Learning Sparse Sharing Architectures for Multiple Tasks

Tianxiang Sun*, Yunfan Shao*, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, Xuanjing Huang

txsun19@fudan.edu.cn

New York City, 2020/02/09
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
What is Multi-Task Learning (MTL)?

“Multi-task Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias.”

How does MTL work?

- Representation Bias (Inductive Bias)

Formulation

- $T$ tasks: $\mathcal{D}_t = \{x^n_t, y^n_t\}_{n=1}^{N_t}$

- Shared layers $\mathcal{E}$ parameterized by $\theta_{\mathcal{E}} = \{\theta_{\mathcal{E},1}, \ldots, \theta_{\mathcal{E},L}\}$

- Task-specific layers $\mathcal{F}^t$ parameterized by $\theta^t_{\mathcal{F}}$

- Parameters: $\theta = (\theta_{\mathcal{E}}, \theta^1_{\mathcal{F}}, \ldots, \theta^T_{\mathcal{F}})$

- Objective: $\mathcal{L}(\theta) = \sum_{t=1}^{T} \lambda_t \sum_{n=1}^{N_t} \mathcal{L}_t(\hat{y}_n^t, y_n^t)$
Multi-Task Sharing Mechanisms

MTL is typically done with *parameter sharing*:

- Soft Sharing ([Misra et al. 2016](#); [Ruder et al. 2019](#))
- Hierarchical Sharing ([Søgaard and Goldberg 2016](#); [Hashimoto et al. 2017](#))
- ...

[Collobert and Weston 2008](#)
[Subramanian et al. 2018](#)
[Liu et al. 2019](#)
[Misra et al. 2016](#)
[Ruder et al. 2019](#)
[Søgaard and Goldberg 2016](#)
[Hashimoto et al. 2017](#)
Hard Sharing

- Stack the task-specific layers on the top of the shared layers
- Inference: $\hat{y}_n^t = F^t(E(x_n^t; \theta_E); \theta_F^t)$
- Advantages:
  1. easy to implement
  2. parameter efficient
- Disadvantages:
  Struggle with loosely related/unrelated tasks
  (Negative Transfer)
Soft Sharing

- Each task has separate model and parameters, but each model can access the information inside other models

- Advantages:
  Makes no assumptions about task relatedness

- Disadvantages:
  Not parameter-efficient
Hierarchical Sharing

- Put different task supervisions at different layers
- Inference: $\hat{y}_n^t = \mathcal{F}_t^t(\mathcal{E}(x_{n^t}; \theta_{\mathcal{E}(1:l)}; \theta_{\mathcal{F}}))$
- Advantages:
  1. more flexible than hard sharing
  2. more parameter-efficient than soft sharing
- Disadvantages:
  Hard to design an effective hierarchy
Limitations of Existing Sharing Mechanisms

- **Hard sharing**: Struggle with loosely related tasks
- **Hierarchical sharing**: Dependent on manual design
- **Soft sharing**: Parameter-inefficient
Motivation

Does there exist a multi-task sharing mechanism:

1. It is compatible with a wide range of tasks, regardless of whether the tasks are related or not.
2. It does not depend on manually designing the sharing structure based on characteristic of tasks.
3. It is parameter efficient.
Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
Sparse Sharing Mechanism

- Base Network: $\mathcal{E}$
- Assign each task a subnet
- Subnet: $\mathcal{E}^t(x) = \mathcal{E}(x; M_t \odot \theta_\mathcal{E})$
- Hard sharing $\rightarrow M_t = 1$
- Hierarchical sharing $\rightarrow$

$\theta_\mathcal{E} = \{\theta_{\mathcal{E},1}, \theta_{\mathcal{E},2}\}$

$M_1 = \{1, 0\}$

$M_2 = \{1, 1\}$
Views of Sparse Sharing

- Over-parameterized base net $\rightarrow$ Large hypothesis space
- Subnet $\rightarrow$ Hypothesis subspace
- Inductive bias $\rightarrow$ Subnet structure
- Parameter overlap $\rightarrow$ Task relatedness
- Biologically intuitive:
  1. Sparse topology (Pessoa 2014)
  2. Different subnets for different tasks (MacLeod 1991)
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
Overview of Our Approach
Generating Subnets for Each Task

- **Iterative Magnitude Pruning (IMP)**

1. Randomly initialize a neural network $f(x; \theta_0)$ (where $\theta_0 \sim \mathcal{D}_\theta$).
2. Train the network for $j$ iterations, arriving at parameters $\theta_j$.
3. Prune $p\%$ of the parameters in $\theta_j$, creating a mask $m$.
4. Reset the remaining parameters to their values in $\theta_0$, creating the winning ticket $f(x; m \odot \theta_0)$. 
Generating Subnets for Each Task

Iterative Magnitude Pruning (IMP)

Algorithm 1 Sparse Sharing Architecture Learning

Require: Base Network $\mathcal{E}$; Pruning rate $\alpha$; Minimal sparsity $S$; Datasets for $T$ tasks $\mathcal{D}_1, \ldots, \mathcal{D}_T$, where $\mathcal{D}_t = \{x^t_n, y^t_n\}_{n=1}^{N_t}$.

1: Randomly initialize $\theta_{\mathcal{E}}$ to $\theta_{\mathcal{E}}^{(0)}$.
2: \textbf{for} $t = 1 \cdots T$ \textbf{do}
3: \hspace{1em} Initialize mask $M^z_t = 1_{|\theta_{\mathcal{E}}|}$, where $z = 1$.
4: \hspace{1em} Train $\mathcal{E}(x; M^z_t \odot \theta_{\mathcal{E}})$ for $k$ steps with data sampled from $\mathcal{D}_t$, producing network $\mathcal{E}(x; M^z_t \odot \theta_{\mathcal{E}}^{(k)})$. Let $z \leftarrow z + 1$.
5: \hspace{1em} Prune $\alpha$ percent of the remaining parameters with the lowest magnitudes from $\theta_{\mathcal{E}}^{(k)}$. That is, let $M^z_t[j] = 0$ if $\theta_{\mathcal{E}}^{(k)}[j]$ is pruned.
6: \hspace{1em} If $\frac{\|M^z_t\|_0}{|\theta_{\mathcal{E}}|} \leq S$, the masks for task $t$ are $\{M^z_t\}_{i=1}^z$.
7: \hspace{1em} Otherwise, reset $\theta_{\mathcal{E}}$ to $\theta_{\mathcal{E}}^{(0)}$ and repeat steps 4-6 iteratively to learn more sparse subnetwork.
8: \textbf{end for}
9: \textbf{return} $\{M^z_1\}_{i=1}^z, \{M^z_2\}_{i=1}^z, \cdots, \{M^z_T\}_{i=1}^z$. 
Select Subnets

- Pick the subnet that performs best on the dev set.
- If there are multiple best-performing subnets, take the subnet with the lowest sparsity.

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>CHUNK</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50.12%</td>
<td>44.67%</td>
<td>56.23%</td>
</tr>
</tbody>
</table>
Training Subnets in Parallel

1. Select the next task $t$.
   - Proportional sampling \cite{Sanh2019}

2. Select a random mini-batch for task $t$.

3. Feed this batch of data into the subnetwork corresponding to task $t$, i.e. $\mathcal{E}(x; M_t \odot \theta_E)$.

4. Update the subnetwork parameters for this task by taking a gradient step with respect to this mini-batch.

5. Go to 1.

\[
\mathcal{L}(\theta) = \sum_{t=1}^{T} \lambda_t \sum_{n=1}^{N_t} \mathcal{L}_t(\hat{y}_n^t, y_n^t)
\]
Multi-Task Warmup (MTW)

Algorithm 1 Sparse Sharing Architecture Learning

**Require:** Base Network $\mathcal{E}$; Pruning rate $\alpha$; Minimal sparsity $S$; Datasets for $T$ tasks $\mathcal{D}_1, \cdots, \mathcal{D}_T$, where $\mathcal{D}_t = \{x_n^t, y_n^t\}_{n=1}^{N_t}$.

1. Randomly initialize $\theta_{\mathcal{E}}$ to $\theta_{\mathcal{E}}^{(0)}$.
2. for $t = 1 \cdots T$ do
3. Initialize mask $M_t^z = \mathbf{1}_{|\theta_{\mathcal{E}}|}$, where $z = 1$.
4. Train $\mathcal{E}(x; M_t^z \odot \theta_{\mathcal{E}})$ for $k$ steps with data sampled from $\mathcal{D}_t$, producing network $\mathcal{E}(x; M_t^z \odot \theta_{\mathcal{E}}^{(k)})$. Let $z \leftarrow z + 1$.
5. Prune $\alpha$ percent of the remaining parameters with the lowest magnitudes from $\theta_{\mathcal{E}}^{(k)}$. That is, let $M_t^z[j] = 0$ if $\theta_{\mathcal{E}}^{(k)}[j]$ is pruned.
6. If $\frac{\|M_t^z\|_0}{|\theta_{\mathcal{E}}|} \leq S$, the masks for task $t$ are $\{M_t^z\}_{i=1}^z$.
7. Otherwise, reset $\theta_{\mathcal{E}}$ to $\theta_{\mathcal{E}}^{(w)}$ and repeat steps 4-6 iteratively to learn more sparse subnetwork.
8. end for
9. return $\{M_1^z\}_{i=1}^z, \{M_2^z\}_{i=1}^z, \cdots, \{M_T^z\}_{i=1}^z$. 
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
Experiments

- **Tasks**: Part-of-Speech, NER, Chunking

- **Datasets**
  
  Exp1: CoNLL-2003
  
  Exp2: OntoNotes 5.0
  

- **Model Settings**
  Base model: **CNN-BiLSTM** ([Ma and Hovy 2016](#))
  
  Multi-Task baselines: hard/soft/hierarchical sharing
## Exp1 & Exp2

<table>
<thead>
<tr>
<th>Systems</th>
<th>POS Test Acc.</th>
<th>NER Test F1</th>
<th>Chunking Test F1</th>
<th># Params</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Δ</td>
<td>Δ</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exp1: CoNLL-2003</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single task</td>
<td>95.09</td>
<td>-</td>
<td>89.36</td>
<td>1602k</td>
</tr>
<tr>
<td>Single task (subnet)</td>
<td>95.11</td>
<td>+0.02</td>
<td>89.39 +0.03</td>
<td>811k</td>
</tr>
<tr>
<td>Hard sharing</td>
<td>95.34</td>
<td>+0.25</td>
<td>88.68 −0.68</td>
<td>534k</td>
</tr>
<tr>
<td>Soft sharing</td>
<td>95.16</td>
<td>+0.07</td>
<td>89.35 −0.01</td>
<td>1596k</td>
</tr>
<tr>
<td>Hierarchical sharing</td>
<td>95.09</td>
<td>+0.00</td>
<td>89.30 −0.06</td>
<td>1497k</td>
</tr>
<tr>
<td><strong>Sparse sharing (ours)</strong></td>
<td><strong>95.56</strong></td>
<td>+0.47</td>
<td><strong>90.35</strong> +0.99</td>
<td><strong>396k</strong></td>
</tr>
<tr>
<td><strong>Exp2: OntoNotes 5.0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single task</td>
<td>97.40</td>
<td>-</td>
<td>82.72</td>
<td>4491k</td>
</tr>
<tr>
<td>Single task (subnet)</td>
<td>97.42</td>
<td>+0.02</td>
<td>82.94 +0.22</td>
<td>1459k</td>
</tr>
<tr>
<td>Hard sharing</td>
<td>97.46</td>
<td>+0.06</td>
<td>82.95 +0.23</td>
<td>1497k</td>
</tr>
<tr>
<td>Soft sharing</td>
<td>97.34</td>
<td>−0.06</td>
<td>81.93 −0.79</td>
<td>4485k</td>
</tr>
<tr>
<td>Hierarchical sharing</td>
<td>97.22</td>
<td>−0.18</td>
<td>82.81 +0.09</td>
<td>1497k</td>
</tr>
<tr>
<td><strong>Sparse sharing (ours)</strong></td>
<td><strong>97.54</strong></td>
<td>+0.14</td>
<td><strong>83.42</strong> +0.70</td>
<td><strong>662k</strong></td>
</tr>
</tbody>
</table>
## Exp1 & Exp2

<table>
<thead>
<tr>
<th>Systems</th>
<th>POS Test Acc.</th>
<th>NER Test F1</th>
<th>Chunking Test F1</th>
<th># Params</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Δ</td>
<td>Δ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp1: CoNLL-2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single task</td>
<td>95.09</td>
<td>-</td>
<td>89.92</td>
<td>1602k</td>
</tr>
<tr>
<td>Single task (subnet)</td>
<td>95.11</td>
<td>+0.02</td>
<td>89.36</td>
<td>+0.03</td>
</tr>
<tr>
<td>Hard sharing</td>
<td>95.34</td>
<td>+0.25</td>
<td>88.68</td>
<td>-0.68</td>
</tr>
<tr>
<td>Soft sharing</td>
<td>95.16</td>
<td>+0.07</td>
<td>89.35</td>
<td>-0.01</td>
</tr>
<tr>
<td>Hierarchical sharing</td>
<td>95.09</td>
<td>+0.00</td>
<td>89.30</td>
<td>-0.06</td>
</tr>
<tr>
<td>Sparse sharing (ours)</td>
<td><strong>95.56</strong></td>
<td>+0.47</td>
<td><strong>90.35</strong></td>
<td>+0.99</td>
</tr>
<tr>
<td>Exp2: OntoNotes 5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single task</td>
<td>97.40</td>
<td>-</td>
<td>95.21</td>
<td></td>
</tr>
<tr>
<td>Single task (subnet)</td>
<td>97.42</td>
<td>+0.02</td>
<td>82.94</td>
<td>+0.22</td>
</tr>
<tr>
<td>Hard sharing</td>
<td>97.46</td>
<td>+0.06</td>
<td>82.95</td>
<td>+0.23</td>
</tr>
<tr>
<td>Soft sharing</td>
<td>97.34</td>
<td>-0.06</td>
<td>81.93</td>
<td>-0.79</td>
</tr>
<tr>
<td>Hierarchical sharing</td>
<td>97.22</td>
<td>-0.18</td>
<td>82.81</td>
<td>+0.09</td>
</tr>
<tr>
<td>Sparse sharing (ours)</td>
<td><strong>97.54</strong></td>
<td>+0.14</td>
<td><strong>83.42</strong></td>
<td>+0.70</td>
</tr>
</tbody>
</table>
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
About Negative Transfer

In the future studies, there are several issues to be addressed. Firstly, outlier tasks, which are unrelated to other tasks, are well known to hamper the performance of all the tasks when learning them jointly. There are some methods to alleviate negative effects outlier tasks bring. However, there lacks principled ways and theoretical analyses to study the resulting negative effects. In order to make MTL safe to be used by human, this is an important issue and needs more studies.

About Negative Transfer

- Construct an unrelated multi-task setting
  - Real: Named Entity Recognition (NER)
  - Synthetic: Position Prediction (PP)

<table>
<thead>
<tr>
<th></th>
<th>NER</th>
<th>Δ</th>
<th>PP</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single task</td>
<td>71.05</td>
<td>-</td>
<td>99.21</td>
<td>-</td>
</tr>
<tr>
<td>Hard sharing</td>
<td>61.62</td>
<td>-9.43</td>
<td>99.50</td>
<td>0.29</td>
</tr>
<tr>
<td>Sparse sharing</td>
<td>71.46</td>
<td>0.41</td>
<td>99.45</td>
<td>0.24</td>
</tr>
</tbody>
</table>
About Task Relatedness

- Define mask overlap ratio (OR) as:

\[
\text{OR}(M_1, M_2, \cdots, M_T) = \frac{\| \bigcap_{t=1}^{T} M_t \|_0}{\| \bigcup_{t=1}^{T} M_t \|_0}
\]

<table>
<thead>
<tr>
<th>Task Pairs</th>
<th>Mask OR</th>
<th>( \Delta(S^2 - HS) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS &amp; NER</td>
<td>0.18</td>
<td>0.4</td>
</tr>
<tr>
<td>NER &amp; Chunking</td>
<td>0.20</td>
<td>0.34</td>
</tr>
<tr>
<td>POS &amp; Chunking</td>
<td>0.50</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 7: Mask Overlap Ratio (OR) and the improvement for sparse sharing \( (S^2) \) compared to hard sharing \( (HS) \) of tasks on CoNLL-2003. The improvement is calculated using the average performance on the test set.
About Sparsity

- Combinations of subnets with different sparsity

![Graph showing average performance and mask overlap ratio against proportion of parameters remaining. The graph highlights hard sharing.](image)
Outline

Multi-Task Learning: Sharing Mechanisms

Sparse Sharing Mechanism

Approach: Learning Sparse Sharing Architectures

Experiments

Analysis and Discussions

Conclusion
Conclusion

Does sparse sharing architecture meets the requirements?

1. It is compatible with a wide range of tasks, regardless of whether the tasks are related or not.
2. It does not depend on manually designing the sharing structure based on characteristic of tasks.
3. It is parameter efficient.

It seems **YES!**
Thanks!

Q & A

txsun19@fudan.edu.cn
## Sequence Labeling Tasks

- POS, NER and Chunking

<table>
<thead>
<tr>
<th>Words</th>
<th>Results of</th>
<th>South</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>NNS</td>
<td>IN</td>
<td>JJ</td>
</tr>
<tr>
<td>NER</td>
<td>O</td>
<td>O</td>
<td>B-MISC</td>
</tr>
<tr>
<td>Chunk.</td>
<td>B-NP</td>
<td>B-PP</td>
<td>B-NP</td>
</tr>
</tbody>
</table>
CNN-BiLSTM: A popular architecture in sequence labeling tasks