Paper Review: Data-Independent Neural Pruning via Coresets

By Ben Mussay et al.

Mitchell Gordon

Table of contents

1. **Neural Pruning Problem**
2. **Approx Single Output Neuron**
   - Neural Network is a Query Space
   - Coreset Algorithm
   - Additive Error Analysis
   - Other Results
3. **Coreset Per Layer**

Facts

- BERT (popular NLP model) costs ~$7k to train
- A recent NLP paper used ~$150k worth of compute
- BERT-Large does not fit in GPU memory, need Google TPU
- BERT parameters are 3 GB on disk
- Training BERT produces ~1.5k lbs of CO_2
- We would train bigger networks if we could

Model Compression

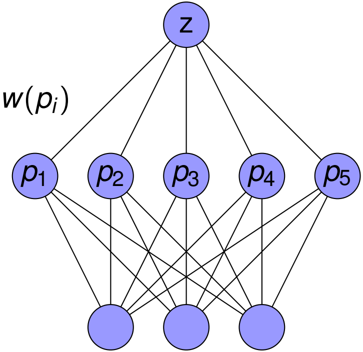
Definition (Model Compression)

Making a trained neural network smaller and faster. In practice, most trained neural networks are highly redundant.

Definition (Neural Pruning)

Removing neurons from a neural network

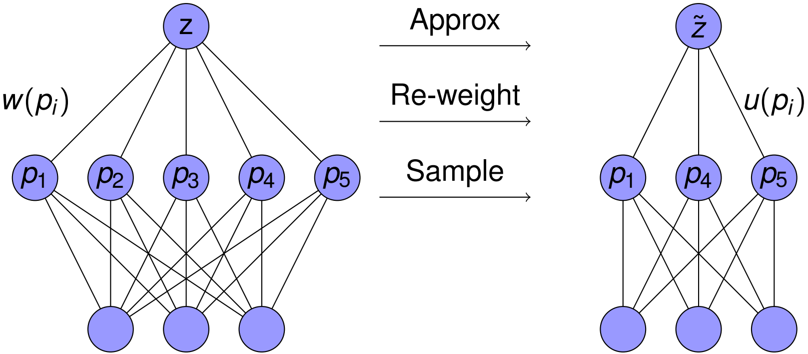
Neuron Pruning via Coresets



$$x \in \mathbb{R}^d, p_i \in P \subseteq \mathbb{R}^d$$

$$z = \sum_{p_i \in P} w(p_i)\varphi(p_i^T x)$$

Neuron Pruning via Coresets



$$x \in \mathbb{R}^d, p_i \in P \subseteq \mathbb{R}^d \qquad Q \subset P$$

$$z = \sum_{p_i \in P} w(p_i)\varphi(p_i^T x) \qquad \tilde{z} = \sum_{q_i \in Q} u(q_i)\varphi(q_i^T x)$$

Table of contents

1 Neural Pruning Problem

2 **Approx Single Output Neuron**
- Neural Network is a Query Space
- Coreset Algorithm
- Additive Error Analysis
- Other Results

3 Coreset Per Layer

Coreset Framework

Definition (Weighted Set)

Let $P \subseteq \mathbb{R}^d$ be a set, and w be a function that maps every $p \in P$ to a weight $w(p)$. The pair (P, w) is called a weighted set.

Definition (Query Space)

Let $P' = (P, w)$ be a weighted set called the input set. Let $X \subset \mathbb{R}^d$ be a set, and $f : P \times X \to [0, \infty)$ be a loss function. The tuple (P, w, X, f) is called a query space.

Translation Table

Query Space	Neural Network
Weighted Input Set (P, w)	Hidden neurons P + weights $w(p)$
Query Set X	Set of possible inputs, X
Loss Function $f : P \times X \to [0, \infty)$	$f(p, x) = \varphi(p^T x)$
Additive ϵ-Coreset Guaranetee	$\lvert z - \tilde{z} \rvert < \epsilon$

Coreset Algorithm

Input: weighted hidden neurons (P, w)
integer sample size $m \geq 1$
an (activation) function $\varphi : \mathbb{R} \to [0, \infty)$
an upper bound $\beta > ||x|| > 0$

Output: weighted neuron coreset (C, u)

Coreset Algorithm

Input: weighted hidden neurons (P, w)
integer sample size $m \geq 1$
an (activation) function $\varphi : \mathbb{R} \to [0, \infty)$
an upper bound $\beta > ||x|| > 0$

Output: weighted neuron coreset (C, u)

for every $p \in P$ **do**
  $pr(p) := \frac{w(p)\varphi(\beta||p||)}{\sum_{q \in P} w(q)\varphi(\beta||q||)}$
  $u(p) := 0$
end for
$C \leftarrow \emptyset$
for m iterations **do**
  Sample q from P w.p. $pr(q)$.
  $C := C \cup q$
  $u(q) := u(q) + \frac{w(q)}{m \cdot pr(q)}$
end for
return (C, u)

Analysis

Theorem (Additive Error Coreset - Braverman et al. (2016))

Let d be the VC-dimension of a query space (P, w, X, f). Suppose $s : P \to [0, \infty)$ such that $s(p) \geq w(p) sup_{x \in X} f(p, x)$. Let $t = \sum_{p \in P} s(p)$, and $\epsilon, \delta \in (0, 1)$. Let $c \geq 1$ be a sufficiently large constant that can be determined from the proof, and let C be a sample (multi-set) of

$$m \geq \frac{ct}{\epsilon^2}(d \log t + \log(\frac{1}{\delta}))$$

i.i.d. points from P, where for every $p \in P$ and $q \in C$ we have $pr(p = q) = s(p)/t$. Then, with probability at least $1 - \delta$,

$$\forall x \in X : |\sum_{p \in P} w(p) f(p, x) - \sum_{q \in C} \frac{w(q)}{m\, pr(q)} \dot{f}(q, x)| \leq \epsilon$$

Analysis

VC-dimension

We know the VC-dimension of neural networks with common activation functions (ReLU, sigmoid) is $O(d)$ (Anthony & Bartlett, 2009)

Weighted Query Space Loss (Weighted NN Activation)

We need an upper-bound on the weighted activation for each neuron. Assume:

- $X \subseteq \mathbb{B}_\beta$.
- $P \subseteq \mathbb{B}_\alpha$

Then the upper-bound on the activation is just a simple application of Cauchy-Schwartz.

$$f(p, x) = \varphi(p^T x) \leq \varphi(||p||||x||) \leq \varphi(||p||\beta) \leq \varphi(\alpha\beta)$$

Other Results

- Extension to Negative Weights
- Multiplicative Error Approximation Impossible

Table of contents

1 Neural Pruning Problem

2 Approx Single Output Neuron
- Neural Network is a Query Space
- Coreset Algorithm
- Additive Error Analysis
- Other Results

3 Coreset Per Layer

Coreset Per Layer Algorithm

Input: weighted sets $(P, w_1), \ldots, (P, w_k)$
integer sample size $m \geq 1$
an (activation) function $\varphi : \mathbb{R} \to [0, \infty)$
an upper bound $\beta > ||x|| > 0$

Output: weighted neuron coreset $(C, u_1, \ldots, u_k)$

for every $p \in P$ **do**

$pr(p) := \frac{max_{i \in [k]} w_i(p)\varphi(\beta||p||)}{\sum_{q \in P} max_{i \in [k]} w_i(q)\varphi(\beta||q||)}$

$u_i(p) := 0$

end for

$C \leftarrow \emptyset$

for m iterations **do**

Sample q from P w.p. $pr(q)$.

$C := C \cup q$

$\forall i \in [k] : u_i(q) := u_i(q) + \frac{w_i(q)}{m \dot{p} r(q)}$

end for

return $(C, u_1, \ldots, u_k)$

Coreset Per Layer

Corollary (Coreset per Layer)

Let $(P, w_1, \mathbb{B}_\beta(0), f), \ldots, (P, w_k, \mathbb{B}_\beta(0), f)$ *be k query spaces, each of VC-dimension* $O(d)$*, such that* $f(p, x) = \varphi(p^T x)$ *for some non-decreasing* $\varphi : \mathbb{R} \to [0, \infty)$ *and* $P \subseteq \mathbb{B}_\beta(0)$. *Let*

$$s(p) = max_{i \in [k]} sup_{x \in X} w_i(p) \varphi(p^T x)$$

Let $c \geq 1$ *be a sufficiently large constant that can be determined from the proof, and* $t = \sum_{p \in P} s(p)$

$$m \geq \frac{ct}{\epsilon^2}(d \log t + \log(\frac{1}{\delta}))$$

Coreset Per Layer

Corollary (Coreset per Layer cont'd)

Let $(C, u_1, \ldots, u_k)$ be the outtput of a call to CORESET$(P, w_1, \ldots, w_k, m, \varphi, \beta)$. Then, $|c| \leq m$ and, with probability at least $1 - \delta$,

$$\forall i \in [k], x \in \mathbb{B}_\beta : |\sum_{p \in P} w_i(p) f(p, x) - \sum_{q \in C} u_i \dot{f}(q, x)| \leq \epsilon$$

Proof.

The proof follows directly from the observation that

$$s(p) \geq w(p) sup_{x \in X} f(p, x)$$

References I

Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, Dan Feldman
Data-Independent Neural Pruning via Coresets, 2019.
https://arxiv.org/abs/1907.04018