PCNS: PCA Network Shrinking

Research draft

P. Rolet

Sep 29, 2022

Abstract

This is a work in progress to explore an idea of algorithm that would significantly reduce both the size and the training time of deep learning models. Other shrinking methods such as distillation (e.g. Gou et. al., 2021) and weight pruning (e.g. Renda et al., 2020) require a full training to be done beforehand and do not offer proven adjustable performance guarantees. If this idea, called PCA network shrinking or *PCNS*, is successful, the expected benefits are:
a/ to significantly reduce training time;
b/ to reach near-minimal network size while maintaining good performance;
c/ to be simpler to operate than other shrinking methods;
d/ to provide performance bounds.

Outline of the PCNS algorithm on a fully-connected layer (oversimplified)
Given the output matrix of the layer on a large batch, i.e. its feature map (with features as columns), consider its rows (each row corresponds to the output of one neuron over the batch) and approximate them by performing PCA and retaining the largest eigenvectors; change the layer's weights so that applying the forward pass on the batch generates approximated eigenvectors instead of the transposed outputs (this is done via ordinary least squares and this is what actually reduces the size of the network); then change the next layer's weights so that for each former output the corresponding linear combination of the new outputs is computed during future forward passes.

1 Introduction

The cost of training deep models is often high, and many look for ways to reduce it. Reducing the model size is also a common objective–for faster and

cheaper inference, and ability to fit the model in limited-memory devices. However, the goals are usually considered separately. Methods to reduce training time rarely focus on reducing the model size during training[1]. Conversely, methods to reduce the model size usually rely on fully training the model first–reducing training time is not their stated goal.

This, combined with the growing availability of computing power tailored to deep learning, has led to model size reduction research being seen as application-specific, restricted to mobile or embedded computing. But if there was a way to shrink models during training, it may potentially make training much faster—e.g. being able to halve the size of the layers during training while maintaining approximately the same performance, would divide by 4 the remaining training time.

Additionally, it can be hoped that reducing the model size is not only quantitatively useful, but may also bear qualitative benefits: it allows even deeper architectures with less parameters, and intuitively, learning something using a smaller description can mean learning it "better"[2] (the use of "Minimum Description Length" approaches in machine learning seems to stem from a similar intuition).

This work explores a new algorithmic method of model size reduction called *PCA network shrinking* or *PCNS* that could be applied soundly during training, and as such can significantly reduce training time. As mentioned in the abstract, other deep learning model shrinking methods such as distillation (e.g. Gou et. al., 2021) and weight pruning (e.g. Renda et al., 2020) require a full training to be done beforehand and do not offer provable adjustable performance guarantees. If the PCNS algorithm is successful, the expected benefits are:

- to significantly reduce training time;
- to reach near-minimal network size while maintaining good performance;
- to be simpler to operate than other shrinking methods;
- to provide performance bounds.

[1] An exception would be using sparse-inducing penalties in the training loss, e.g. proportional to the L0-norm (TODO provide the ref of the L0 paper), but the approach does not seem conclusive as it has not been picked up by the research community despite being known for a long time. This is not to say that the approach doesn't deserve further study—however, it is not the focus of this work.

[2] TODO explain this argument is admittedly informal but anyways secondary, the main immediate goal being the model size reduction

The idea behind this algorithm stems from *redundancy* that can be observed inside network layers, explained in the subsection below.

1.1 Redundancy & intuition behind the algorithm

We can reduce the network during the training if we observe *redundancy* in a layer, i.e. groups of neurons performing similar computations. The simplest example would be 2 neurons having the exact same weights–in this case it would make sense to remove one. Another basic example would be a neuron that outputs 0 almost all the time: removing it would not change the network's output. A neuron outputting almost always a value close to 1, which may occur with saturating activation functions, can also be accounted for by the bias of the layer.

The PCNS algorithm is an attempt to generalize these observations and systematically remove redundancy during the training process, when it makes sense[3].

A direct approach would be to look into directly into weights matrices looking for neurons with similar weights, but this is not general enough since it does not incorporate the inputs distribution in the approximation. Even if the neuron weights are not the same, we consider them redundant if they compute similar outputs on most possible inputs. Therefore, we will spot neuron redundancies by comparing neuron outputs. An underlying assumption of the algorithm is that if two neurons generate almost the same ouput on a large input batch, they generate almost always the same output *in general*–and one can be removed. The algorithm will therefore focus on *transposed ouputs*. A transposed output is the output of a single neuron on an entire input batch (whereas an output usually refers to the output of all neurons on a single input).

The second idea in PCNS is to generalize "removing neurons with similar outputs". Looking at the matrix of transposed outputs, saying that there are 2 neurons with similar outputs means the rank of the matrix is one less. This leads to a generalization embodied by using PCA on the transposed outputs : approximating them by projecting them on a basis made of the largest eigenvectors. We then would like to use these eigenvectors in the network.

Since transposed ouptuts are approximated well by a linear combination of the eigenvectors, the third idea is to change the weight matrix of the layer so that instead of generating the transposed outputs, it generates those

[3] it does not always make sense; TODO when does it?

eigenvectors; and the fourth one is to use those linear combinations in the next layer to make it so that its inputs are approximated versions of the former inputs (approximated via aforementioned linear combinations).

1.2 Structure

This work is currently in progress. It is layed out as follows:

- section 2 describes the PCNS algorithm in more details;
- section 3 explains core assumptions behind the algorithm and what makes it a potentially good idea;
- section 4.1 provides detail on when and how to run the algorithm and choose its main parameters;
- section 7 explains next steps to complete this work;
- sections 5 (Example: how PCNS works on a toy network) and 6 (Experiments) are not yet written.

1.3 Notations

We consider a fully-connected neural network of L layers, such as the one described in Figure 1.

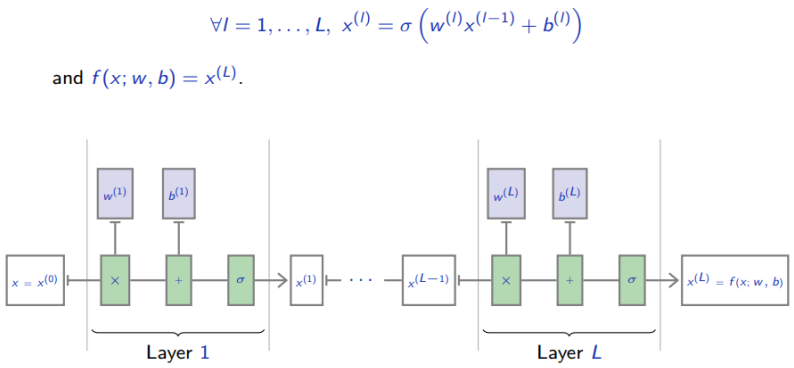

Figure 1: Multi-layer perceptron formalization (from F. Fleuret)

In this work, the following conventions are used:

- uppercase letters are used to denote matrices; uppercase X is used to denote a batch of input vectors; $X^{(0)}...X^{(L-1)}$ are input batches of size B for layers 1 to L, and $X^{(L)}$ is the output of layer L;
- subscript index is used to denote a column vector, therefore a matrix is a sequence of column vectors, e.g. $X^{(j)} = X_1^{(j)}...X_B^{(j)}$ (consequently, $X_i^{(j)} = (X_{i1}^{(j)}...X_{iN_j}^{(j)})$ denotes a single input vector of dimension N_j);
- vectors may also be designated with bold lowercacse letters (e.g. $\mathbf{y} = \mathbf{x}^{(l+1)} = X_1^{(l+1)}$), and scalar with non-bold lowercase letters (e.g. $\mathbf{y} = (y_1, \ldots, y_n)$);
- $W^{(i)}$ are the weights for layer i, noted with a capital letter since it is a matrix (as opposed to lowercase w in Fig. 1);
- σ is the activation function, the $\odot$ operator may be used to remind it is applied per-coordinate to vectors and matrices;
- at every layer, every input's last coefficient is 1, so that the last value of a neuron's weight is actually a bias–therefore there is no explicit bias addition operation[4] as opposed to Fig. 1;
- $S^{(i)}$ denote the pre-rectifier outputs, with capital letters since they are batch of vectors (as opposed to lowercase letter in Fig. 1).

In other words:

$$\begin{cases} S^{(i)} = W^{(i)}.X^{(i-1)} \\ X^{(i)} = \sigma(S^{(i)}) \text{ performed per coordinate} \end{cases}$$

2 Algorithm description

PCNS is described here for deep networks of fully-connected layers[5].

Given a partially trained network, the algorithm creates a shrinked network with smaller layers[6] whose output is a sound approximation of the original network's output (therefore its error rate is comparable).

[4] Including the bias in the weight matrix in this manner is only aesthetic : it simplifies notations, but does not semantically alter computations presented below (nor forward/backward propagations).

[5] Adaptations to convolutional kinds of layers (e.g. convolutional) seem clearly possible and will be studied in a next step (TODO rewrite this)

[6] how much smaller depends on the layer redundancy, see section 4.1

Training with the new, shrinked network can be continued, and the shrinking operation can be repeated multiple times during training [7]. If the shrinking gains are large, training time will be significantly reduced. This is a major difference w.r.t. distillation / pruning techniques that require full training and sometimes re-training afterwards.

Each hidden layer is shrinked by performing the 4 steps below (layer superscripts may be dropped when obvious from context).

2.1 Step 1: transposed outputs

Compute the output matrix $X^{(l)}$ of hidden layer l on a large[8] batch $\mathcal{B}$ and the *transposed outputs* (TOs) of the matrix—that is, the column vectors of ${}^{T}X^{(l)}$. Transposed output vector $TO_i^{(l)}$ of size B can be understood as the output of a single neuron (neuron i in layer l) over the batch, as shown below.

$$\sigma \odot \begin{pmatrix} - & W_1 & - \\ & \vdots & \\ - & W_N & - \end{pmatrix} \begin{pmatrix} | & & | \\ X_1^{(l-1)} & \cdots & X_B^{(l-1)} \\ | & & | \end{pmatrix} \begin{pmatrix} X_{1,1}^{(l)} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & X_{N,B}^{(l)} \end{pmatrix} = \begin{pmatrix} - & TO_1 & - \\ & \vdots & \\ - & TO_n & - \end{pmatrix} \tag{1}$$

2.2 Step 2: PCA approximation

Approximate the set of the transposed output vectors (TOs) using PCA, that is using the eigenvectors (EVs) of the experiment matrix of the TOs. The first M eigenvectors with the largest eigenvalues are selected such as the sum of the M eigenvalues matches a desired variance threshold[9] (thus M may be different for each layer):

$$\begin{aligned} TO_1 &\approx \hat{TO} + \sum_1^M \lambda_{1i} EV_i \\ &\vdots \\ TO_N &\approx \hat{TO} + \sum_1^M \lambda_{Ni} EV_i \end{aligned} \tag{2}$$

[7] The computational complexity of the algorithm is equivalent to a few training epochs on a batch. When to apply it for best results and fast training is discussed in section 4.4

[8] 'Large' here refers to being at least a few times bigger than the number of neurons of the layer. This is discussed more precisely in 4.3

[9] TODO

where $\hat{TO} = \sum_i TO_i$ is the mean vector of transposed outputs and $\lambda_{ji} =< TO_j | EV_i >$, dot product of TO_j and EV_i, is the i th coefficient of the projection of TO_j on the eigenvectors.

2.3 Step 3: Shrinked weight matrix

In the forward pass operation $\sigma(W.X)$, replace W by the *PCA-shrinked weight matrix* $\tilde{W}$ such that multiplying $\tilde{W}$ by X and applying σ yields approximated eigenvectors rather than transposed outputs. Compute $\tilde{W}$ by finding the least squares solution to the linear system below for each eigenvector EV_i as well as for $\hat{TO}$.

$$\tilde{W}_i.X = \sigma^{-1}(EV_i) \tag{3}$$

where we solve for $\tilde{W}_i$

Note that σ^{-1}, the inverse of σ, may not be defined or directly applicable depending on the bijectivity and domain of σ. For some activation functions there is no issue, e.g. *LeakyReLU* can be used directly (bijective with domain spanning $\mathbb{R}$) and *tanh* and *hardtanh* only need a minor adaptation for being applied on $\hat{TO}$ (since the EVs have there values in (-1,1) and both functions are bijective on the (-1,1) domain). The rest of this section assumes LeakyReLU is used.

For most other common activation functions the issue can also be overcome, see section 4.5.

$\tilde{W}$ has $M+1$ rows and is smaller than W (which has N rows). Replacing W by $\tilde{W}$ completely changes the layers' output on batch $\mathcal{B}$, since it yields approximated EVs (noted $\tilde{EV}_i$) rather than TOs, but each TO can be approximated by a linear combination of the EVs. A core assumption of PCNS is that the linear combination of EVs approximating TO_i (i.e. the output of neuron i on batch B) can be a good approximation of neuron i's output *in general*–that is, on other batches than B.

$$\sigma\odot \begin{pmatrix} - & \tilde{W}_1 & - \\ & \vdots & \\ - & \tilde{W}_M & - \\ - & \tilde{W}_{\hat{TO}} & - \end{pmatrix} \begin{pmatrix} | & & | \\ X_1 & \cdots & X_B \\ | & & | \end{pmatrix} \begin{pmatrix} - & \tilde{EV}_1 & - \\ & \vdots & \\ - & \tilde{EV}_M & - \\ - & \tilde{\hat{TO}} & - \end{pmatrix} \tag{4}$$

2.4 Step 4: Next layer projection

Replace the weight matrix $W^{(l+1)}$ of the following layer by $\bar{W}^{(l+1)} = W^{(l+1)}.P^{(l)}$ where $P^{(l)}$ is the projection matrix of the TOs on the EVs—therefore of dimension $(N_l, M+1)$; the resulting dimension of $\bar{W}^{(l+1)}$ being $(N_{l+1}, M+1)$. By doing this, $\tilde{\mathbf{s}}^{(l+1)}$, the pre-$\sigma$ output of layer $l+1$, will be a sound approximation[10] of $\mathbf{s}^{(l+1)}$, the pre-σ output before transforming layers with the PCNS algorithm, as schematised below.

$$\sigma \odot \begin{pmatrix} - & W_1^{(l)} & - \\ & \vdots & \\ - & W_{N_l}^{(l)} & - \end{pmatrix} \quad \begin{pmatrix} | \\ \mathbf{x}^{(l-1)} \\ | \end{pmatrix} \quad \begin{pmatrix} x_1^{(l)} = y_1 \\ \vdots \\ y_{N_l} \end{pmatrix} \qquad (5)$$

$$\begin{pmatrix} - & W_1^{(l+1)} & - \\ & \vdots & \\ - & W_{N_{l+1}}^{(l+1)} & - \end{pmatrix} \quad \begin{pmatrix} | \\ \mathbf{s}^{(l+1)} \\ | \end{pmatrix}$$

Before: Given a new input $\mathbf{x}^{(l-1)}$ to layer l, the forward pass yields $\mathbf{y} = \mathbf{x}^{(l)}$, then the following layer's pass is applied via $W^{(l+1)}$ to yield $s^{(l+1)}$

[10] The overall efficiency of PCNS relies on the soundness of this approximation, discussed in section 3

$$
\begin{pmatrix} | \\ \mathbf{x}^{(l-1)} \\ | \end{pmatrix}
$$

$$
\sigma \odot \begin{pmatrix} - & \tilde{W}_1^{(l)} & - \\ & \vdots & \\ - & \tilde{W}_{M+1}^{(l)} & - \end{pmatrix} \quad \begin{pmatrix} e_1 \\ \vdots \\ e_{M+1} \end{pmatrix}
$$

$$
\bar{W}^{(l+1)} \begin{cases} P^{(l)} = \begin{pmatrix} \lambda_{11} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \lambda_{N_l M+1} \end{pmatrix} & \begin{pmatrix} \sum_i \lambda_{1i} e_i \approx y_1 \\ \vdots \\ \sum_i \lambda_{N_l i} e_i \approx y_{N_l} \end{pmatrix} \\ W^{(l+1)} = \begin{pmatrix} - & W_1^{(l+1)} & - \\ & \vdots & \\ - & W_{N_{l+1}}^{(l+1)} & - \end{pmatrix} & \begin{pmatrix} | \\ \tilde{\mathbf{s}}^{(l+1)} \approx \mathbf{s}^{(l+1)} \\ | \end{pmatrix} \end{cases} \tag{6}
$$

After: Changing $W^{(l)}$ to $\tilde{W}^{(l)}$ the output of layer l, noted $\mathbf{e}$, and projecting using $P^{(l)}$ we fall back on approximated y then using $W^{(l+1)}$ continues forward propagation: $\bar{W}^{(l+1)} = P^{(l)}.W^{(l+1)}$ can be soundly substituted to $W^{(l+1)}$

Repeating the operation on all the hidden layers shrinks the whole network.

3 Core assumptions behind PCNS

4 Criteria for PCNS success

4.1 In-layer redundancy

- At a given layer, there is a lot of redundancy–that is different neurons are used partly to do the same computations. The most extreme examples would be 2 neurons with almost identical weights, or a neuron that is either always off, or always on (thus being redundant with the bias).
- Directly looking for these on weights does not incorporate the inputs distribution in the approximation, which is a fat mistake
- But such redundancy can be spotted by examining for each neuron sequences of what it outputs. On a long enough sequence, if there is

few differences between 2 neurons one can be killed–and more generally PCA can tell how much neurons we really need

- Potential effectiveness of PCNS can be measured by checking the number of required eigenvectors to reach a given variance

4.2 Soundness of approximation

Applying PCNS on layer l works if the outputs of layer $l+1$ after changing $W^{(l)}$ and $W^{(l+1)}$ are close to what they were before. This requires the 3 approximations below to perform well:

- the linear combination of eigenvectors of step 2 must approximate the transposed outputs well;
- the new weights of layer l as computed by step 3, must generate (via the forward pass) good approximations of the eigenvectors on the PCNS batch $\mathcal{B}$;
- the fact that the approximation is good on batch $\mathcal{B}$ must translate to it being good in general.

1. **TODO** The first one is ensured by the variance threshold discussed TODO, the second is a mystery, the third one is why we want a large batch see below section 4.3

4.3 Batch size

The PCNS batch size will determine the quality of the approximation (notably the 2nd & 3rd ones in section 4.2). A small batch size will imply a good approximation of the TO by step 2, but a poor generalization to other batches (similarly to overfitting). Therefore, a batch as large as possible ensures the best possible generalization–just as a training set as large as possible is often desirable. The limitation is the computational cost. The batch should be at least a few times bigger than the layer size, otherwise the risk of overfitting would be too great. Apart from that, the goal of PCNS being to reduce both the model size and training time, the chosen batch size would depend on the overall training methodology–e.g. if initial training is meant to run over tens thousands of epochs, choosing a PCNS batch size of 10 times the layer size and running PCNS every tens of thousands of epochs will be suitable since the pcns cost would be a small fraction of the training cost. Deciding precisely how the size of the batch impacts the precision of the algorithm is a question to be tackled.

4.4 Algorithm usage during training & complexity

The most costly step of the algorithm is the least square computations. The complexity of those are $O(N_{l+1}*N_l*B)$ and since B is supposed to be a few times bigger than N_l, this amounts to a cost of a few training epochs. Therefore, applying the algorithm every few epochs * 10 would ensure that its cost stays minimal as compared to the global cost of training.

4.5 Adapting PCNS to various activation functions

1. **TODO** Explain how bijective functions can be handled using a shift, and plain relu can be handled by decomposing into positive and negative components of eigenvectors.

4.6 last layer not concerned

(TODO rewrite) The operation can be performed on all *hidden* layers, therefore shrinking the whole network except for the last layer. PCNS on layer l operates on weights of both l and $l+1$ and the intermediate activations (the ones in between the 2 layers) have no direct relationship–therefore PCNS cannot be straightforwardly applied to the last layer. This operation should a priori not include the last layer, since its output is the final output which usually has a precise semantic–e.g. logprobabilities for multiclass classification. PCNS would mess with this semantic. However, note that the weights of the last layer will be changed by performing PCNS on the second-to-last layer, since PCNS on a layer changes not only weights of the layer itself, but also those of the following layer.

5 Example: how PCNS works on a toy network trained on MNIST

This section will illustrate PCNS on a toy network. TODO Work in progress

- Train a 2-layers fully connected network on MNIST
- Apply PCNS on the first layer, illustrate and explain what happens
- Apply PCNS on the second layer and show the performance is still quite good

- Train for a few more epochs and show the performance improves, for a fraction of the training time since the shrinked network is faster to train

6 Experiments

This section will test PCNS on more sophisticated Work in progress

- Test the algorithm on fully-connected networks applied to toy problems in which we expect redundancy;

6.1 TODO say the following at the right place

- next steps for this work: determine precisely the batch size, etc.
- Choose variance threshold
- Estimate network reduction, ensure a proper bound (can we measure if it worked / how well it worked?)

7 Next steps for this work

The next steps that are being taken to validate the PCNS algorithm are:

- Adapt the algorithm to convolutional networks (which would most likely translate in reducing the number of channels, not the size of the kernel)
- Check the algorithm is not redundant with other deep network optimization techniques (e.g. batch normalization, dropout, resnets) and can be seamlessly integrated with them;
- Measure redundancy (cf chap TODO) of various classical deep networks on standard problems (e.g. AlexNet on ImageNet) to see if there is one that would most benefit from being shrinked by PCNS
- Test PCNS on such a network and complete this work with the results.

8

9 DRAFT

9.1 Layer ordering

PCNS on the whole network should be performed in descending order starting with the second-to-last layer. Indeed, PCNS reduces the size of a neural layer by replacing the computation of the outputs by the computation of 'approximate eigenvectors' of the outputs. The approximation is made by computing eigenvectors of a batch of outputs (transposed see below), corresponding to a batch of inputs. These batches for the whole network have been computed via an initial forward pass. If we were to perform PCNS on layer l-1 before performing it on layer l, TBD

10 Citations