1.2.1 [Inception-v3] Notes by YK

Rethinking the Inception Architecture for Computer Vision

a. Main Contributions

Scale up networks in ways that aim at utilizing the added computation as efficiently as possible by suitably **factorized convolutions** and **aggressive regularization**.

b. Key Points

(i)Design Principles
-- Avoid representational bottlenecks(e.g. 32x32x128-->16x16x256)
-- Higher dimensional representations are easier to process locally within a network
-- Spatial aggregation can be done over lower dimensional embeddings without much or any loss in representational power
-- Balance the widen and depth of the network

(ii)Factorizing Convolutions with Large Filter
-- Factorization into smaller convolutions: 5x5Conv+relu——>3x3Conv+relu + 3x3Conv+relu
-- Spatial Factorization into Asymmtric Convolutions: 7x7Conv-->1x7Conv+7x1Conv

(iii)Efficient Grid Size Reduction
from $d\mathrm{x}d\mathrm{x}k$ to $\frac{d}{2}\mathrm{x}\frac{d}{2}\mathrm{x}2k$: $d\mathrm{x}d\mathrm{x}k$ --> $d\mathrm{x}d\mathrm{x}2k$ --> $\frac{d}{2}\mathrm{x}\frac{d}{2}\mathrm{x}2k$ (large computation instead of bottlenecks)
Better Choice: 35x35x320 --> (conv)17x17x320-[concat]-(pool)17x17x320 --> 17x17x640

(iv)Model Regularization via Label Smoothing [todo]

c. Experiments and Results

In one word, Inception-v3 is smaller and more accurate than GoogLeNet, VGG and Inception-v2.

d. English Writing

(i) Nice phrases
contributing factors
architectural improvements
computational tricks
prove prohibitive or unreasonable in practical scenarios
mitigate the impact of ...
grave deviations
use them judiciously
exploit translation invariance
(ii)Nice sentences
-- Since 2014 very deep convolutional networks started to become **mainstream**, **yielding substantial gains in various benchmarks**.

-- Although increased model size and computational cost tend to **translate** to **immediate quality gains** for most tasks, **computational efficiency** and **low paramter count** are still **enabling factors** for various use cases such as **mobile vision** and **big-data scenarios**.

-- This makes it much harder to adapt it to new use-cases while maintaining its efficiency

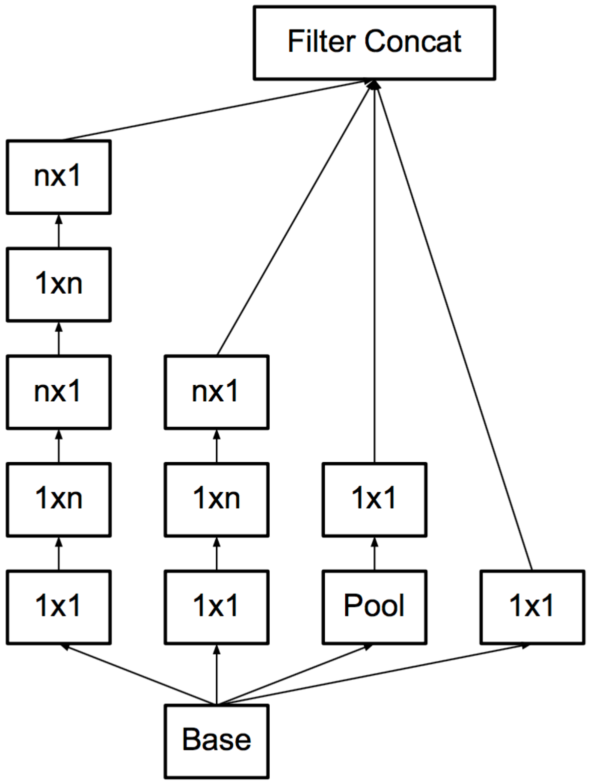

Figure 6. Inception modules after the factorization of the $n \times n$ convolutions. In our proposed architecture, we chose $n = 7$ for the 17×17 grid. (The filter sizes are picked using principle 3)

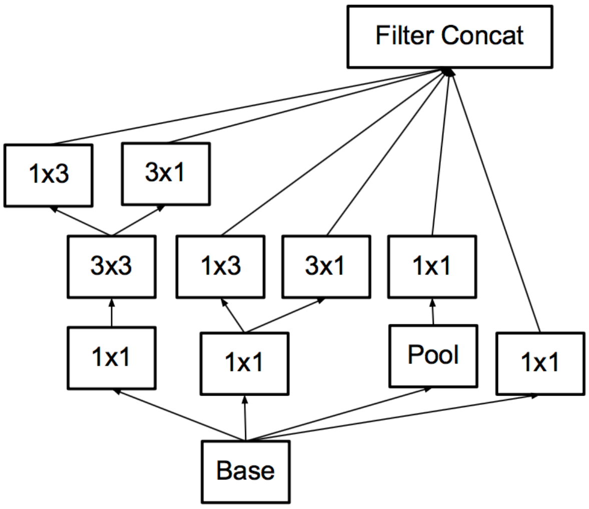

Figure 7. Inception modules with expanded the filter bank outputs. This architecture is used on the coarsest (8×8) grids to promote high dimensional representations, as suggested by principle 2 of Section 2. We are using this solution only on the coarsest grid, since that is the place where producing high dimensional sparse representation is the most critical as the ratio of local processing (by 1×1 convolutions) is increased compared to the spatial aggregation.

...