Splitting Steepest Descent for Growing Neural Architectures

Qiang Liu, Lemeng Wu* and Dilin Wang*

Department of Computer Science
UT Austin
Our goal: finding small & accurate neural networks

- lightweight
- energy-efficient

- large models
- energy-intensive
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Splitting yields adaptive net structure optimization
Intuition: escaping local minima

- A simple network:

\[ \mathcal{L}(\theta) := \mathbb{E}_{x \sim D} \left[ \Phi( \sigma(\theta, x) ) \right]. \]
Intuition: escaping local minima

- Splitting \( \theta \) into \( m \) copies \( \{w_i, \theta_i\}_{i=1}^m \):

\[
\mathcal{L}(\{\theta_i, w_i\}) := \mathbb{E}_{x \sim D} \left[ \Phi \left( \sum_{i=1}^m w_i \sigma(\theta_i, x) \right) \right]
\]

- A simple network:

\[
\mathcal{L}(\theta) := \mathbb{E}_{x \sim D} \left[ \Phi \left( \sigma(\theta, x) \right) \right].
\]
Intuition: escaping local minima

- **Splitting $\theta$ into $m$ copies $\{w_i, \theta_i\}_{i=1}^m$:**

  $$\mathcal{L}(\{\theta_i, w_i\}) := \mathbb{E}_{x \sim D} \left[ \Phi\left( \sum_{i=1}^{m} w_i \sigma(\theta_i, x) \right) \right]$$

- **A simple network:**

  $$\mathcal{L}(\theta) := \mathbb{E}_{x \sim D} \left[ \Phi\left( \sigma(\theta, x) \right) \right].$$

- **Smooth loss change:**

  $$\sum_{i=1}^{m} w_i = 1, \quad ||\theta_i - \theta||_2 \leq \epsilon.$$
Splitting Steepest Descent

- How to choose $m$ and $\{\theta_i, w_i\}$ optimally?

$$\min_{m, \{\theta_i, w_i\}_{i=1}^m} \left\{ \mathcal{L}(\{\theta_i, w_i\}) - \mathcal{L}(\theta) \quad \text{s.t.} \quad ||\theta_i - \theta||_2 \leq \epsilon, \sum_{i=1}^m w_i = 1, \ w_i > 0, \ \forall \ i \right\}. $$
Splitting Steepest Descent

How to choose $m$ and $\{\theta_i, w_i\}$ optimally?

$$
\min_{m, \{\theta_i, w_i\}_{i=1}^m} \left\{ \mathcal{L}(\{\theta_i, w_i\}) - \mathcal{L}(\theta) \text{ s.t. } ||\theta_i - \theta||_2 \leq \epsilon, \sum_{i=1}^m w_i = 1, w_i > 0, \forall i \right\}.
$$

Splitting-index, minimum eigenvalue

$$
= \frac{\epsilon^2}{2} \min \left\{ \lambda_{\min}(S(\theta)), 0 \right\} + \mathcal{O}(\epsilon^3)
$$

CLOSED-FORM

with

$$
S(\theta) = \mathbb{E}_{x \sim D} \left[ \nabla_\sigma \Phi(\sigma(\theta, x)) \nabla^2_{\theta\theta} \sigma(\theta, x) \right],
$$

Splitting-matrix
Splitting Steepest Descent

- How to choose $m$ and $\{\theta_i, w_i\}$ optimally?

$$
\min_{m,\{\theta_i,w_i\}_{i=1}^m} \left\{ \mathcal{L}(\{\theta_i,w_i\}) - \mathcal{L}(\theta) \quad \text{s.t.} \quad ||\theta_i - \theta||_2 \leq \epsilon, \sum_{i=1}^m w_i = 1, w_i > 0, \forall i \right\}.
$$

**Splitting-index, minimum eigenvalue**

$$
= \frac{\epsilon^2}{2} \min \left\{ \lambda_{\min}(S(\theta)), 0 \right\} + \mathcal{O}(\epsilon^3)
$$

**CLOSED-FORM**

with

$$
S(\theta) = \mathbb{E}_{x \sim D} \left[ \nabla_\sigma \Phi(\sigma(\theta, x)) \nabla^2_{\theta\theta} \sigma(\theta, x) \right],
$$

- Optimal splitting strategy

$$
\lambda_{\min}S(\theta) \geq 0, \quad \text{no splitting}
$$

$$
\lambda_{\min}S(\theta) < 0, \quad m = 2, \quad \theta_1 = \theta + \epsilon v_{\min}(S(\theta)), \quad \theta_2 = \theta - \epsilon v_{\min}(S(\theta)), \quad w_1 = w_2 = 1/2.
$$
Our Algorithm

Continuous optimization!
Growing Interpretable Networks

- Training the interpretable neural network by Li et al., 2018\(^1\).

---

Results on CIFAR10

- Compare with pruning methods: batch-normalization-based pruning (Bn-prune) (Liu et al., 2017\(^1\)) and L1-based pruning (L1-prune) (Li et al., 2017\(^2\))

![Graphs showing test accuracy for MobileNet and VGG19 models compared with different pruning methods.](image)

2. Li et al., Pruning filters for efficient convnets. ICLR. 2017
Keyword Spotting on Microcontrollers

- Identifying a set of keywords from speech signal, e.g. “hey siri"
- use benchmark from Zhang et al., 2017

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>Params (K)</th>
<th>Ops (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>86.94</td>
<td>495.7</td>
<td>1.0</td>
</tr>
<tr>
<td>CNN</td>
<td>92.64</td>
<td>476.7</td>
<td>25.3</td>
</tr>
<tr>
<td>BasicLSTM</td>
<td>93.62</td>
<td>492.6</td>
<td>47.9</td>
</tr>
<tr>
<td>LSTM</td>
<td>94.11</td>
<td>495.8</td>
<td>48.4</td>
</tr>
<tr>
<td>GRU</td>
<td>94.72</td>
<td>498.0</td>
<td>48.4</td>
</tr>
<tr>
<td>CRNN</td>
<td>94.21</td>
<td>485.0</td>
<td>19.3</td>
</tr>
<tr>
<td>DS-CNN</td>
<td>94.85</td>
<td>413.7</td>
<td>56.9</td>
</tr>
<tr>
<td>Ours</td>
<td>95.36</td>
<td>282.6</td>
<td>39.2</td>
</tr>
</tbody>
</table>

Conclusion

- Incremental neural structure optimization with splitting gradient
- Simple and fast, promising in practice

Thank you!

Poster #35, Today 10:45am – 12:45am @East Exhibition Hall B+C