Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation

This talk was presented by Brian Thompson at WMT at EMNLP 2018

It is based on this paper:
http://www.aclweb.org/anthology/W18-6313
bib: http://aclweb.org/anthology/W18-6313.bib
Freezing Subnetworks
to Analyze Domain Adaptation
in Neural Machine Translation

Brian Thompson†, Huda Khayrallah†, Antonios Anastasopoulos‡, Arya D. McCarthy†, Kevin Duh†, Rebecca Marvin†, Paul McNamee†, Jeremy Gwinnup°, Tim Anderson°, and Philipp Koehn†

†Johns Hopkins University, ‡University of Notre Dame, 
°Air Force Research Laboratory
Continued Training

- Random Initialized NMT Model
  - Train on general domain data
  - General Domain NMT Model
  - Continue training on in-domain data
  - Domain Adapted NMT Model

brian.thompson@jhu.edu

Freezing Subnetworks to Analyze Domain Adaptation
### Corpora

<table>
<thead>
<tr>
<th>Languages</th>
<th>General Domain (WMT + OpenSubtitles)</th>
<th>In Domain (Patents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-En</td>
<td>5.8M + 22M</td>
<td>820k</td>
</tr>
<tr>
<td>Ko-En</td>
<td>0 + 1.4M</td>
<td>81k</td>
</tr>
<tr>
<td>Ru-En</td>
<td>25M + 26M</td>
<td>29k</td>
</tr>
</tbody>
</table>

(size in lines)

In-domain data: Patent abstracts from the World Intellectual Property Organization (WIPO)
### General-Domain:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenSubtitles</td>
<td>You’re gonna need a bigger boat.</td>
</tr>
<tr>
<td>WMT</td>
<td>Intensified communication and sharing of information between the project partners enables the transfer of expertise in rural tourism.</td>
</tr>
</tbody>
</table>

### In-Domain:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents</td>
<td>The films coated therewith, in particular polycarbonate films coated therewith, have improved properties with regard to scratch resistance, solvent resistance, and reduced oiling effect, said films thus being especially suitable for use in producing plastic parts in film insert molding methods.</td>
</tr>
</tbody>
</table>
Models

Freezing Subnetworks to Analyze Domain Adaptation

- BLEU (Patents)
  - German
  - Korean
  - Russian

- General Domain

brian.thompson@jhu.edu
Freezing Subnetworks to Analyze Domain Adaptation

**BLEU (Patents)**

- **German:** +26.3
- **Korean:** +29.0
- **Russian:** +13.6

**Models**

- General Domain
- Continued Training
Background
Continued Training
Corpora
Models

Subnetworks
Analysis-1
Distance Sensitivity
Analysis-2
Freeze 1/5
Freeze 4/5
Discussion

Models

Freezing Subnetworks to Analyze Domain Adaptation

---

**Models**

---

**BLEU (Patents)**

- **German**: +0.4
- **Korean**: +1.8
- **Russian**: +10.1

---

**Legend**

- General Domain
- Continued Training
- In-Domain

---

brian.thompson@jhu.edu

Freezing Subnetworks to Analyze Domain Adaptation
Freezing Subnetworks to Analyze Domain Adaptation

BLEU (Patents)

<table>
<thead>
<tr>
<th>Language</th>
<th>General Domain</th>
<th>Continued Training</th>
<th>In-Domain</th>
<th>Online A</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td></td>
<td>+11.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td>+4.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td>+7.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

brian.thompson@jhu.edu
Subnetworks

- **Target Embedding**: 15.1M
- **Softmax**: 15.1M
- **Decoder**: 6.8M
- **Encoder**: 3.7M
- **Source Embedding**: 15.4M

- **Wash**
- **your**
- **hands**
- **Wasch**
- **dir**
- **die**
- **Hände**

Freezing Subnetworks to Analyze Domain Adaptation
Freezing Subnetworks to Analyze Domain Adaptation

Subnetworks

Target Embedding
15.1M
Softmax
15.1M
Decoder
6.8M
Encoder
3.7M
Source Embedding
15.4M

Wash
your
hands

Wasch
dir
die
Hände

brian.thompson@jhu.edu
Subnetworks

Freezing Subnetworks to Analyze Domain Adaptation

Wash your hands
Wasch dir die Hände
Softmax 15.1M
Decoder 6.8M
Source Embedding 15.4M
Target Embedding 15.1M
Encoder 3.7M

Discussion
Freeze 1/5
Freeze 4/5
Freezing Subnetworks to Analyze Domain Adaptation

- **Target Embedding**: 15.1M
- **Softmax**: 15.1M
- **Decoder**: 6.8M
- **Encoder**: 3.7M
- **Source Embedding**: 15.4M

**Words andConcepts**

- Wash
- your
- hands
- Wasch
- dir
- die
- Hände
Subnetworks

<table>
<thead>
<tr>
<th>Subnetworks</th>
<th>Wash</th>
<th>your</th>
<th>hands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Embedding</td>
<td>15.1M</td>
<td>Softmax</td>
<td>15.1M</td>
</tr>
<tr>
<td>Decoder</td>
<td>6.8M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoder</td>
<td>3.7M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Embedding</td>
<td>15.4M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Wash your hands
Wasch dir die Hände

6.8M Decoder
3.7M Encoder
15.4M Source Embedding
15.1M Softmax

brian.thompson@jhu.edu

Freezing Subnetworks to Analyze Domain Adaptation
Freezing Subnetworks to Analyze Domain Adaptation

Target Embedding
15.1M
Softmax
15.1M
Decoder
6.8M
Encoder
3.7M
Source Embedding
15.4M
Wasch
your
Hands
Wash
dir
die
Hände

brian.thompson@jhu.edu
Subnetworks

Freezing Subnetworks to Analyze Domain Adaptation

- Background
  - Continued Training
  - Corpora
  - Models

- Subnetworks

- Analysis-1
  - Distance
  - Sensitivity

- Analysis-2
  - Freeze 1/5
  - Freeze 4/5

- Discussion

<table>
<thead>
<tr>
<th>Subnetworks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Embedding</strong></td>
</tr>
<tr>
<td><strong>Softmax</strong> 15.1M</td>
</tr>
<tr>
<td><strong>Decoder</strong> 6.8M</td>
</tr>
<tr>
<td><strong>Encoder</strong> 3.7M</td>
</tr>
<tr>
<td><strong>Source Embedding</strong></td>
</tr>
<tr>
<td><strong>15.4M</strong></td>
</tr>
</tbody>
</table>

- Wash 15.1M
- your
- hands

- Wash your hands
- Wasch dir die Hände

- Softmax
- Decoder
- Encoder
- Source Embedding

- brian.thompson@jhu.edu
How much do parameters change during continued training?

(RMS Change)
Per-Component Sensitivity Analysis

Performance (BLEU) as a function of noise (standard deviation) added to a given component.

(Russian)

<table>
<thead>
<tr>
<th>Component</th>
<th>$L^2$ Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax</td>
<td>0.14</td>
</tr>
<tr>
<td>Encoder</td>
<td>0.22</td>
</tr>
<tr>
<td>Decoder</td>
<td>0.24</td>
</tr>
<tr>
<td>Src. Emb</td>
<td>0.20</td>
</tr>
<tr>
<td>Tgt. Emb</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Freezing One Component at a Time

Question: How much does the model / training procedure depend on any **single** component for adaptation?
Freezing One Component at a Time

Question: How much does the model / training procedure depend on any **single** component for adaptation?

![Bar Chart]

- **General Domain**: 23.32
- **Encoder**: 20
- **Decoder**: 25
- **Source Embed**: 20
- **Target Embed**: 20
- **Softmax**: 30
- **Continued Training**: +11.4

**Russian**

1. When initial general-domain model is reasonably good
Freezing One Component at a Time

Question: How much does the model / training procedure depend on any single component for adaptation?
Answer: Not much\(^1\)

![Graph showing BLEU scores for different components](image)

\(^1\)When initial general-domain model is reasonably good
Freezing One Component at a Time

Question: How much does the model / training procedure depend on any **single** component for adaptation?

Answer: **Not much**

When initial general-domain model is reasonably good

\[ \text{(German)} \]
Question: How much does the model / training procedure depend on any **single** component for adaptation?

Answer: **Not much**\(^1\)

(Korean)

---

\(^1\)When initial general-domain model is reasonably good
Freezing All But One Component at a Time

Question: How much can the model / training procedure adapt using only a single component?

Answer: A lot!
Freezing All But One Component at a Time

Question: How much can the model / training procedure adapt using only a **single** component?

Answer: **A lot!**\(^1 \text{ } ^2\)

---

1When initial general-domain model is reasonably good
2Except for the target embeddings
Question: How much can the model / training procedure adapt using only a **single** component?

Answer: **A lot!**

1. When initial general-domain model is reasonably good
2. Except for the target embeddings

(Bar chart showing BLEU scores for different components.)
Freezing All But One Component at a Time

Question: How much can the model / training procedure adapt using only a single component?

Answer: A lot!\(^1\)\(^2\)

\(^1\)When initial general-domain model is reasonably good
\(^2\)Except for the target embeddings

(Korean)

<table>
<thead>
<tr>
<th>Component</th>
<th>BLEU (Patents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Domain</td>
<td>2.7</td>
</tr>
<tr>
<td>Encoder</td>
<td>+14.3</td>
</tr>
<tr>
<td>Decoder</td>
<td>+13.1</td>
</tr>
<tr>
<td>Source Embed</td>
<td>+10.1</td>
</tr>
<tr>
<td>Target Embed</td>
<td>+3.3</td>
</tr>
<tr>
<td>Softmax</td>
<td>+8.2</td>
</tr>
<tr>
<td>Continued Training</td>
<td>+29.0</td>
</tr>
</tbody>
</table>
Discussion

- Single components capable of adapting entire system
- Could effect be replicated without parallel data?
Discussion

- Single components capable of adapting entire system
  - Could effect be replicated without parallel data?

- Adaptation successful with small subset of parameters
  - Regularization techniques (Khayrallah et al. 2018)
  - Adapt subsets of parameters (Vilar, 2018)
Discussion

- Single components capable of adapting entire system
  - Could effect be replicated without parallel data?

- Adaptation successful with small subset of parameters
  - Regularization techniques (Khayrallah et al. 2018)
  - Adapt subsets of parameters (Vilar, 2018)

- DNNs are difficult to inspect/understand
  - But we can run experiments!
Acknowledgements

Thanks to:

- Lane Schwartz and Graham Neubig for organizing MTMA
- Michael Denkowski and David Vilar for Sockeye help
- NDSEG Fellowship, NSF Award 1464553, and DARPA LORELEI for funding