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
### 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>

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

<table>
<thead>
<tr>
<th>General-Domain:</th>
<th>In-Domain:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OpenSubtitles</strong></td>
<td><strong>Patents</strong></td>
</tr>
<tr>
<td>You’re gonna need a bigger boat.</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>

**WMT**
Intensified communication and sharing of information between the project partners enables the transfer of expertise in rural tourism.
Models

Freezing Subnetworks to Analyze Domain Adaptation

[Bar chart showing BLEU scores for German, Korean, and Russian in General Domain and Continued Training.]
Models

FREEZING SUBNETWORKS TO ANALYZE DOMAIN ADAPTATION

BLEU (PATENTS)

<table>
<thead>
<tr>
<th>Language</th>
<th>General Domain</th>
<th>Continued Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td></td>
<td>+26.3</td>
</tr>
<tr>
<td>Korean</td>
<td></td>
<td>+29.0</td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td>+13.6</td>
</tr>
</tbody>
</table>

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

brian.thompson@jhu.edu
Models

Freezing Subnetworks to Analyze Domain Adaptation

The diagram shows the BLEU scores (Patents) for German, Korean, and Russian domains. The scores are compared across General Domain, Continued Training, and In-Domain conditions.

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

The model's performance improves with continued training and in-domain adaptation, as evidenced by the higher scores in these conditions compared to the general domain.
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

Wash
your
hands
Wasch
dir
die
Hände

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

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

Words and their corresponding subnetworks:
- **Wash**
- **your**
- **hands**
- **Wasch**
- **dir**
- **die**
- **Hände**

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
Subnetworks

Freezing Subnetworks to Analyze Domain Adaptation

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

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

Background
Continued Training
Corpora
Models

Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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

brian.thompson@jhu.edu

Freezing Subnetworks to Analyze Domain Adaptation
Subnetworks

Freezing Subnetworks to Analyze Domain Adaptation
Change During Adaptation

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.

<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>
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?

![Graph showing BLEU scores for various components](image_url)

1 When initial general-domain model is reasonably good.

---

brian.thompson@jhu.edu

Freezing Subnetworks to Analyze Domain Adaptation
Freezing One Component at a Time

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

1When 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\(^1\)

---

1When 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\(^1\)

(Korean)

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

---

<table>
<thead>
<tr>
<th>BLEU (Patents)</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Domain</td>
<td>23.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoder</td>
<td>+11.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoder</td>
<td>+9.9</td>
<td>+10.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Embed</td>
<td></td>
<td>+4.0</td>
<td>+9.3</td>
<td>+11.4</td>
</tr>
<tr>
<td>Target Embed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Softmax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continued Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Russian)
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!**

<table>
<thead>
<tr>
<th>Component</th>
<th>BLEU (Patents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Domain</td>
<td>36</td>
</tr>
<tr>
<td>Encoder</td>
<td>+20.0</td>
</tr>
<tr>
<td>Decoder</td>
<td>+18.8</td>
</tr>
<tr>
<td>Source Embed</td>
<td>+18.1</td>
</tr>
<tr>
<td>Target Embed</td>
<td></td>
</tr>
<tr>
<td>Softmax</td>
<td>+16.7</td>
</tr>
<tr>
<td>Continued Training</td>
<td>+26.3</td>
</tr>
</tbody>
</table>

(German)

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

Answer: **A lot!**

![BLEU (Patents) graph]

(Korean)

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