1 Introduction

此文在CNN的基础上对hidden layer进行supervision。实际上是对CNN的隐层采用squared hinge loss （L2 Loss）使得CNN的隐层具有discriminative。 本文的方法有一个先验的经验，那就是如果能够让特征discriminative，则分类效果一定就好。（不过这一点我觉得不能完全认同）

Introduction部分指出了现在CNN存在的一些问题，觉得有必要列出来：

- 隐层的特征的 transparency 和 criminativeness 特性减少。
- 由 exploding and vanishing gradients 导致的训练难度
- 缺乏对DL算法的数学理论解释
- 对大量训练数据的依赖
- 调参复杂

2 算法说明

2.1 符号规定

input training data: $S = \{(X_i, y_i), i = 1, \ldots, N\}$ 其中$X_i \in \mathcal{R}^n$表示原始数据，$y_i \in \{1, \ldots, K\}$表示对应的label。
网络层数：M
学习到的filters/weights：$W^{(m)}, m = 1, \ldots M$
$m-1$层产生的feature map：$Z^{(m-1)}$
convolved/filtered responses：$Q^{(m)}$
Pooling function：$f()$
输出层的SVM weights：$\mathrm{w}^{(out)}$

2.2 公式说明

对每一层$m = 1, \ldots M$有：

$$Z^{(m)} = f(Q^{(m)}),\ and\ Z^{(0)} \equiv X, \tag{1}$$

$$Q^{(m)} = W^{(m)} * Z^{(m-1)}, \tag{2}$$

亦即从网络的结构上来看文中的结构与传统的CNN网络并无差别。 合并所有的层的weights有：

$$W = (W^{(1)}, \ldots, W^{(M)}),$$

对于each hidden layer有：

$$\mathrm{w} = (\mathrm{w}^{(1)}, \ldots, \mathrm{w}^{(M-1)}),$$

%%这个w是怎么来的我目前还没完全确定，文章写的不是十分清楚，智能说一下自己的理解，这个w很可能是将$Z^{(m-1)}$作为输入，输出为各不相同的label到L2SVM中，也就是说利用svm将一层的Z完全区分开来，这样学到一个对应的w。因此在训练之前这个w是未知的，他与每一次filters update时一起update，其update方法与CNN+SVM的策略相同，在《Deep Learning using Linear Support Vector Machines》中有详细介绍（本文后面附有链接）。

本文的核心motivation就是在hidden layer加入约束

总体上的目标函数是：

$$\|\mathrm{w}^{(out)}\|^2 + \mathcal{L}(W, \mathrm{w}^{(out)}) + \sum_{m=1}^{M-1} \alpha_m \left[\|\mathrm{w}^{(out)}\|^2 + \ell(W, \mathrm{w}^{(m)}) - \gamma\right]_+, \tag{3}$$

这其中：

$$\mathcal{L}(W, \mathrm{w}^{(out)}) = \sum_{y_k \neq y} \left[1 - \langle \mathrm{w}^{(out)}, \phi(Z^{(M)}, y) - \phi(Z^{(M)}, y_k)\rangle\right]_+^2 \tag{4}$$

$$\ell(W, \mathrm{w}^{(m)}) = \sum_{y_k \neq y} \left[1 - \langle \mathrm{w}^{(m)}, \phi(Z^{(m)}, y) - \phi(Z^{(m)}, y_k)\rangle\right]_+^2 \tag{5}$$

对于这个目标函数有如下解释：

- (3)左边两项$\|\mathrm{w}^{(out)}\|^2$和$\mathcal{L}(W, \mathrm{w}^{(out)})$与传统CNN相似，不同的是在这里不是用传统CNN的softmax做分类器而是用SVM。
- (3)右边一项括号中的$\|\mathrm{w}^{(out)}\|^2$和$\ell(W, \mathrm{w}^{(m)})$是本文中提出的方法
- 从整体上来看这个函数，一方面它照顾到了经典CNN中误差反向传导所必须的自下至上的传递结构，另一方面还使得filters之间满足response discriminative的特性。

未完待续。。。。

3 题外话

这篇文章值得注意的地方不仅仅是方法，更重要的是思路，虽然很多人认为中层监督相比普通CNN相对退化，也就是说现在主流研究还应该是用理论解

释how does DL work？而不是有退回了原来的特征+聚类+分类的思路。但是我认为中层监督实际上还是存在实际应用上的价值，尤其是语意上的中层监督，可以将人工的引导引入CNN的中层，这样好处有3：（1）大量label大有用武之地（2）帮助解释中层black box（3）直觉上能够提高收敛速度。