Abstract

Pose Machines provide a sequential prediction framework for learning rich implicit spatial models,this work show a systematic design for how convolutional networks ca be incorporated into the pose machine framework.
The contribution of this paper is to implicitly model long-range dependencies between variables in structured prediction tasks。achieved this by designing a sequential architecture composed of convolutional networks.
for vanishing gradients, wo providing a natural learning objective function that enforce intermediate supervision.

Introduction

CPM inherit the pose machine (the implicit learning of long-rane dependencies between image and multi-part cues)
At each stage in a CPM, image features and the belief maps produced by the previous stage are used as input.
CPM隐式的从局部关系中学习图像的空间模型
In order to capture loong-range interactions between parts, we achieve a large receptive field on both image and the belief maps. CPMs naturally suggest a systematic framework that replenishs gradients and guides the network to produce increasingly accurate belief map by enforcing intermediate supervision periodically through the network. Main Contribution:

1. sequential CNN architectures.
2. learning ability for both image features and image-dependent spatial models for structured prediction by a systematic approach.

Method

Pose Machines

A pose machine consists of a sequence of multi-class predictors $g_t(\cdot)$,在每个阶段$t \in \{1...T\}$,g_t 在当前位置z的图片x上提取的特征以及从之前或相邻分类器中的上下文信息的基础上预测每个部分的置信度，例如步骤1产生下列置信值：

$$g_1(X_z) \rightarrow \{b_1^p(Y_p = z)\}_{p\in\{0...P\}} \quad (1)$$

在接下来步骤中，每个预测的belief来源于两部分（之前提到):

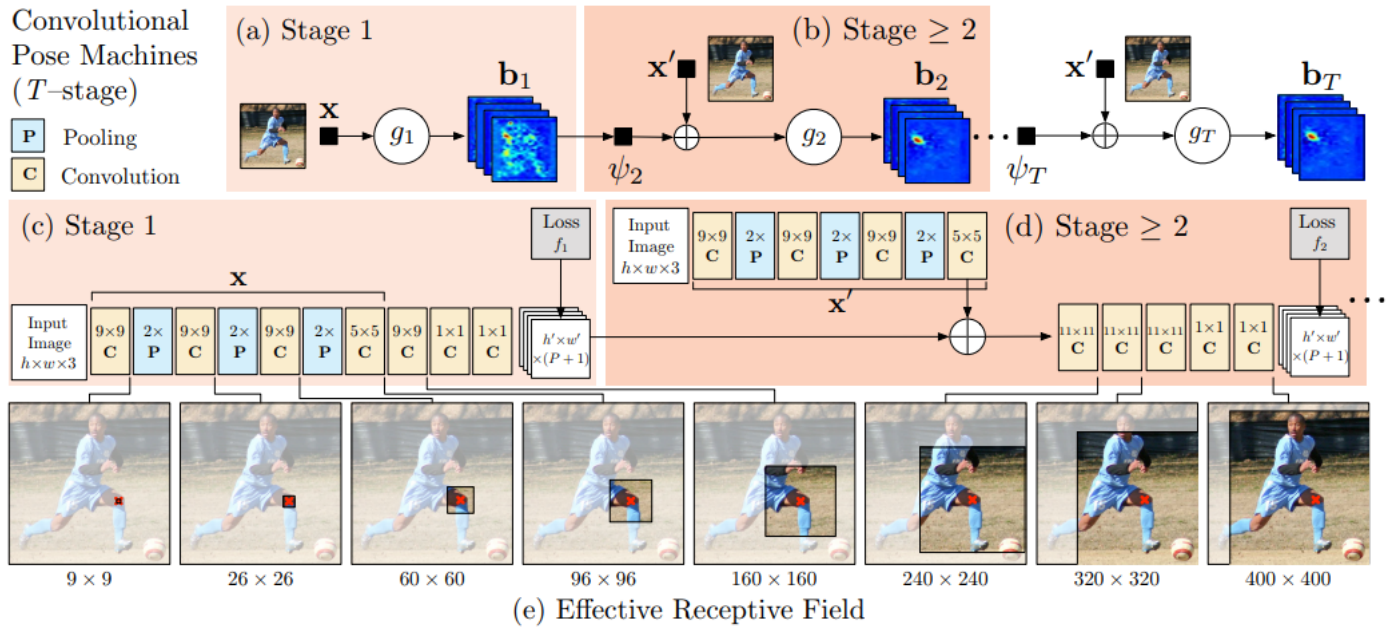
$$g_t(X_z', \phi_t(z, b_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p\in\{0...P+1\}} \quad (3)$$

The pose machine proposed used boosted random forests for prediction.

Convolutional Pose Machines

Keypoint Localization Using Local Image Evidence

从最底层的局部开始获取信息(总共有5层，最后两层为1x1），每个网络可接受的范围为160x160,所以一个网络不能获取全部信息，需要多个输出的组合。



用CNN代替了Pose Machine中的阶层

Sequential Prediction with learned Spatial Context Features

因为外形和配置上较大的方差， 在运动中的关节识别准确率是很低的，所以依据相邻结构和之前的上下文信息能提高准确率。
一般获得较大的感受范围能或者较大的准确率(但是当范围提高到一定程度会饱和)， 为了达到这个效果，我们增加网络的层数。

Learning in Convolutional Pose Machines

为了减少梯度消失问题，每次都会重复训练局部位置上重新训练生成置信映射。
然后在每一个阶段之后设计一个损失函数减小没部分预测的结果和理想置信映射的l2距离.

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in Z} \|b_t^p(z) - b_*^p(z)\|_2^2 \quad (4)$$

总体的目标函数就是每个阶段的损失相加，此外，在stage大于等于2的阶段的CNN的参数共享。

附：Pose Machines: Articulated Pose Estimation via Inference Machines

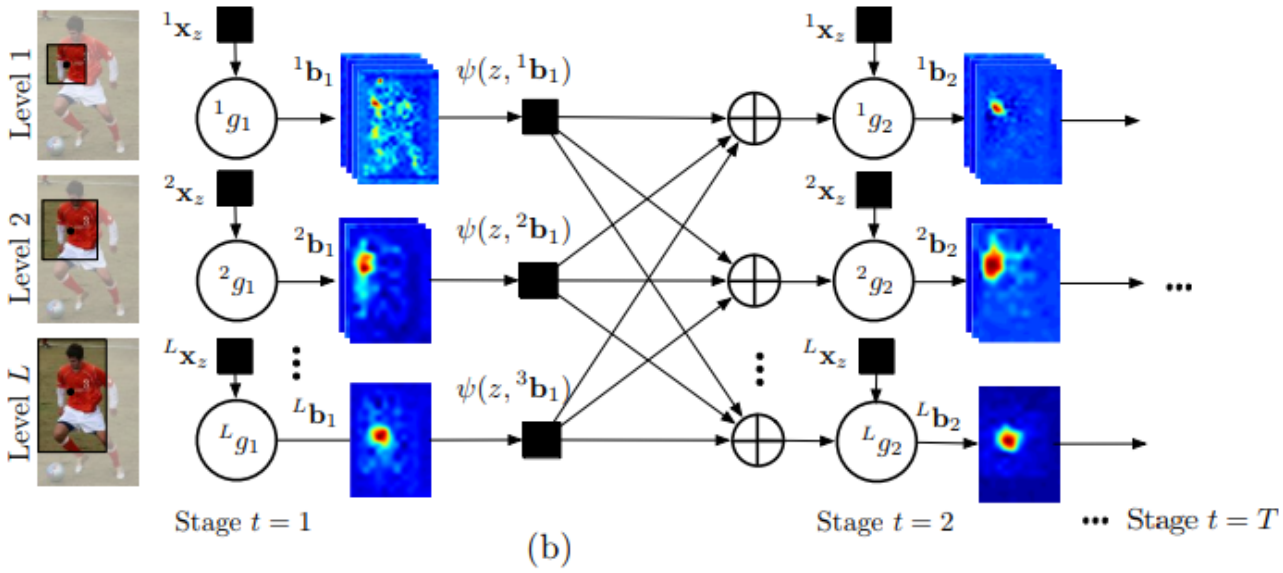
1 通过博客的简要理解

模拟图模型的推断的机制，本质是多个分层的多类分类器的级联。对于关节的位置的预测，宣称能解决遮挡问题。
传统的方法往往是树或者星状图，不能很好的捕捉身体部位之间的关系，往往会为了问题的可解性，放弃问题的复杂度，而本文用了一种图模型的inference方法，称为推断机，可以有效的对遮挡的关节进行建模。

1.1 结构：

1. 对每一层的多分类预测
2. 一个姿态估计有两个阶段
   整体的网络结构为多个层次，最低层次表示某个部件，越往上表达越抽象（比如最顶层表示一个人），然后

在图像中每层都用（a）中的多分类预测，得到置信度map,然后将map作为下一个阶段的输入。从下图中，可以看到，层次是横向的，stage是纵向的(一般3个stage效果即满足)。



然后作者提到了两种处置信度map的方法，处理完成后才能作为下一阶段的输入，后面详细说其方法。

2 论文针对拾遗

通过上图可以看到，第一阶段，每一层的输入为不同层次的x（子图片），然后第二阶段及以后的阶段，除了包含图片本身的输入外，还有上一阶段各层的输出的concat。

$$b_t(Y_p = z) = g_t^p(x_z; \bigoplus_{i=1}^{p} \phi(z, b_{t-1}^i))$$

其中： $b_{t-1}^p = \{b_{t-1}(Y_p = z)\}_{z \in Z}$表示前一个分类器在每个location z（图片的位置）上对p'th part（人体的每个部位）的置信度预测。$\bigoplus$表示vector concatenation。

2.1 Incorporating a Hierarchy

design a hierarchical inference machine that similarly encodes these interactions among parts at different scales in the image.
如在第1阶段第l层输入为z的得分为：

$$^l g_1(\mathbf{x}_z^l) \rightarrow \left\{ ^l b_1^p(Y_p = z) \right\}_{p \in 0 \ldots P_l},$$

Note that he predictions for a part use features computed on ouputs of al parts and in all levels of the hierarchy($\{b_{t-1}^l\}$)

2.2 Context Features

to capture the spatial correlations between the confidences of each part with respect to its neighbors.

1. **Context Patch Features**

$$\psi_1(z, ^l\mathbf{b}_{t-1}) = \bigoplus_{p \in 0 \ldots P_l} \mathbf{c}_1(z, ^l\mathbf{b}_{t-1}^p).$$

即该图片部分对每个关节的预测值的cancat

2. **Context Offset Features**
   为了获取更长期的关系，**首先**对每一层的part排序后取前K个，**然后**以每个位置对应第k个part的置信映射作为极坐标计算子集向量，自己向量为：

$$\mathbf{c}_2(z, {}^l\mathbf{b}^p_{t-1}) = \left[{}^lo^p_1; \ldots; {}^lo^p_K\right].$$

   ， 然乎如同1：

$$\psi_2(z, {}^l\mathbf{b}_{t-1}) = \bigoplus_{p \in 1 \ldots P_l} \mathbf{c}_2(z, {}^l\mathbf{b}^p_{t-1}).$$

Training

Algorithm 1 `train_pose_machine`

Initialize: $\left\{{}^l\mathbf{b}_0 = \emptyset\right\}_{l \in 1,\ldots,L}$
for $t = 1 \ldots T$ **do**
  for $i = 1 \ldots N$ **do**
    Create $\{{}^l\mathbf{b}_{t-1}\}^L_{l=1}$ for each image i using predictor ${}^lg_{t-1}$ using Eqn. 5.
    Append features extracted from each training image i, and from corresponding $\{{}^l\mathbf{b}_{t-1}\}^L_{l=1}$ (Eqns. 6 & 8), to training dataset $\mathcal{D}_t$, for each image i.
  end for
  Train ${}^lg_t$ using $\mathcal{D}_t$.
end for
Return: Learned predictors $\{{}^lg_t\}$.

针对于数据集，是一个阶段一个阶段训练的。

Stacking

由于每个predictor对训练数据预测后再传给下一层作为下一层的训练数据，很容易导致过拟合，为此，我们使用CV,每个predictor使用其中一份数据训练，然后再预测没有训练过的数据，将结果作为下一阶段的训练数据。

LSTM Pose Machine

Contributions of out work:

- First, build a novel recurrent architecture with LSTM to capture temporal geometric cosisitency and dependency among video frames for pose estimation
- Second, decouples the relationship among network stages and results in much faster inference speed for videos
- Third, probed into the memory cells and visualized how they woud help to imptove the joint predictoins on videos.