



Background
Continued
Training
Corpora
Models

Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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

Background

Continued Training

Corpora
Models

Subnetworks

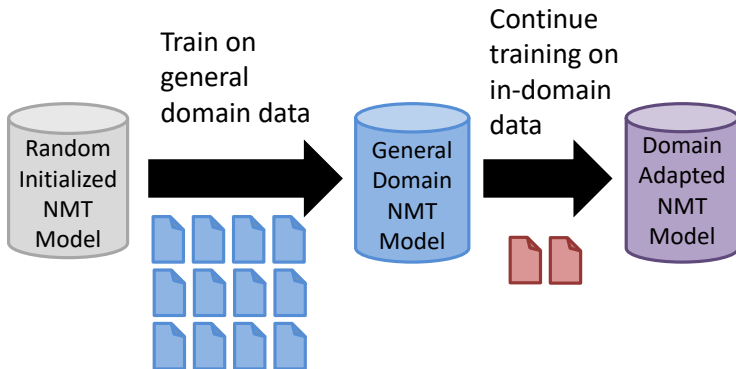
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Background

Continued
Training

Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

Languages	General Domain (WMT + OpenSubtitles)	In Domain (Patents)
De-En	5.8M + 22M	820k
Ko-En	0 + 1.4M	81k
Ru-En	25M + 26M	29k

(size in lines)

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



Data Examples

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

General-Domain:

OpenSubtitles

You're gonna need a bigger boat.

WMT

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

In-Domain:

Patents

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.



Models

Background

Continued
Training

Corpora

Models

Subnetworks

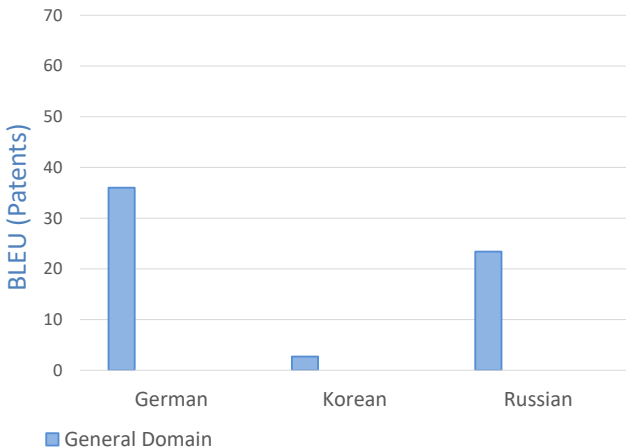
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Models

Background

Continued
Training

Corpora

Models

Subnetworks

Analysis-1

Distance

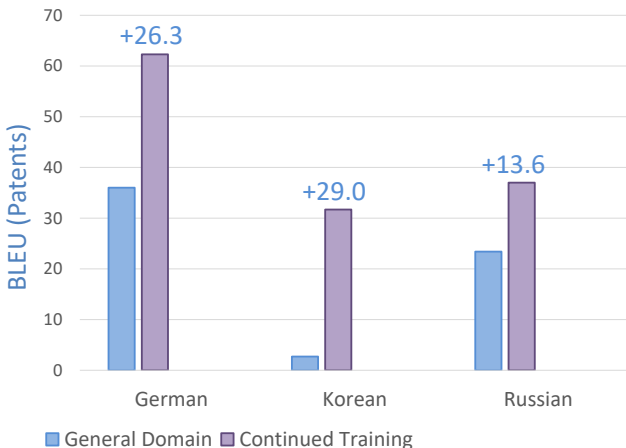
Sensitivity

Analysis-2

Freeze 1/5

Freeze 4/5

Discussion





Models

Background

Continued
Training

Corpora

Models

Subnetworks

Analysis-1

Distance

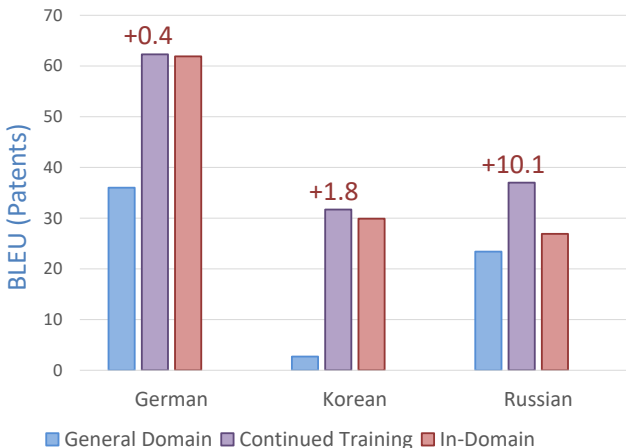
Sensitivity

Analysis-2

Freeze 1/5

Freeze 4/5

Discussion





Models

Background

Continued
Training

Corpora

Models

Subnetworks

Analysis-1

Distance

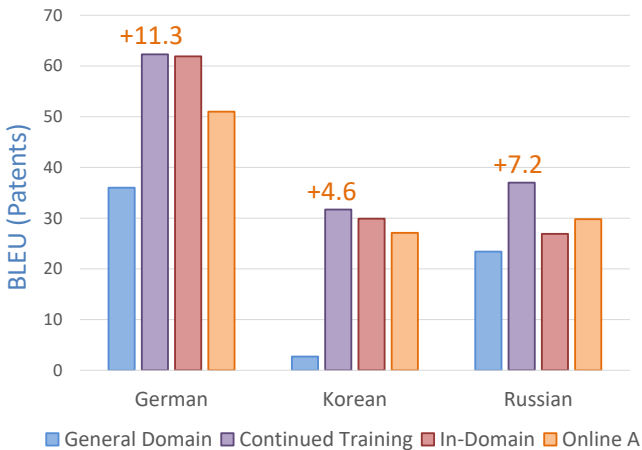
Sensitivity

Analysis-2

Freeze 1/5

Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

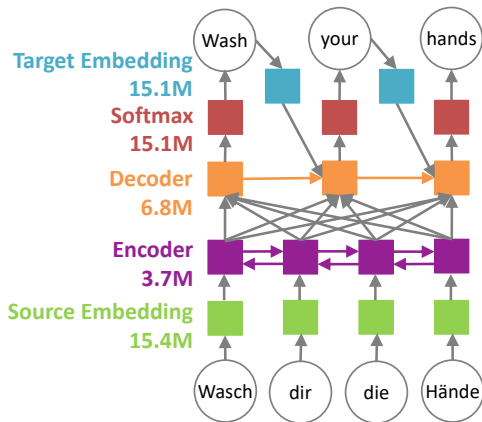
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

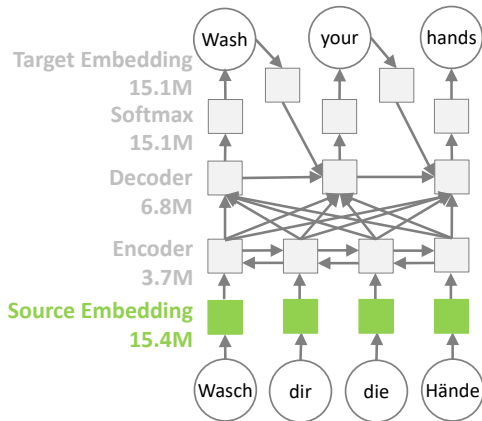
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

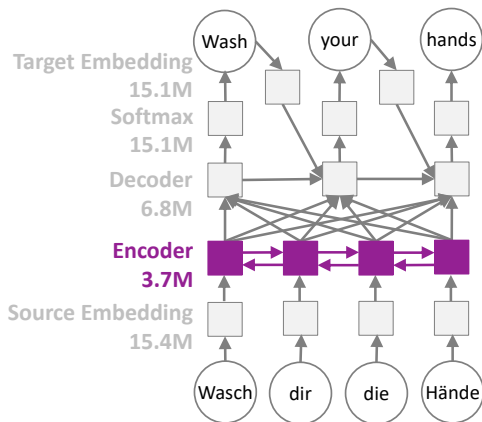
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

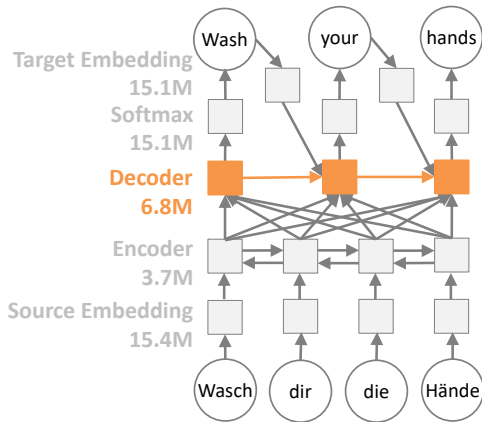
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

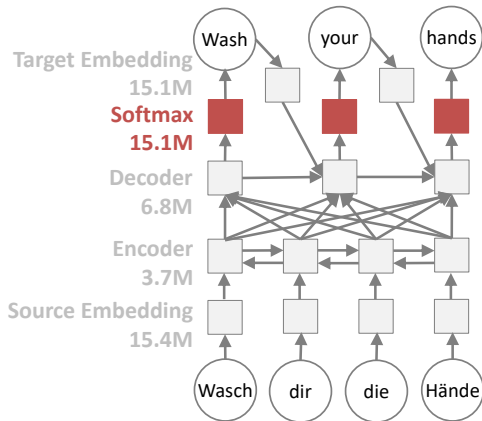
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

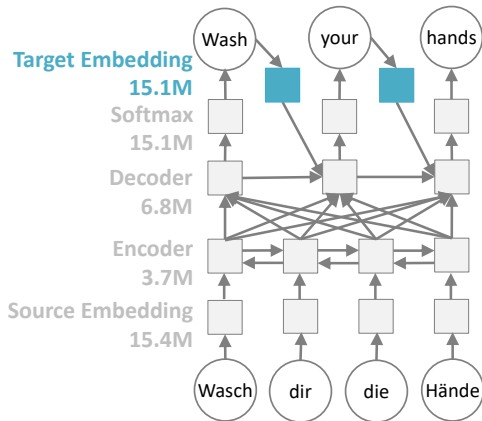
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Subnetworks

Background

Continued
Training
Corpora
Models

Subnetworks

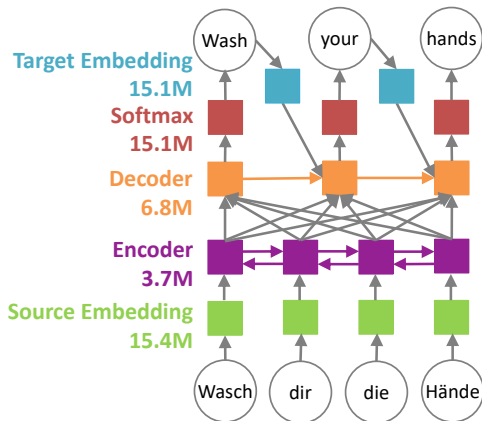
Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion





Change During Adaptation

Background

Continued Training
Corpora
Models

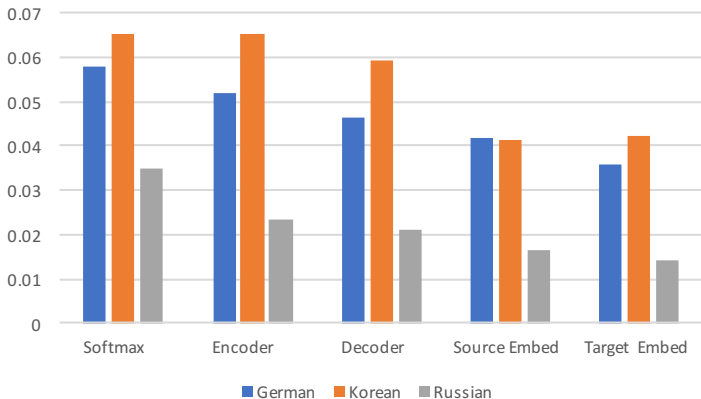
Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

How much do parameters change during continued training?



(RMS Change)



Per-Component Sensitivity Analysis

Background
Continued
Training
Corpora
Models

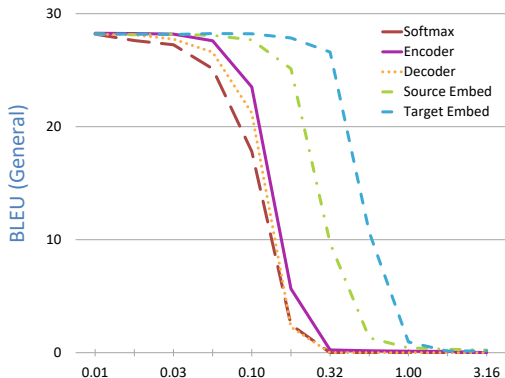
Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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



Component	L^2 Norm
Softmax	0.14
Encoder	0.22
Decoder	0.24
Src. Emb	0.20
Tgt. Emb	0.20

(Russian)



Freezing One Component at a Time

Background

- Continued Training
- Corpora
- Models

Subnetworks

Analysis-1

- Distance
- Sensitivity

Analysis-2

- Freeze 1/5**
- Freeze 4/5

Discussion

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



Freezing One Component at a Time

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

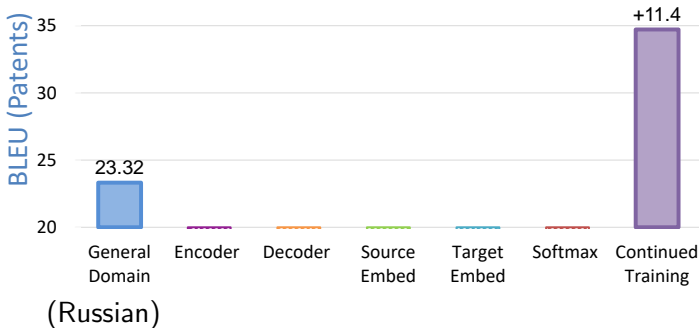
Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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



¹When initial general-domain model is reasonably good



Freezing One Component at a Time

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

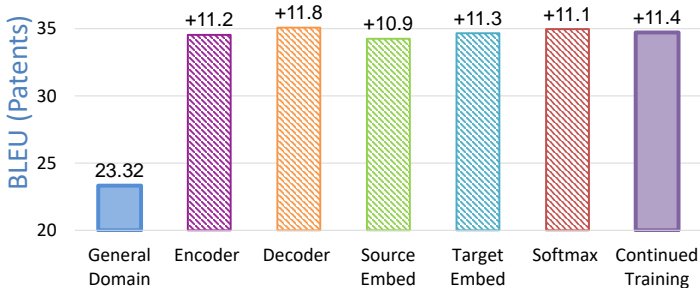
Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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

Answer: **Not much**¹



(Russian)

¹When initial general-domain model is reasonably good



Freezing One Component at a Time

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

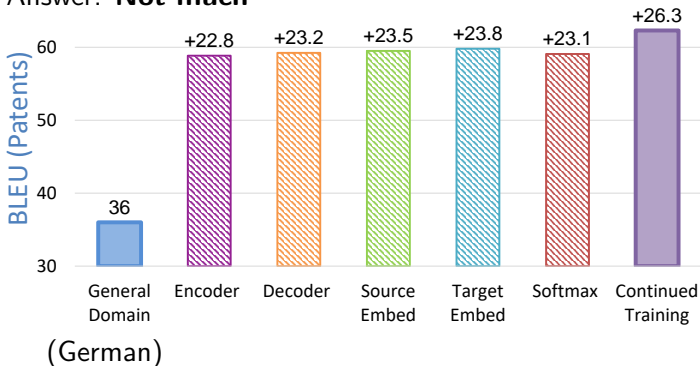
Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

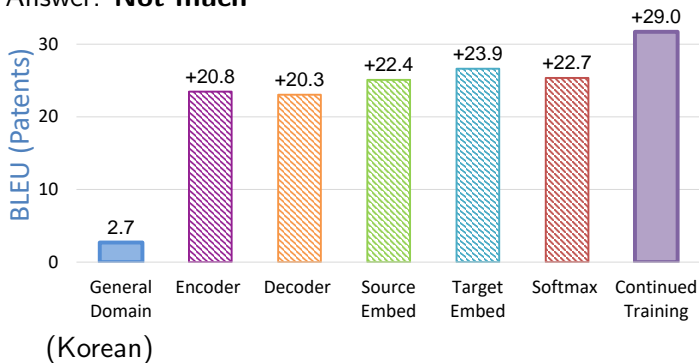
Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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 All But One Component at a Time

Background
Continued
Training
Corpora
Models

Subnetworks

Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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



Freezing All But One Component at a Time

Background
Continued
Training
Corpora
Models

Subnetworks

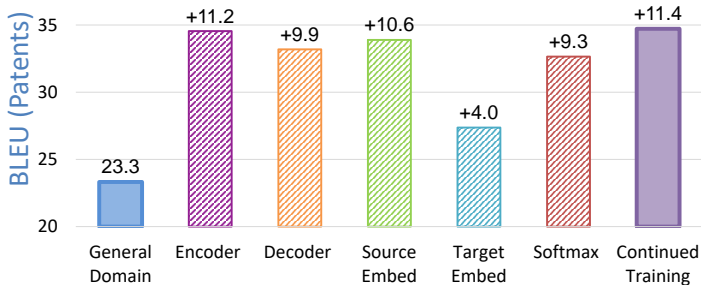
Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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

Answer: **A lot!**^{1 2}



(Russian)

¹When initial general-domain model is reasonably good

²Except for the target embeddings



Freezing All But One Component at a Time

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

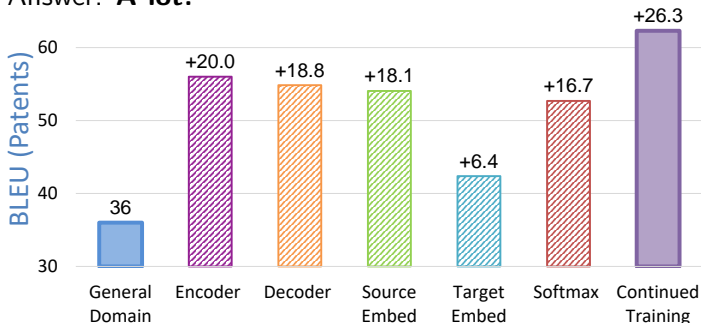
Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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

Answer: **A lot!**^{1 2}



(German)

¹When initial general-domain model is reasonably good

²Except for the target embeddings



Freezing All But One Component at a Time

Background
Continued
Training
Corpora
Models

Subnetworks

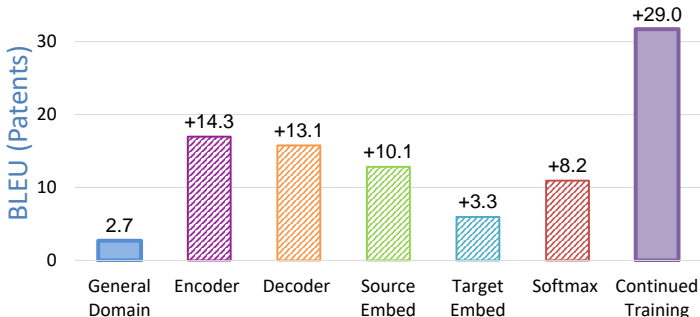
Analysis-1
Distance
Sensitivity

Analysis-2
Freeze 1/5
Freeze 4/5

Discussion

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

Answer: **A lot!**^{1 2}



(Korean)

¹When initial general-domain model is reasonably good

²Except for the target embeddings



Discussion

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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



Discussion

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

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

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

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

Background

Continued
Training
Corpora
Models

Subnetworks

Analysis-1

Distance
Sensitivity

Analysis-2

Freeze 1/5
Freeze 4/5

Discussion

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