Deep LDA-Pruned Nets for Efficient Facial Gender Classification

Qing Tian, Tal Arbel and James J. Clark

Center for Intelligent Machines and ECE Department
McGill University

July 21st, 2017

Outline

Introduction

Deep nets, especially CNNs, are very effective but

- not well suited for PCs/mobile devices w/o a powerful GPU. However, high efficiency is desired in various applications such as HCI, image retrieval, and online video stream analysis.
- GPUs' memory constraints limit the number of heavy nets that can be loaded at the same time.

Introduction

A common practice in deep learning: adopting a general network and fine-tune it for a specified task (usually on a smaller dataset).

However, there is no theory to adjustify the inherited architecture. Do we really need all the structures from pre-trained heavy nets?

Introduction

Our aim is to greatly prune deep networks in a supervised way while maintaining (or even improving on) their classification accuracy.

- Most off-diagonal values in within-gender scatter matrix of last conv layer firing data are (near) zero $\xrightarrow[\text{LDA}]{\text{Fisher's}}$ discarding less useful neurons directly with no information loss.
- Light alternatives to FC layers to further reduce complexity.

Related Work

Gender Recognition

- Global/Local handcrafted features + classic classifiers
  e.g. LBP + SVM/Bayesian (combined with AdaBoost)
- Neural networks: from shallow to deep
  Early works: Golomb (1990), Poggio (1992), Gutta (1999)
  Recent: Levi and Hassner (2015), Mansanet (2016),
  Liu (2015)

Related Work

Deep Neural Networks Pruning

- Traditional approaches targeting at shallow nets:
  e.g. Optimal Brain Damage (LeCun 1990)
- Weight Magnitude based pruning: e.g. Han 2016
- Inspiration for neuron level pruning from neuroscience:
  1. neurons typically receive inputs from a task dependent small set of others (Valiant 2006)
  2. functional columns exist in the cortex (Mountcastle 1957): minicolumns have accompanying functionalities, which becomes clear when seen on the higher macrocolumn level.

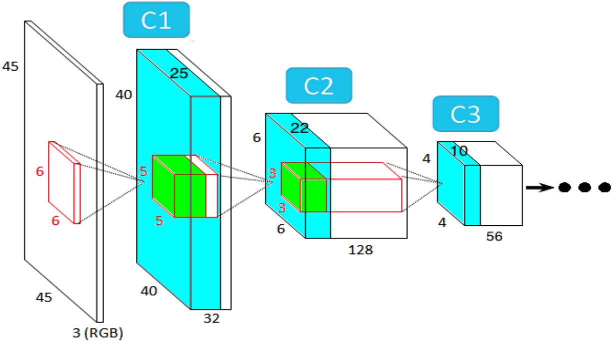
Main Contribution

A Fisher LDA based Deep Net Pruning Approach

- It prunes on the filter level, thus directly leads to space and time savings (key difference from Han 2016).
- It treats pruning as a supervised dimensionality reduction problem.

Fisher LDA based Filter Level Pruning

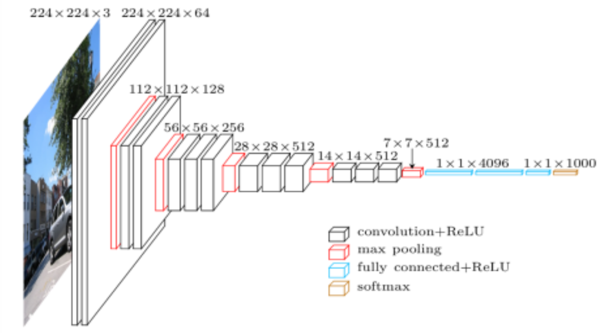
- Pruning CNN on the Filter Level



Demonstration of pruning on filter level (cyan indicates remaining data, green represents the surviving part of a remaining next layer filter).

Fisher LDA based Filter Level Pruning

- Base Network Structure:



VGG-16 Model (Simonyan and Zisserman 2015)

Fisher LDA based Filter Level Pruning

- Dimension Reduction in the Last Conv Layer

  1. Conv5_3 neurons are shown empirically to fire less correlatedly within each class than other conv layers.
  2. Unlike FC layers, Conv5_3 preserves location information, and is not restricted by input dimension.
  3. Babenko (2015) and Zhong (2016) have demonstrated higher accuracies of last conv layer than FC layers in image retrieval and facial traits analysis tasks.

Fisher LDA based Filter Level Pruning

Instead of PCA, we draw inspiration from Fishers Linear Discriminant Analysis (Fisher 1936) and adopt the intra-class correlation (ICC) to better measure the information utility:

$$ICC = \frac{s^2(b)}{s^2(b) + s^2(w)}$$

$s^2(w)$: within-class variance, $s^2(b)$: between-class variance. When reducing dimension, we maximize ICC or $s^2(b)/s^2(w)$.

Fisher LDA based Filter Level Pruning

The direct multivariate generalization of it is:

$$W_{opt} = \arg\max_{W} \frac{| W^T S_b W |}{| W^T S_w W |}$$

where

$$S_w = \sum_{i=0:1} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T$$

$$S_b = \sum_{i=0:1} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

W: orthogonal transformation matrix. By analyzing S_w for LFW and CelebA, we find most off-diagonal values in S_w are (near) zero.

McGill University

Fisher LDA based Filter Level Pruning

We find firings of the last conv layer neurons are highly uncorrelated within each gender mainly because:

- higher layers capture various high-level abstractions.
- high dimensional coincidences can hardly occur by chance.

Since W columns are generalized eigenvectors of S_w, they turn out to be standard basis vectors (with eigenvalues on S_w diagonal).

Fisher LDA based Filter Level Pruning

To maximize the ICC, we simply need to select neuron dimensions of low within-gender variance and high between-gender variance.

Fisher LDA based Filter Level Pruning

Demo of Conv5_3 Neurons' Activation (CelebA trained, N214 highlighted)

Fisher LDA based Filter Level Pruning

Demo of Conv5_3 Neurons' Activation (CelebA trained, N298 highlighted)

Fisher LDA based Filter Level Pruning

We use deconv (Zeiler 2011, 2013) to calculate cross-layer dependency, which consists of series of unpooling, rectification, and reversed convolution.

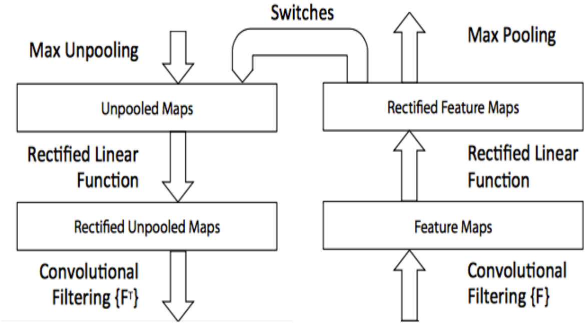


Figure: Unit Deconv Operation

Fisher LDA based Filter Level Pruning

- Light Classifiers on Top of CNN Features
  1. SVM (with linear and RBF kernels)
  2. Bayesian quadratic discriminant analysis

Experiments and Results

- Datasets
  1. LFWA - richly labeled version of the LFW database, covering a large range of pose and background variations.
  2. CelebA - the CelebFaces Attributes Dataset containing 202,599 images of 10,177 identities.

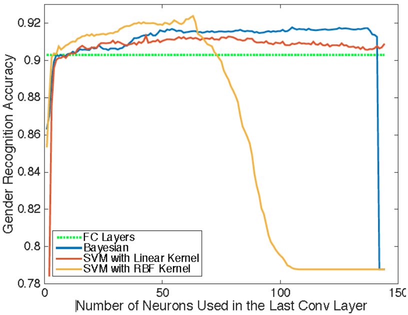
Experiments and Results

- Recognition Accuracy

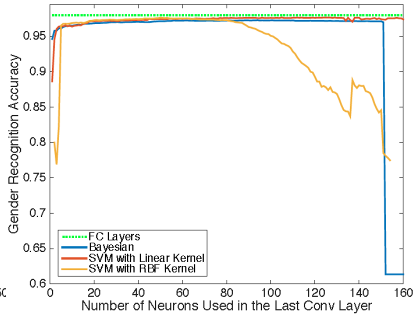
Method	LFW	CelebA
Original Net with FC	90.3% (512)	98.0% (512)
LDA-CNN+Bayesian	91.8% (105)	97.3% (94)
LDA-CNN+SVML	91.3% (43)	97.7% (105)
LDA-CNN+SVMR	92.4% (63)	97.5% (52)

Table: Highest recognition accuracy comparison of different approaches. SVML/SVMR: SVM with linear/RBF kernel. The accuracies reported here are the highest when a smaller number (specified in the parentheses) of neurons are utilized in the last conv layer.

Experiments and Results



(a) Accuracy Comparison using Different Classifiers on LFW

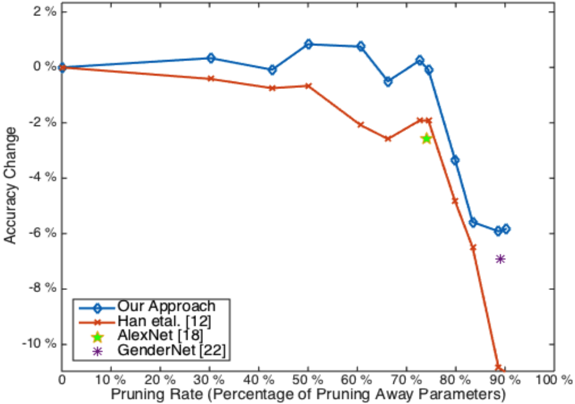


(b) Accuracy Comparison using Different Classifiers on CelebA

McGill University

Experiments and Results

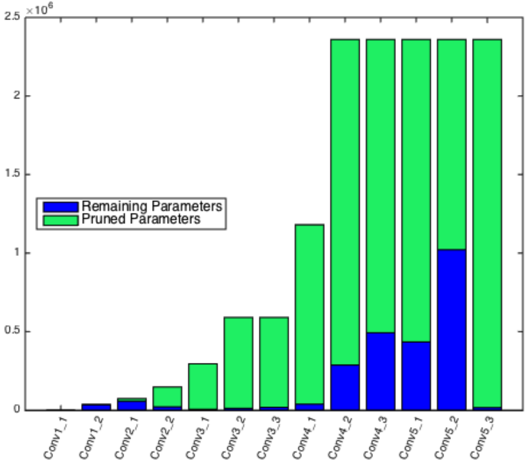
Accuracy Change vs. Parameter Pruning Rate



Accuracy change vs conv layers pruning rate (4 Conv5_3 neurons, LFWA)

Experiments and Results

- Structure Complexity



Layerwise structure complexity reduction (4 Conv5_3 neurons)

Experiments and Results

- Storage
  Compared to the original deep net (**over 500 MB**), our pruned models are very light:
  1. conv filters take up less than **10 MB** in space
  2. the storage overhead can be ignored for Bayesian QDA
  3. the extra storage needed is only about **30KB** for both SVMs.

Experiments and Results

- Recognition Speed

Method \ Layer	Conv1_1	Conv1_2	Conv2_1	Conv2_2	Conv3_1	Conv3_2	Conv3_3	Conv4_1
Original CNN+FC Layers	70.96	405.39	183.60	362.15	171.64	341.23	341.33	166.94
LDA-CNN+Bayesian/SVM	18.02	98.27	39.68	31.96	3.59	6.43	9.92	3.79
Speedup Ratio	3.93	4.13	4.63	11.33	47.83	53.06	34.41	44.08

Method \ Layer	Conv4_2	Conv4_3	Conv5_1	Conv5_2	Conv5_3	FC Layers			**Total**
						BC	SVML	SVMR	
Original CNN+FC Layers	333.75	333.98	85.69	85.70	85.63	283.20			**3306.50**
LDA-CNN+Bayesian/SVM	18.11	28.07	6.79	11.92	0.84	0.04	0.01	0.05	**286.86**
Speedup Ratio	18.43	11.90	12.63	7.19	101.68	7E3	3E4	6E3	**11.53**

Table: Per image recognition time comparison of different approaches in all layers (in milliseconds). BC: the Bayesian classifier, SVML/SVMR: SVM with linear/RBF kernel. The tests are run on the CPU.

Conclusion

In our work, a deep but pruned CNN is developed that, combined with alternative classifiers, can boost efficiency while maintaining accuracy in gender recognition. Advantages over unstructured weights based CNN pruning:

- our framework picks connections that eventually contribute to the discriminating power (connection importance is not necessarily related to pre-trained weights' magnitudes).
- instead of using masks to disregard weights, our pruning approach is able to achieve real space and time savings (desirable for embedded systems).

Future Plan

- Test this approach for other tasks and net structures.
- Make our approach iterative and location aware.
- Investigate the macrocolumn/minicolumn hypothesis in neuroscience for more inspiration for neuron level pruning.
- Extend our approach to prune the whole network including fully connected layers.