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In IEEE CVPR Workshop of Generative Model Based Vision, 2004.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524v5, 2014. Published in Proc. CVPR, 2014.

Gkioxari, G., Girshick, R., and Malik, J. Actions and attributes from wholes and parts. CoRR, abs/1412.2604, 2014.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proc. AISTATS, volume 9, pp. 249–256, 2010.

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In Proc. ICLR, 2014.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

He, K., Zhang, X., Ren, S., and Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR, abs/1406.4729v2, 2014.

Hoai, M. Regularized max pooling for image categorization. In Proc. BMVC., 2014.

Howard, A. G. Some improvements on deep convolutional neural network based image classification. In Proc. ICLR, 2014.

Jia, Y. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.

Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. CoRR, abs/1404.5997, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In NIPS, pp. 1106–1114, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551, 1989.

Lin, M., Chen, Q., and Yan, S. Network in network. In Proc. ICLR, 2014.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. CoRR, abs/1411.4038, 2014.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In Proc. CVPR, 2014.

Perronnin, F., Sá nchez, J., and Mensink, T. Improving the Fisher kernel for large-scale image classification. In Proc. ECCV, 2010.

Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. CNN Features off-the-shelf: an Astounding Baseline for Recognition. CoRR, abs/1403.6382, 2014.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. CoRR, abs/1409.0575, 2014.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In Proc. ICLR, 2014.

Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. CoRR, abs/1406.2199, 2014. Published in Proc. NIPS, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. CoRR, abs/1409.4842, 2014.

Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S. CNN: Single-label to multi-label. CoRR, abs/1406.5726, 2014.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. Published in Proc. ECCV, 2014.

# Deep Learning

‹ VGG论文翻译——中文版

Caffe模型的Forward时间及GPU显存对比 ›