论文题目

Spatial Transformer Networks[PDF]

解决问题

CNN网络结构中，pooling layer能一定程度上具有spatial invariant，但是这种人工设定的变换规则使CNN网络过分依赖先验知识。本文提出了spatial transform net，可以在网络中显示的自动学习空间变换信息，并且不需要对优化过程进行额外的训练监督或修改，取得了很好的效果。

创新

网络结构

空间变换网络的结构如下图：

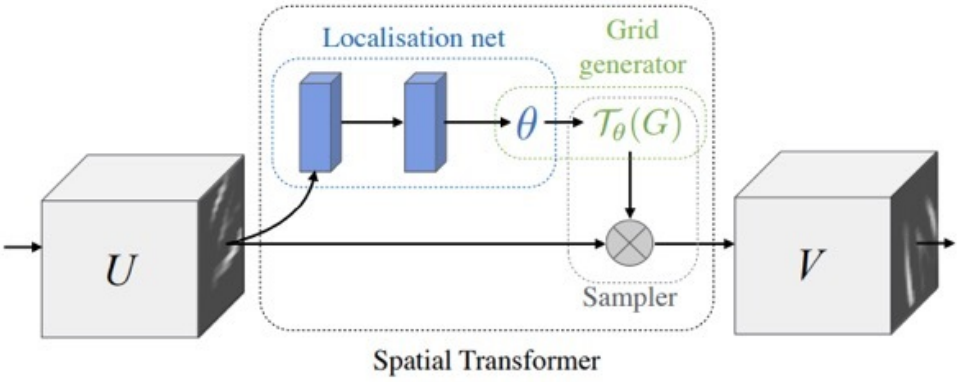

Figure 2: The architecture of a spatial transformer module. The input feature map U is passed to a localisation network which regresses the transformation parameters θ. The regular spatial grid G over V is transformed to the sampling grid $\mathcal{T}_\theta(G)$, which is applied to U as described in Sect. 3.3, producing the warped output feature map V. The combination of the localisation network and sampling mechanism defines a spatial transformer.

ST的结构如上图所示,每一个ST模块由Localisation net, Grid generator和Sample组成, Localisation net决定输入所需变换的参数θ,Grid generator通过θ和定义的变换方式寻找输出与输入特征的映射T(θ),Sample结合位置映射和变换参数对输入特征进行选择并结合双线性插值进行输出

Loscalisation net

自己定义的一个网络，输入为Input feature map.输出为spatial transform的参数 θ。

Grid Generator

输入为V中的坐标点以及变换参数 θ ，计算出U中的坐标点

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathbf{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

Sampler

输入为Input feature map 以及Grid。从Input feature map中进行采样得到output feature map

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_j^s - n; \Phi_y) \forall i \in [1 \cdots H'W'] \forall c \in [1 \cdots C]$$

总结

STN加强了网络对图像空间变换的能力，并且能实现endtoend的训练，具有很好的效果