Stacked Hourglass Networks for Human Pose Estimation

**keywords**

* 人体姿态估计 Human Pose Estimation 给定单张RGB图像，输出人体某些关键点的精确像素位置.
* 全卷积网络 Stacked Hourglass Networks
* 多尺度特征 Features processed across all scales
* 特征用于捕捉人体的空间关系 Capture spatial relationships associated with body
* 中间监督 Intermediate supervision

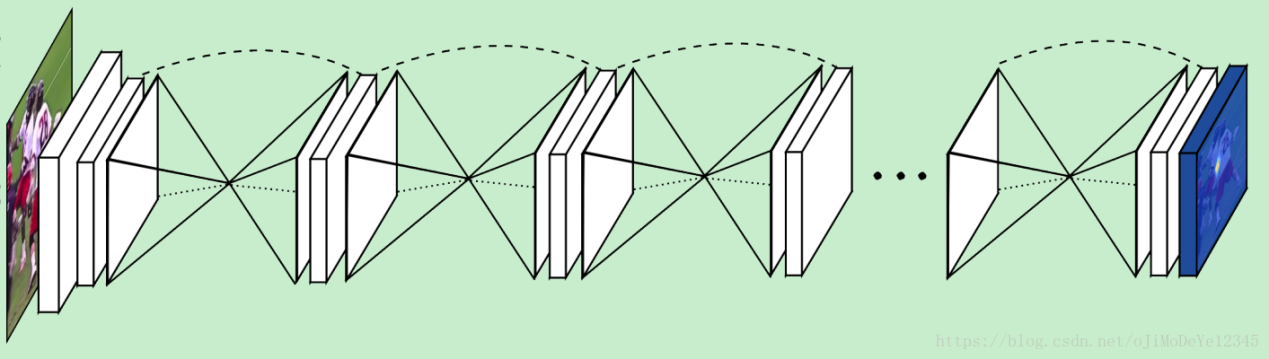


图 - Stacked Hourglass Networks由多个 stacked hourglass 模块组成，通过重复进行bottom-up, top-down推断以估计人体姿态.

**沙漏设计 Hourglass Design**

动机：捕捉不同尺度下图片所包含的信息.

局部信息，对于比如脸部、手部等等特征很有必要，而最终的姿态估计需要对整体人体一致理解. 不同尺度下，可能包含了很多有用信息，比如人体的方位、肢体的动作、相邻关节点的关系等等.

Hourglass设计：

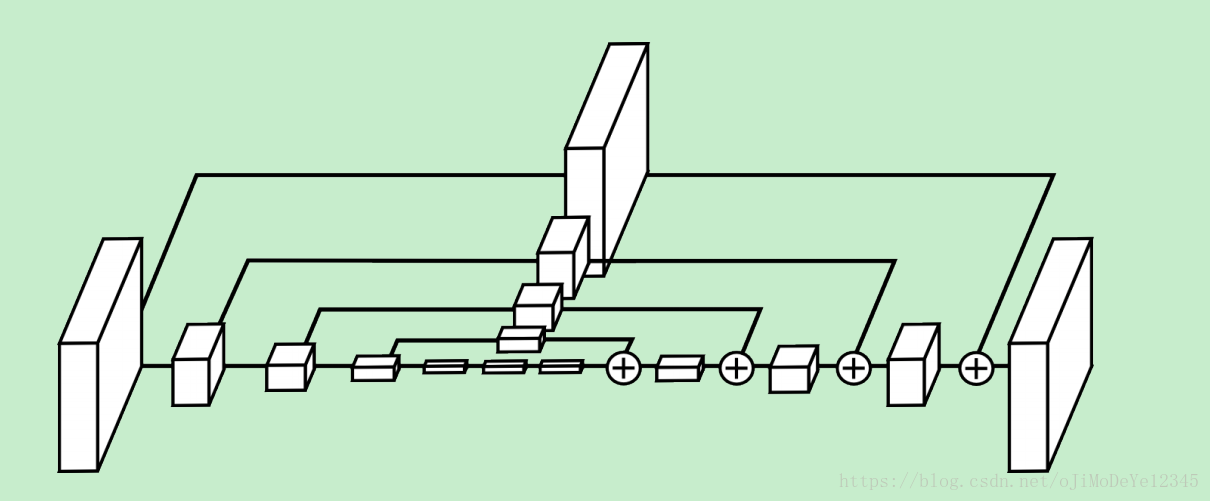


图 - 单个hourglass模块示例. 图中个方框分别对应一个residual模块. 整个hourglass中，特征数是一致的.

hourglass设置：

首先Conv层和Max Pooling层用于将特征缩放到很小的分辨率；

每一个Max Pooling(降采样)处，网络进行分叉，并对原来pre-pooled分辨率的特征进行卷积；

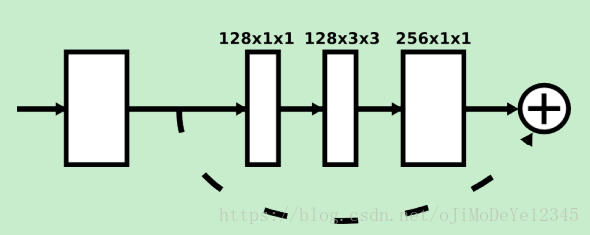
得到最低分辨率特征后，网络开始进行upsampling，并逐渐结合不同尺度的特征信息. 这里对较低分辨率采用的是最近邻上采样(nearest neighbor upsampling)方式，将两个不同的特征集进行逐元素相加.

整个hourglass是对称的，获取低分辨率特征过程中每有一个网络层，则在上采样的过程中相应低就会有一个对应网络层.

得到hourglass网络模块输出后，再采用两个连续的 1×1 Conv层进行处理，得到最终的网络输出.

Stacked Hourglass Networks输出heatmaps的集合，每一个heatmap表征了关节点在每个像素点存在的概率.

**Residual模块 Residual Module**



Hourglass网络中使用的Residual模块.

Residual模块提取较高层次的特征，同时保留原有层次的信息.

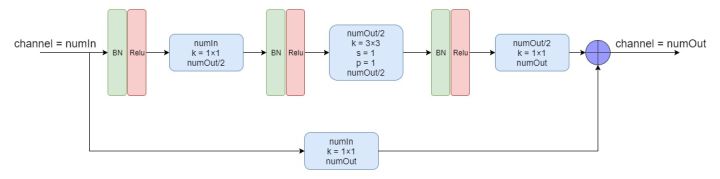
未使用大于3×3的Filters

限制每层的参数总数，以减少总的内存消耗

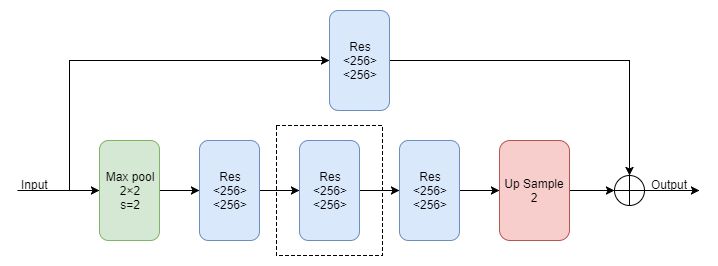
输入256×256尺寸的图片需要较多的GPU显存，hourglass网络的最大分辨率，也就是最终的输出分辨率，为 64×64.

Stacked Hourglass Networks网络开始是步长为2，filters为7×7的Conv层，其后是residual模块和一串Max Pooling层，将分辨率由256降到64. 其后是两个residual和hourglass模块. 整个hourglass中的所有residual模块输出特征数都是256.

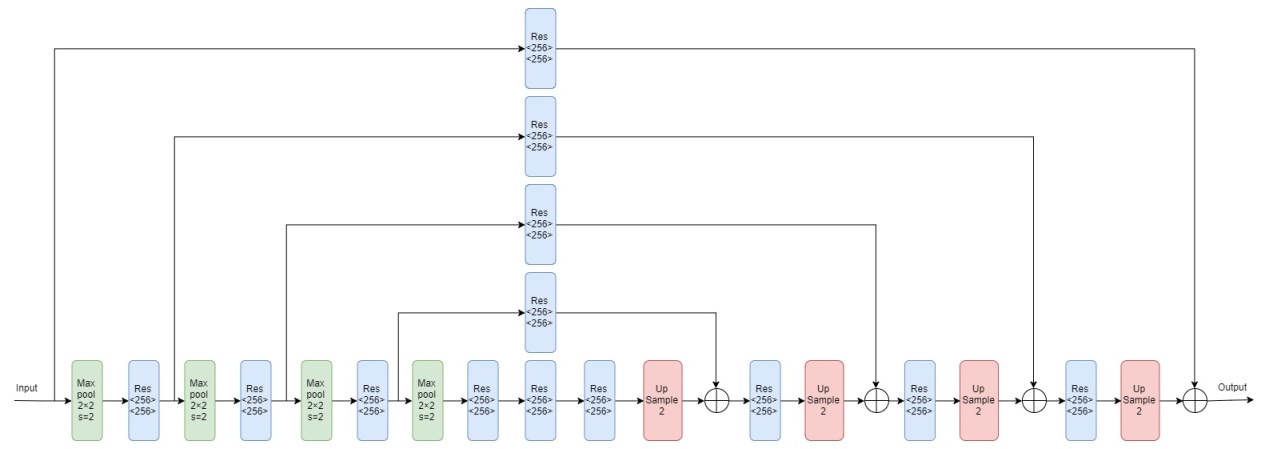
上图的残差块是论文中的原图，描述的不够详细，自己看了下源代码之后，画出了如下图所示的Residual Module：



Hourglass Module由上面的Residual Module组成，由于它是一个递归的结构，所以可以定义一个阶数来表示递归的层数，首先来看一下一阶的Hourglass Module：



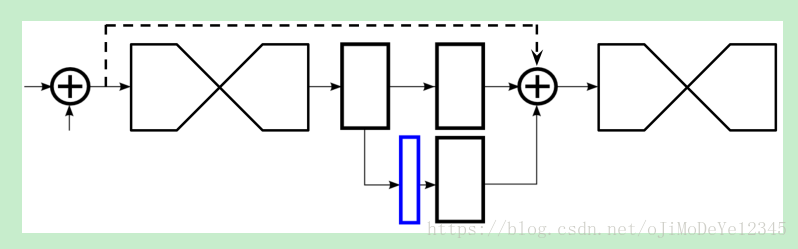
上图中的Max pool代表下采样，Res代表上面介绍的Residual Module，Up Sample代表上采样。多阶的Hourglass Module就是将上图虚线框中的块递归地替换为一阶Hourglass Module，由于作者在实验中使用的是4阶的Hourglass Moudle，所以我们画出了4阶的Hourglass Module的示意图：



**整体结构**

网络输入的图片分辨率为256×256，在hourglass模块中的最大分辨率为64×64，整个网络最开始要经过一个7×7的步长为2的卷积层，之后再经过一个残差块和Max pooling层使得分辨率从256降到64。

**中间监督 Intermediate Supervision**



Hourglass网络输出heatmaps集合(蓝色方框部分)，与真值进行误差计算. 其中利用1×1的Conv层对heatmaps进行处理以保证中间特征的通道数一致. 每一个Hourglass网络都添加Loss层.

在整个网络中，作者共使用了8个hourglass模块，需要注意的是，这些hourglass模块的权重不是共享的，并且所有的模块都基于相同的ground truth添加了损失函数。下面介绍训练过程的细节。

关于中间监督loss的计算，论文中是这么说的：

Predictions are generated after passing through each hourglass where the network has had an opportunity to process features at both local and global contexts. Subsequent hourglass modules allow these high level features to be processed again to further evaluate and reassess higher order spatial relationships.

所以，每个Hourglass Module的loss是单独计算的，这样使得后面的Hourglass Module能够更好地再评估。

**训练过程细节**

作者在FLIC和MPII Human Pose数据集上进行了训练与评估。这篇论文只能用于单人姿态检测，但是在一张图片中经常有多个人，解决办法就是只对图片正中心的人物进行训练。将目标人物裁剪到正中心后再将输入图片resize到256×256。为了进行数据增量，作者将图片进行了旋转（+/-30度）、scaling（.75-1.25）。

网络使用RMSprop进行优化，学习率为2.5e-4. 测试的时候使用原图及其翻转的版本进行预测，结果取平均值。网络对于关节点的预测是heatmap的最大激活值。损失函数使用均方误差（Mean Squared Error,MSE）来比较预测的heatmap与ground truth的heatmap（在节点中心周围使用2D高斯分布，标准差为1）

为了提高高精度阈值的性能，在转换回图像的原始坐标空间之前，预测在其下一个最高邻居的方向上偏移四分之一像素。

**训练 Training**

MPII Human Pose Dataset 数据处理：

- 采用MPII提供的scale和center标注信息，以目标人为中心，裁剪图片

- Resized to 256x256

- Rotation (+/- 30 degrees)

- Scaling (.75-1.25)

未进行平移处理，因为图像中目标人的位置是很重要的信息.

Training：

- Torch7

- rmsprop with a learning rate of 2.5e-4

- drop the learning rate once by a factor of 5 after validation accuracy plateaus

- 3 days on a 12GB NVIDIA TitanX GPU

- Batch normalization

- Mean Squared Error (MSE) loss 计算估计的heatmap和参考heatmap的误差

- single forward pass of the network takes 75 ms

Testing：

- 原始图片和其翻转图片，输入到网络，取输出heatmaps的平均值

- 最终输出的各heatmap的最大值位置作为关节点位置

Results



关于hourglass模块使用的个数，作者也进行了对比实验，分别采用2、4、8个堆叠的hourglass模块进行对比实验，结果如下所示：

