Freezing Subnetworks to Analyze Domain Adaptation in Neural Machine Translation

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

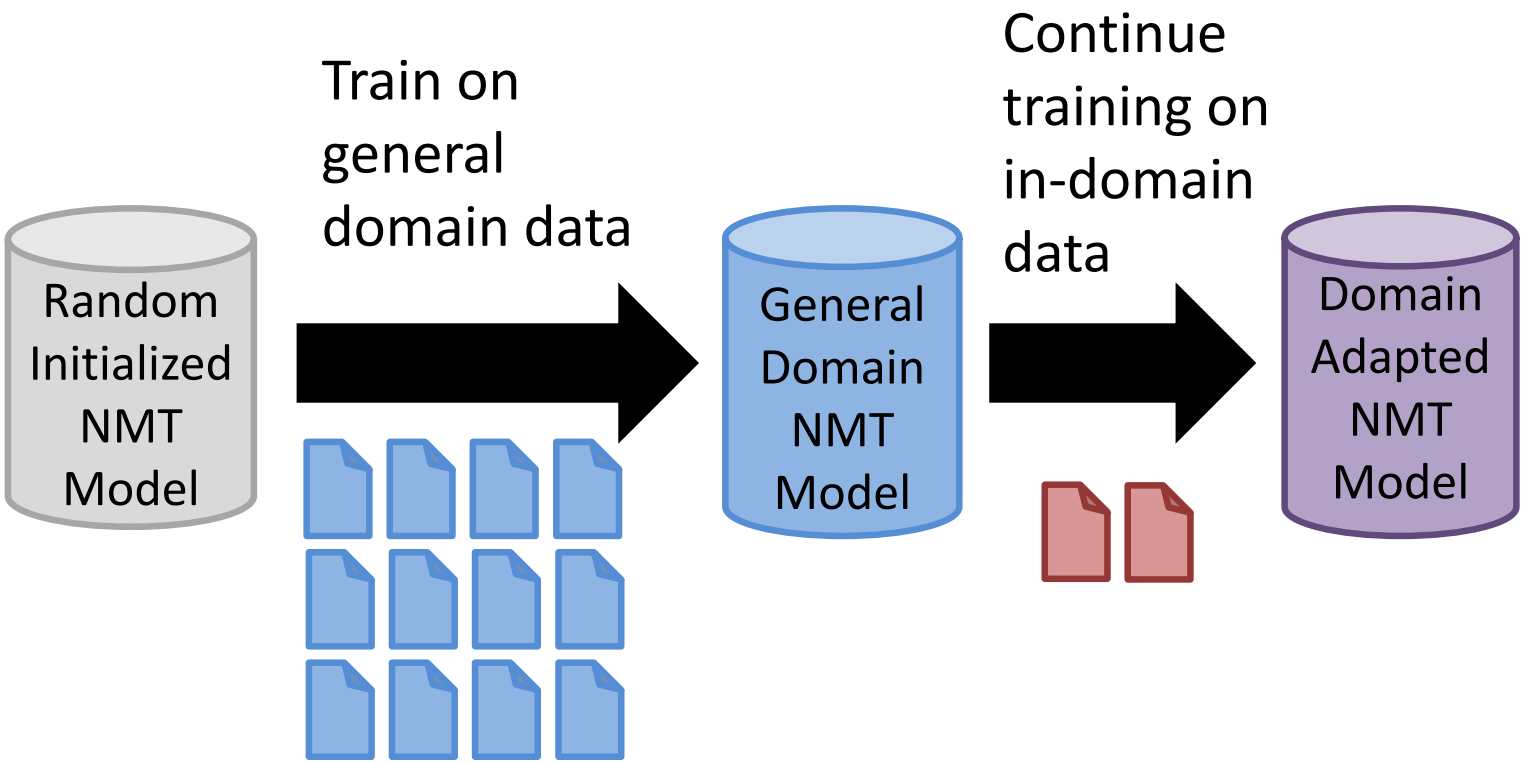
It is based on this paper:
http://www.aclweb.org/anthology/W18-6313
bib: http://aclweb.org/anthology/W18-6313.bib

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



Corpora

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)

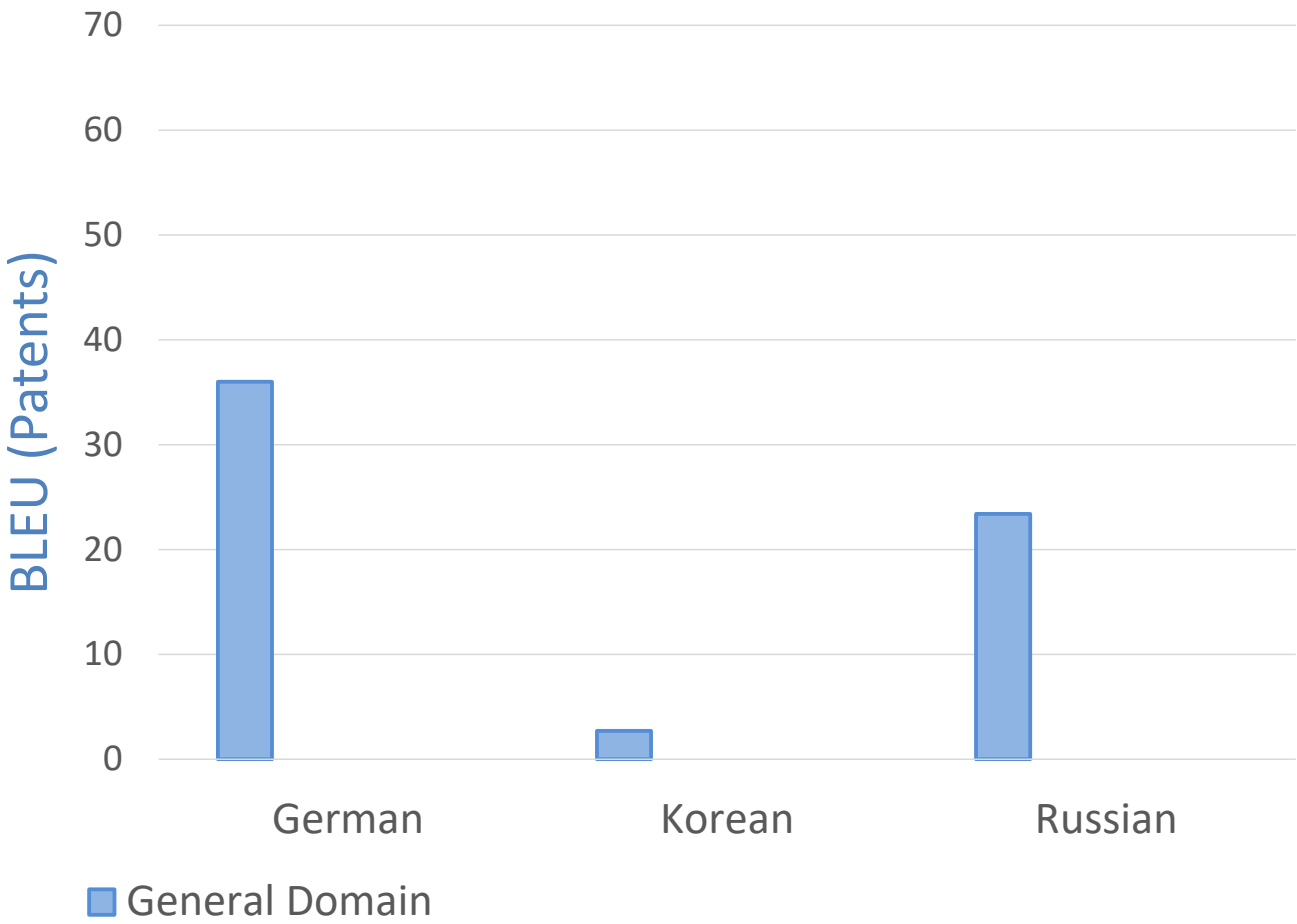
In-domain data: Patent abstracts from the World Intellectual Property Organization (WIPO)

Data Examples

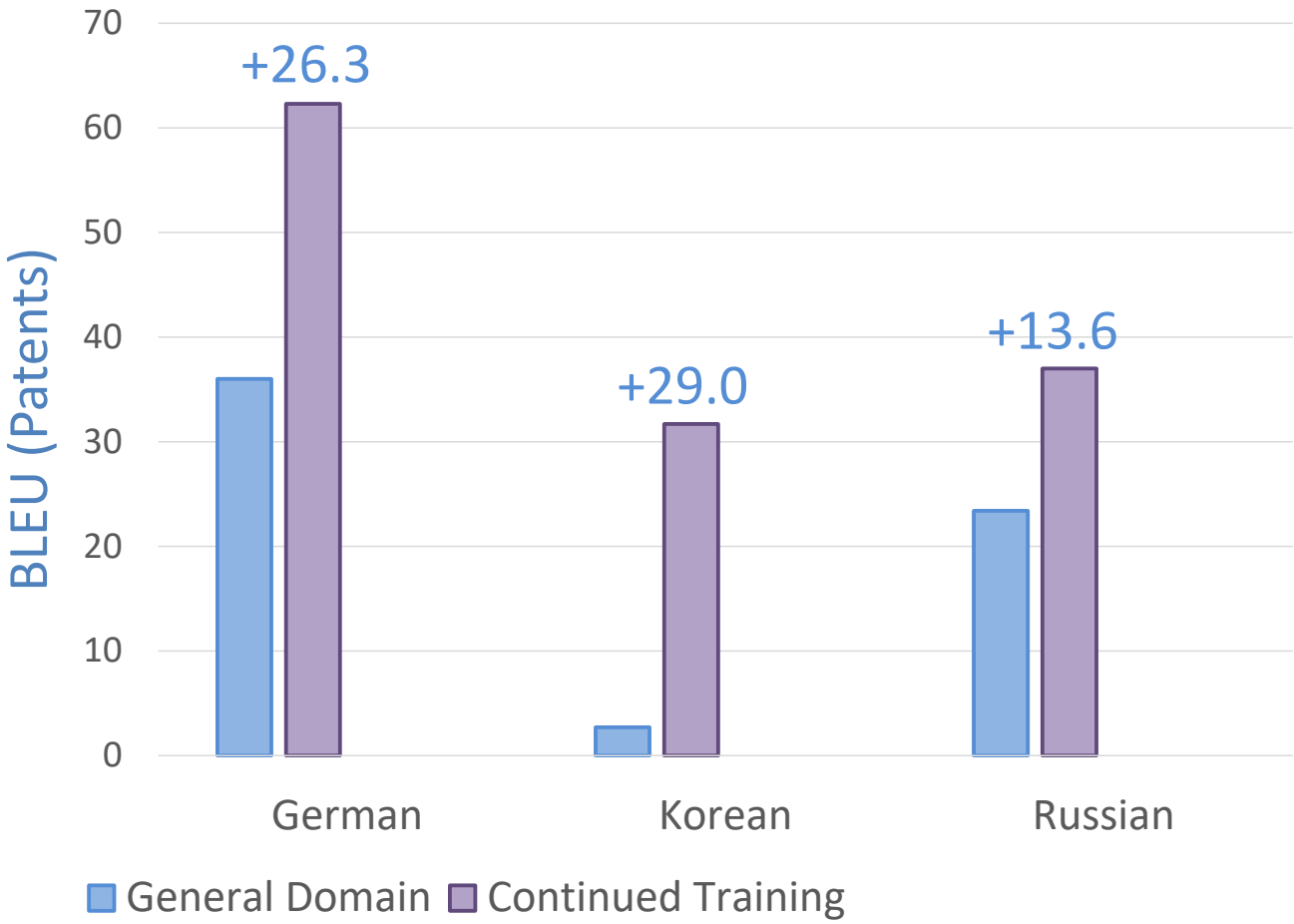
General-Domain:	
OpenSubtitles	You're gonna need a bigger boat.
WMT	Intensified communication and sharing of information between the project partners enables the transfer of expertise in rural tourism.

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.

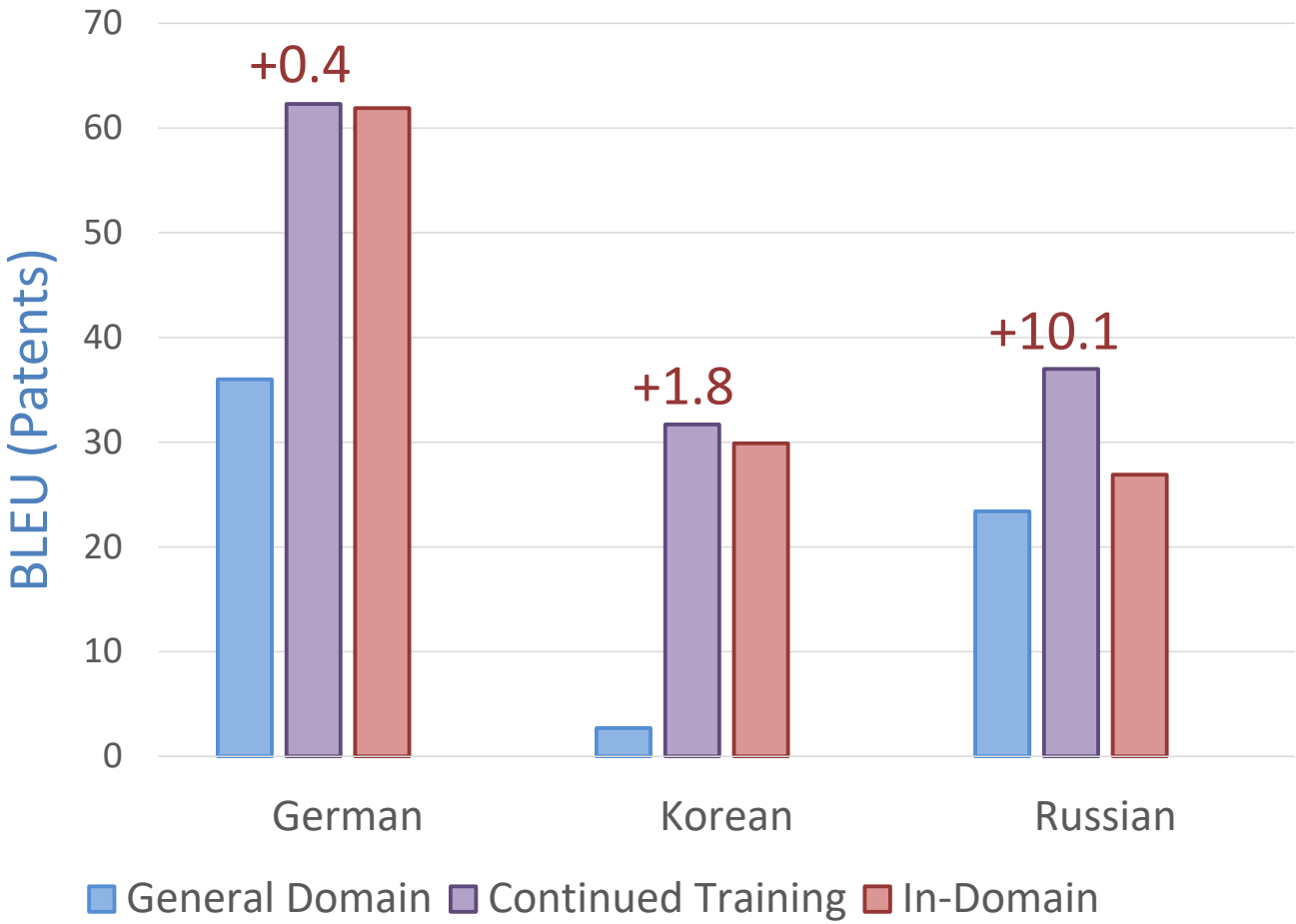
Models



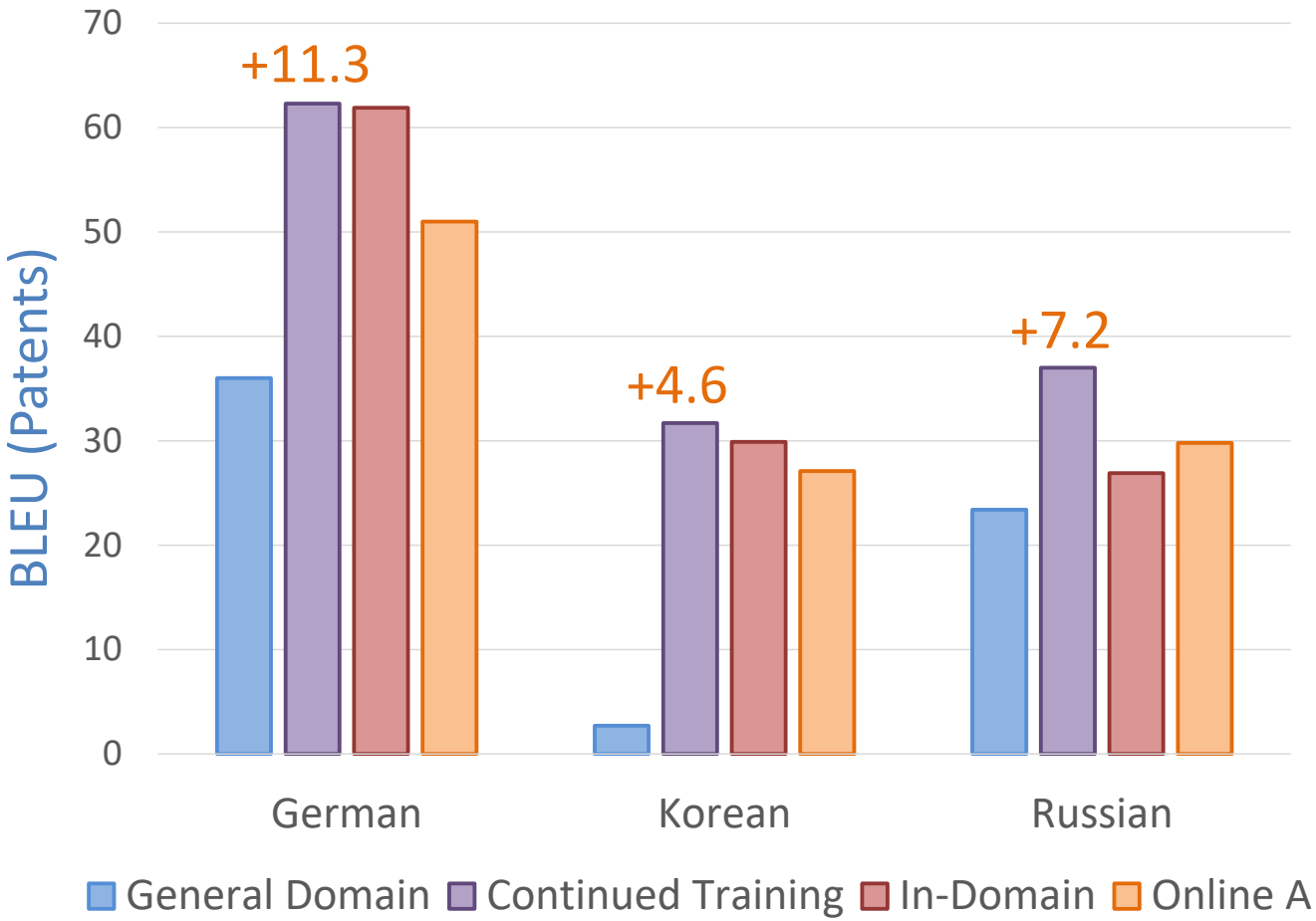
Models



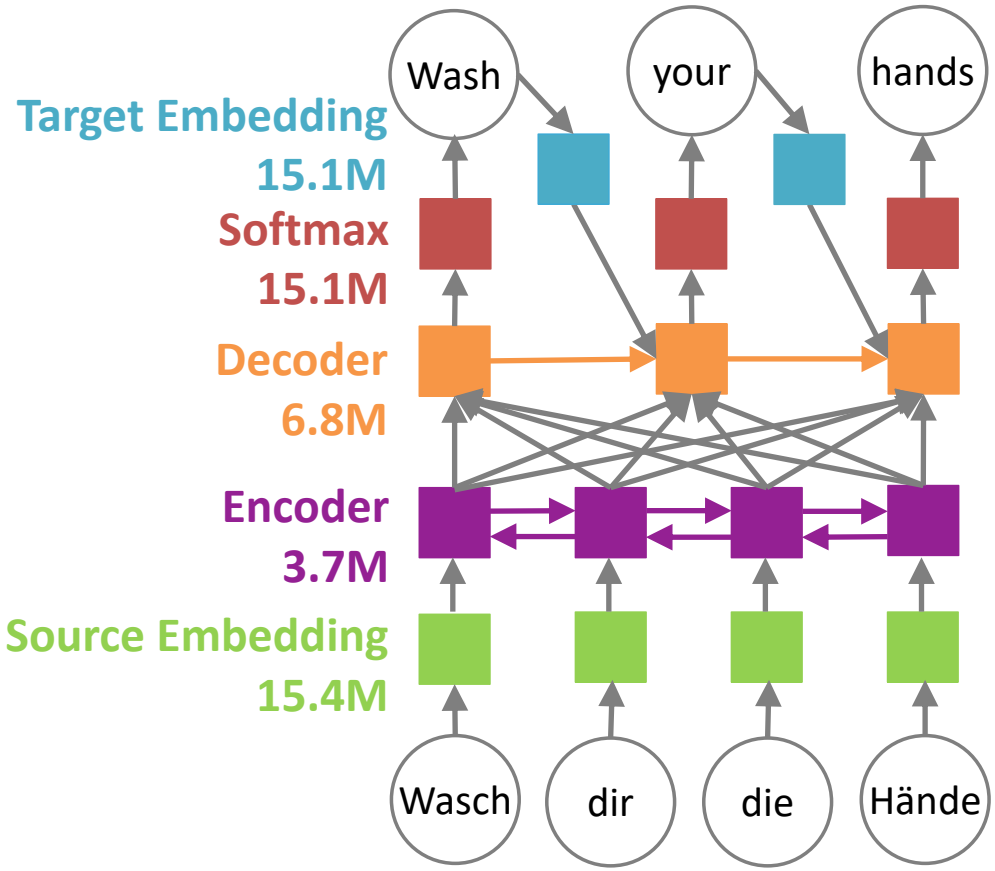
Models



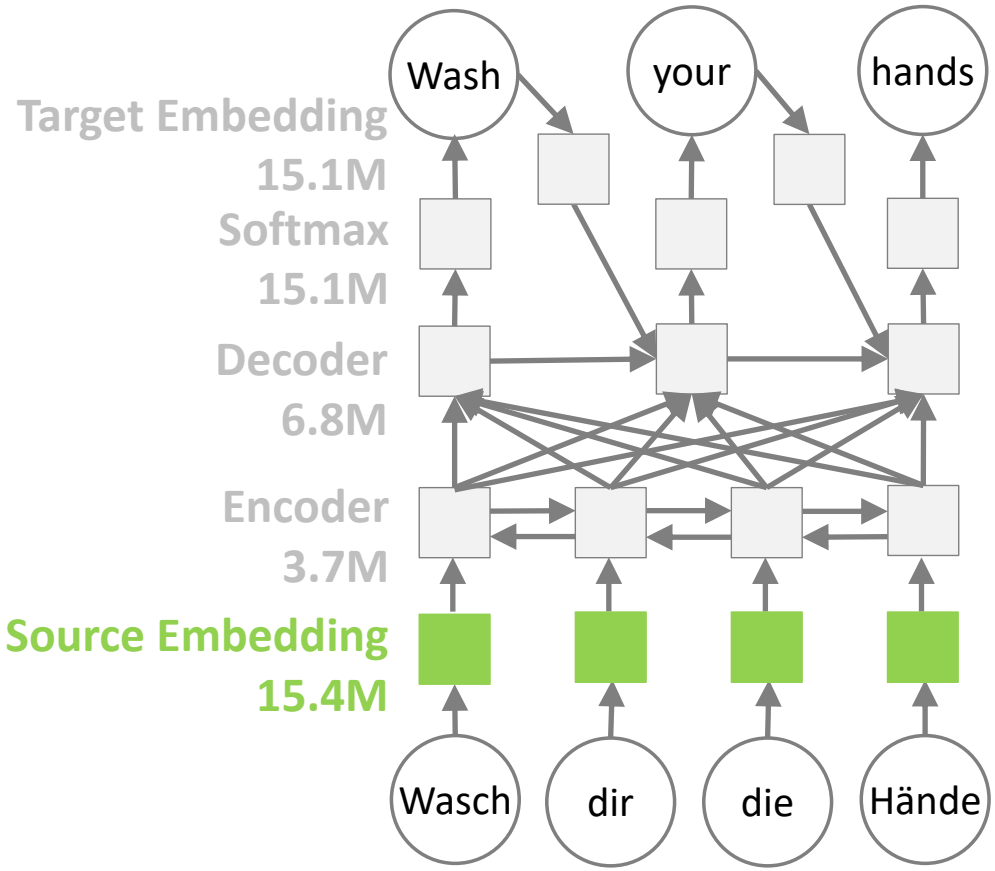
Models



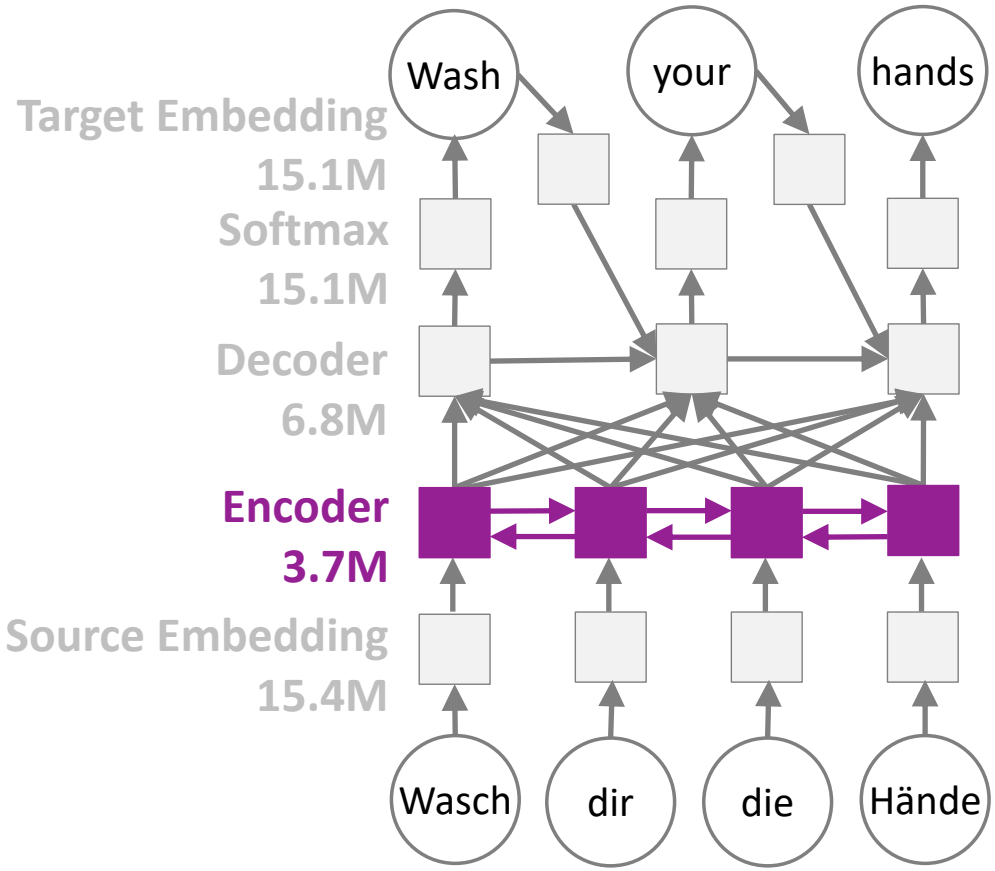
Subnetworks



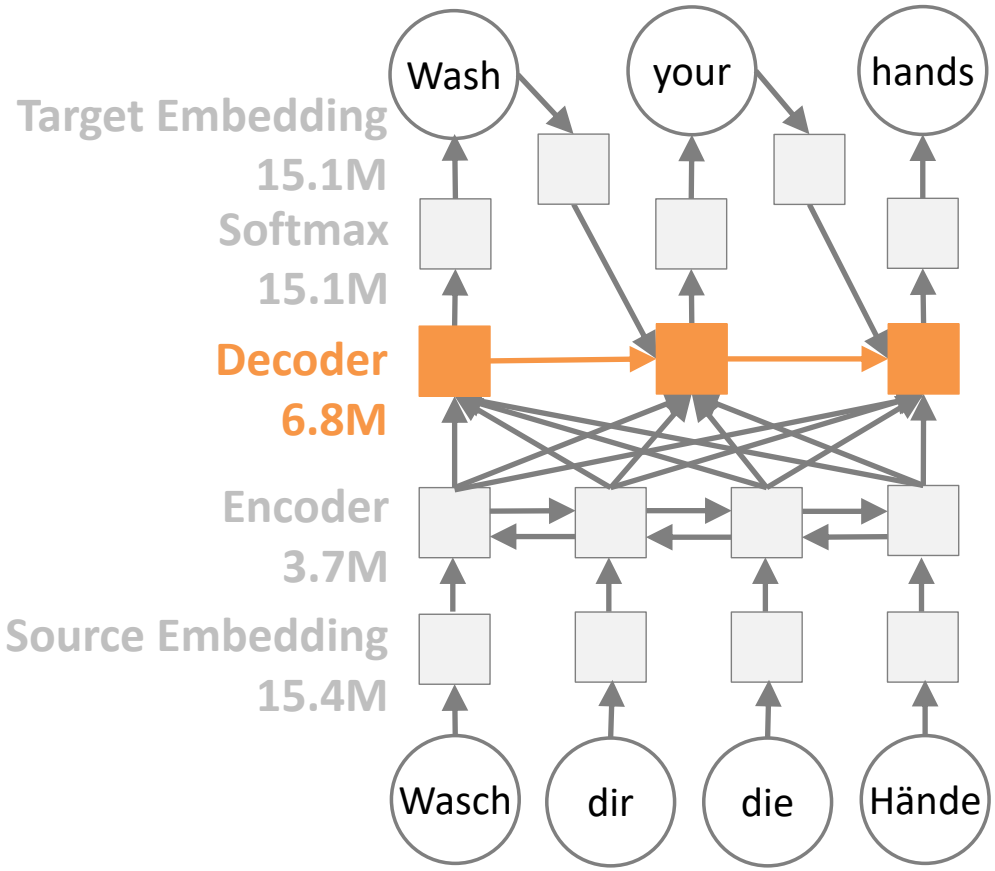
Subnetworks



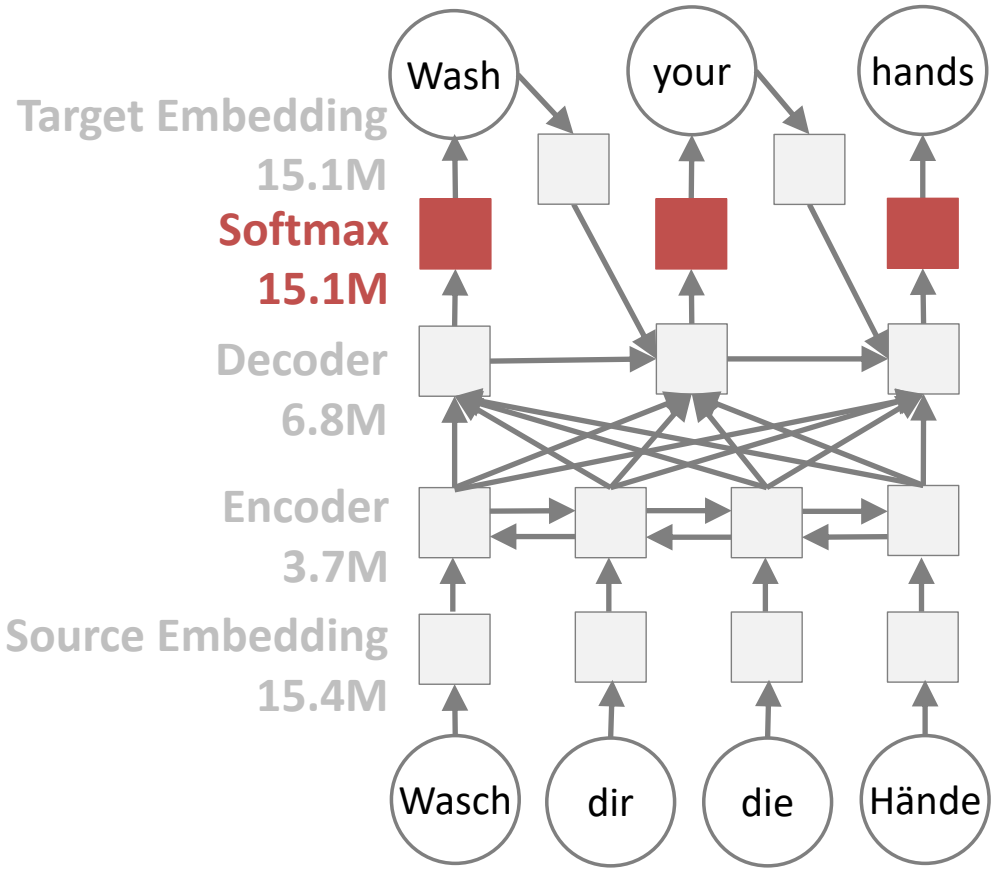
Subnetworks



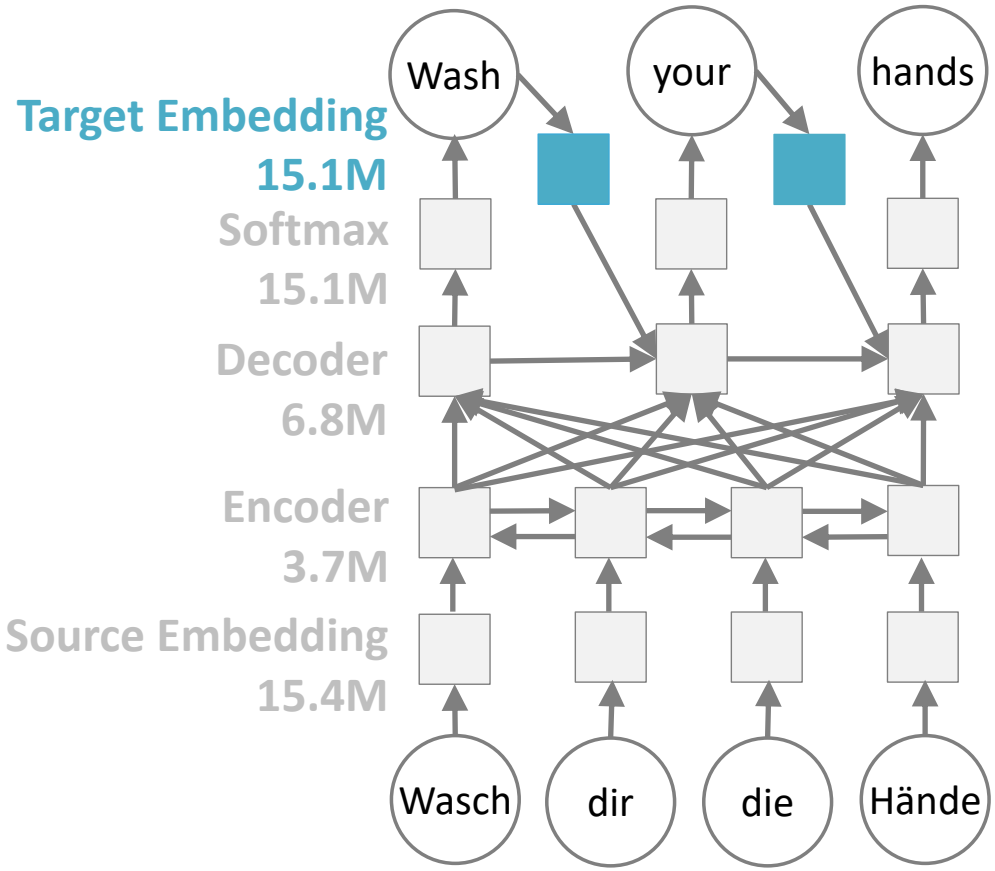
Subnetworks



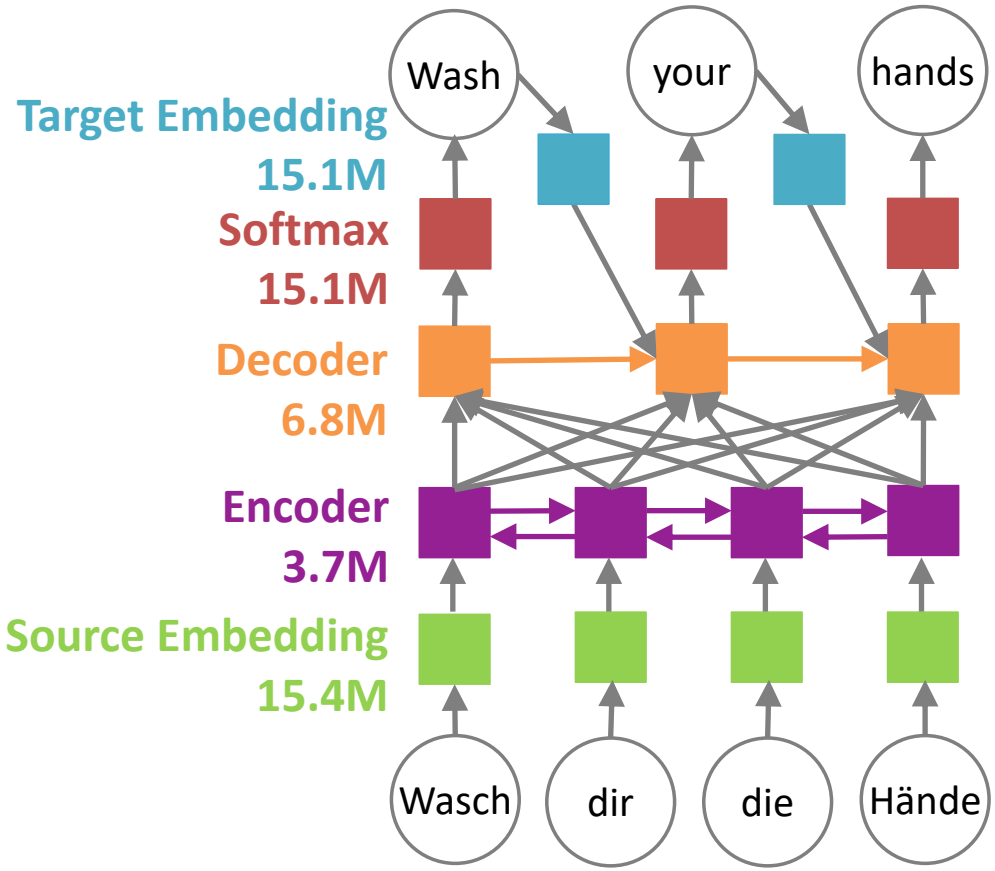
Subnetworks



Subnetworks

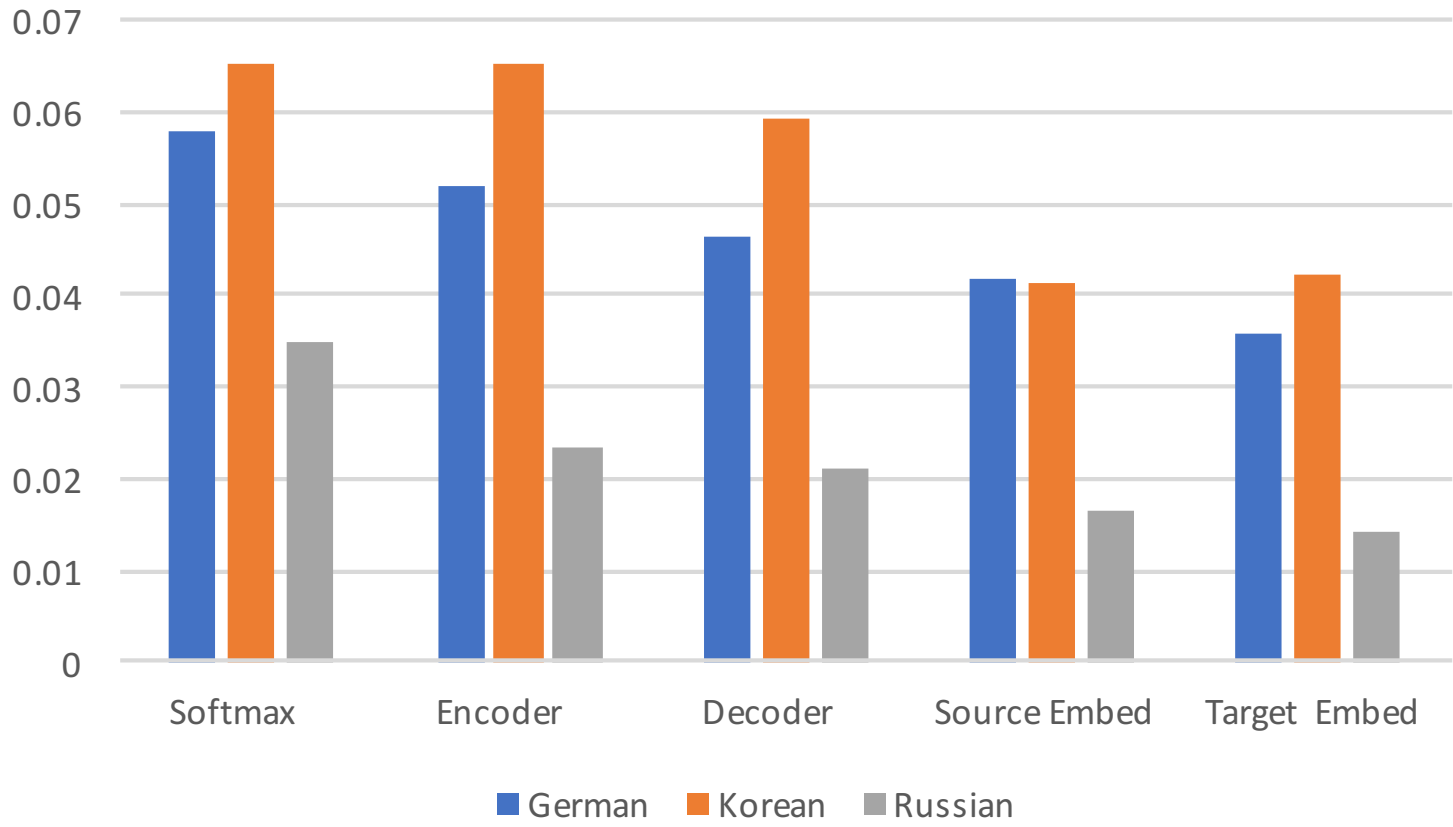


Subnetworks



Change During Adaptation

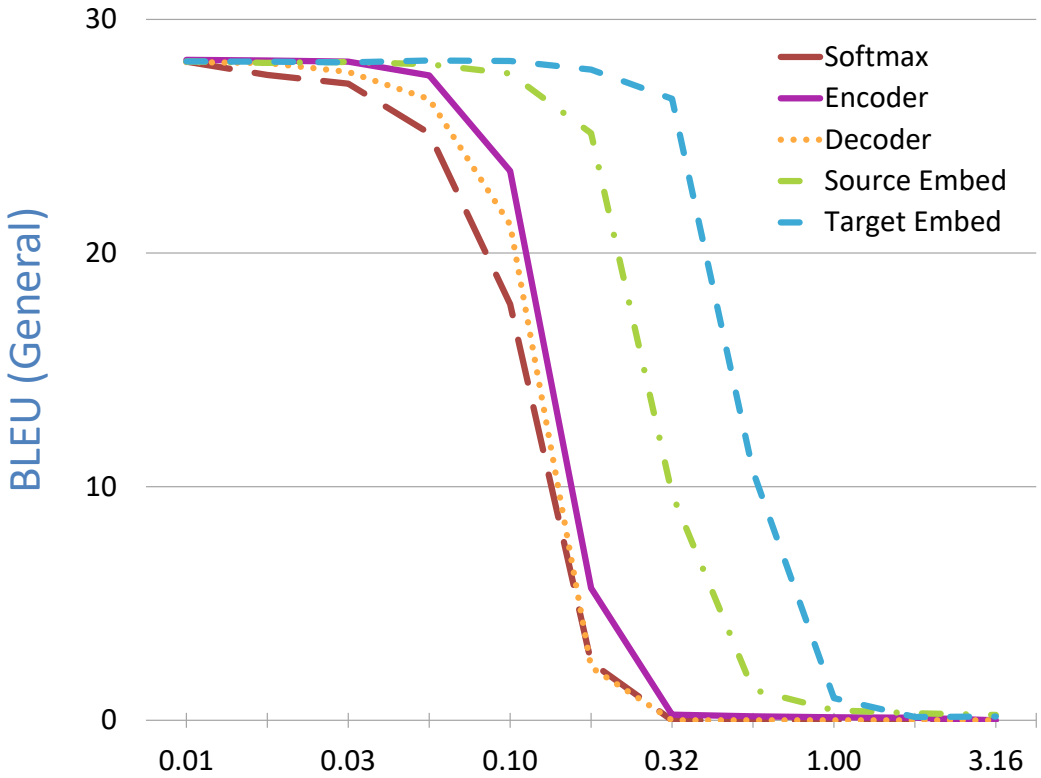
How much do parameters change during continued training?



(RMS Change)

Per-Component Sensitivity Analysis

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



(Russian)

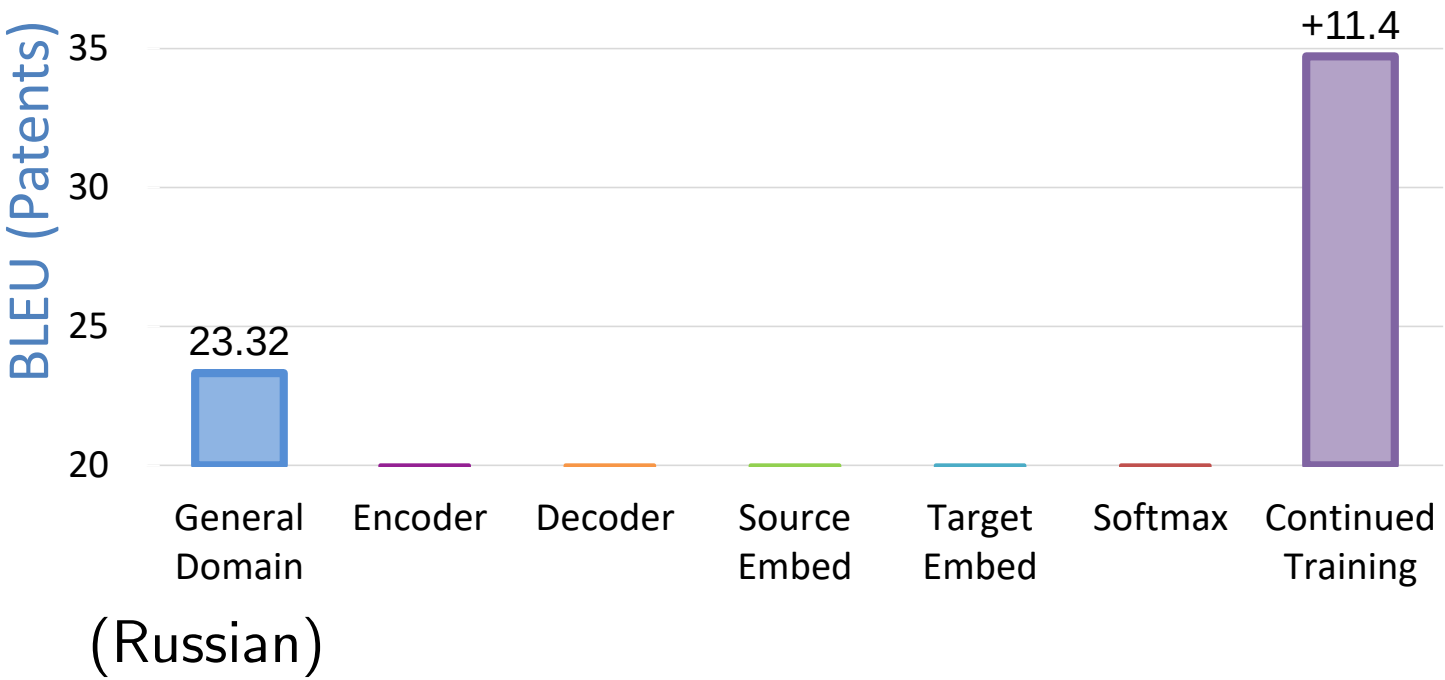
Component	L^2 Norm
Softmax	0.14
Encoder	0.22
Decoder	0.24
Src. Emb	0.20
Tgt. Emb	0.20

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

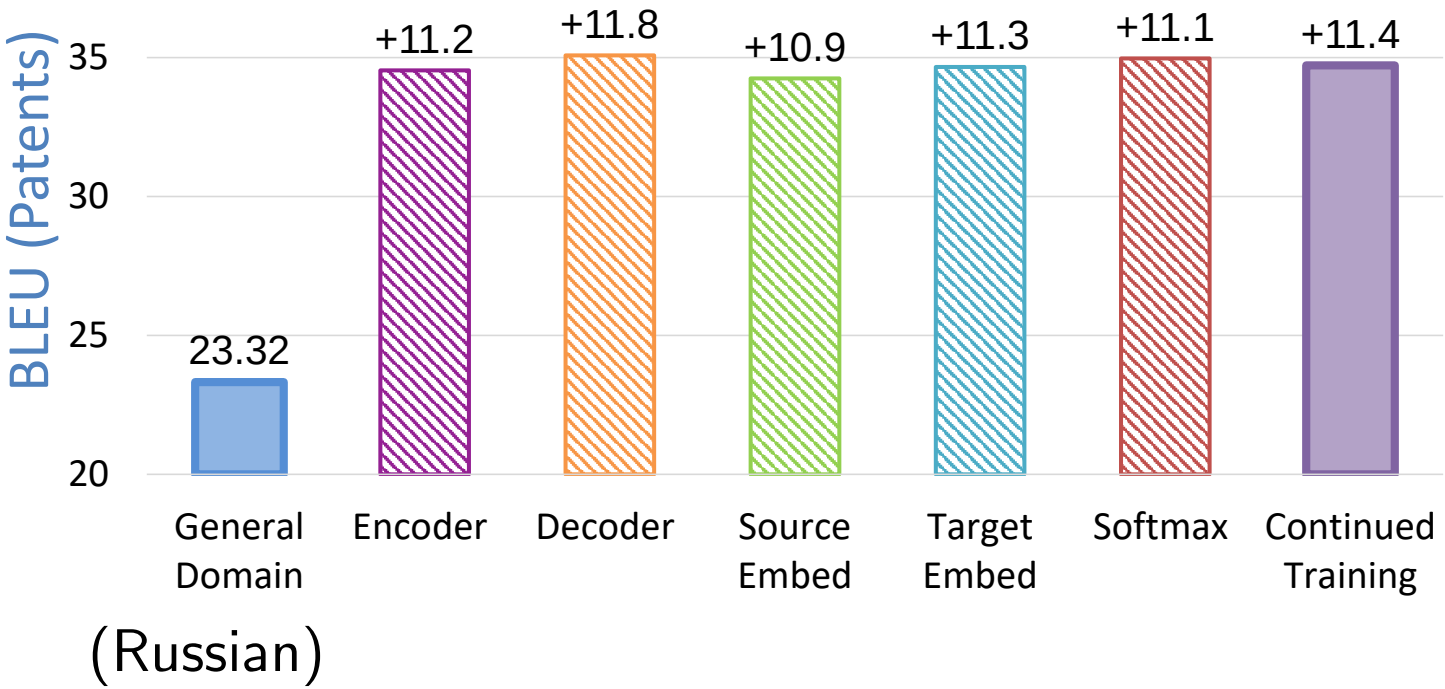
Question: How much does the model / training procedure depend on any **single** component for adaptation?



[1]When initial general-domain model is reasonably good

Freezing One Component at a Time

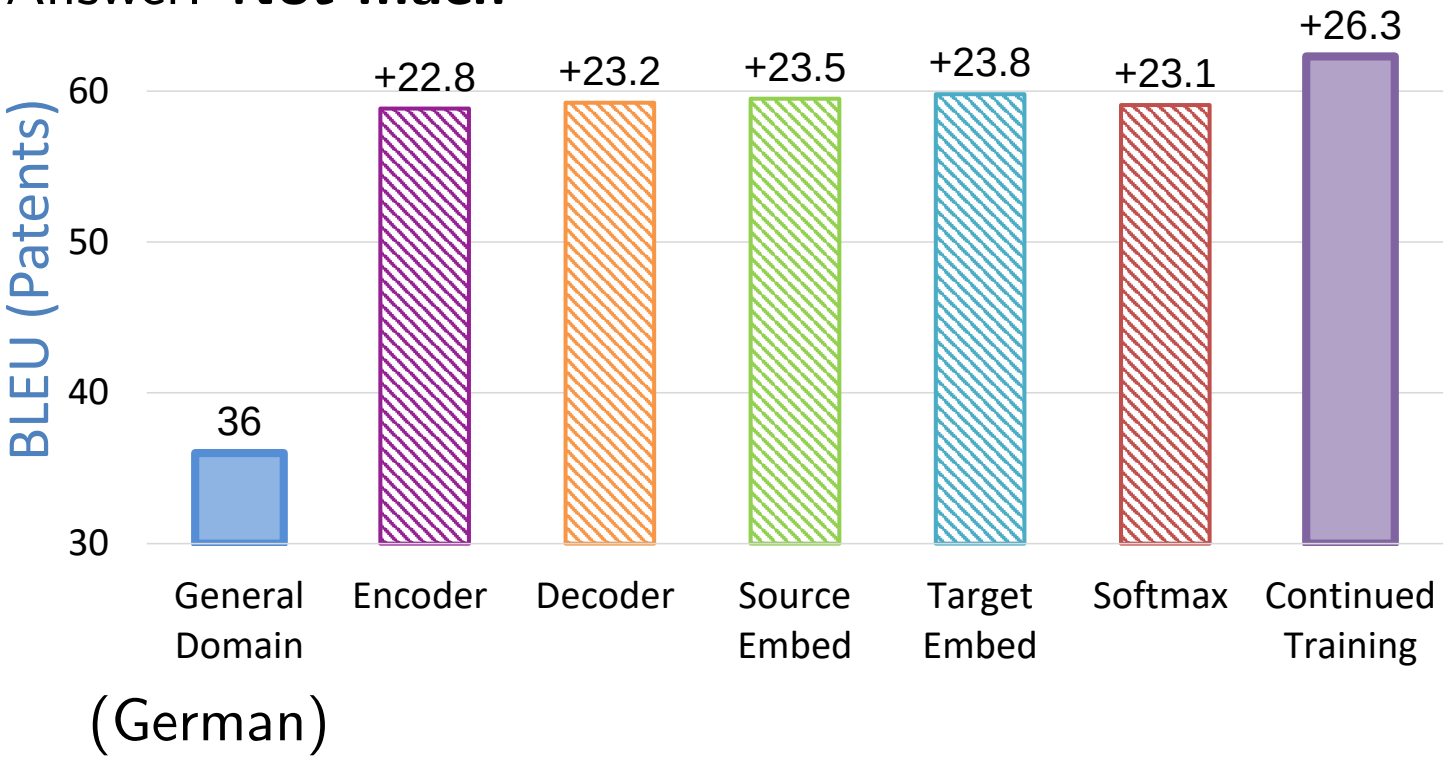
Question: How much does the model / training procedure depend on any **single** component for adaptation?
Answer: **Not much**[1]



[1]When initial general-domain model is reasonably good

Freezing One Component at a Time

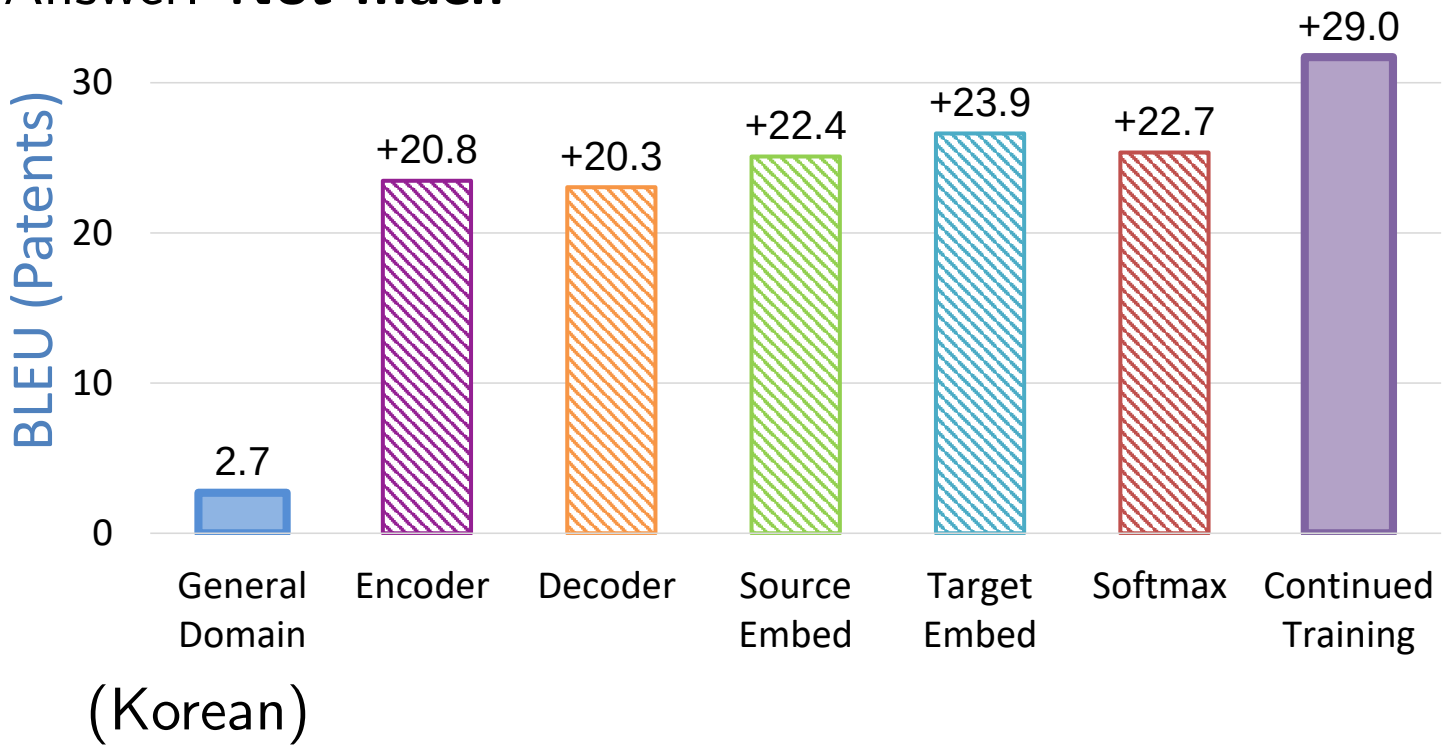
Question: How much does the model / training procedure depend on any **single** component for adaptation?
Answer: **Not much**[1]



[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]



(Korean)

[1]When initial general-domain model is reasonably good

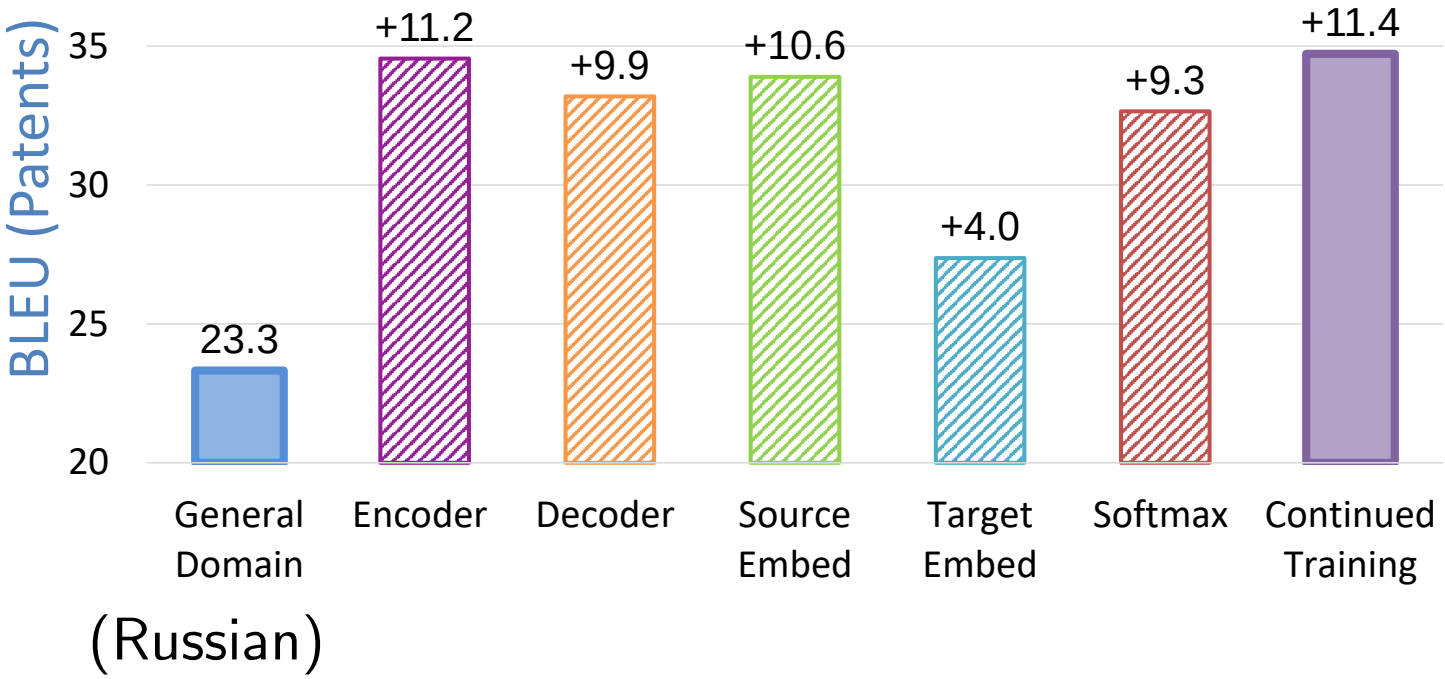
Freezing All But One Component at a Time

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

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]



(Russian)

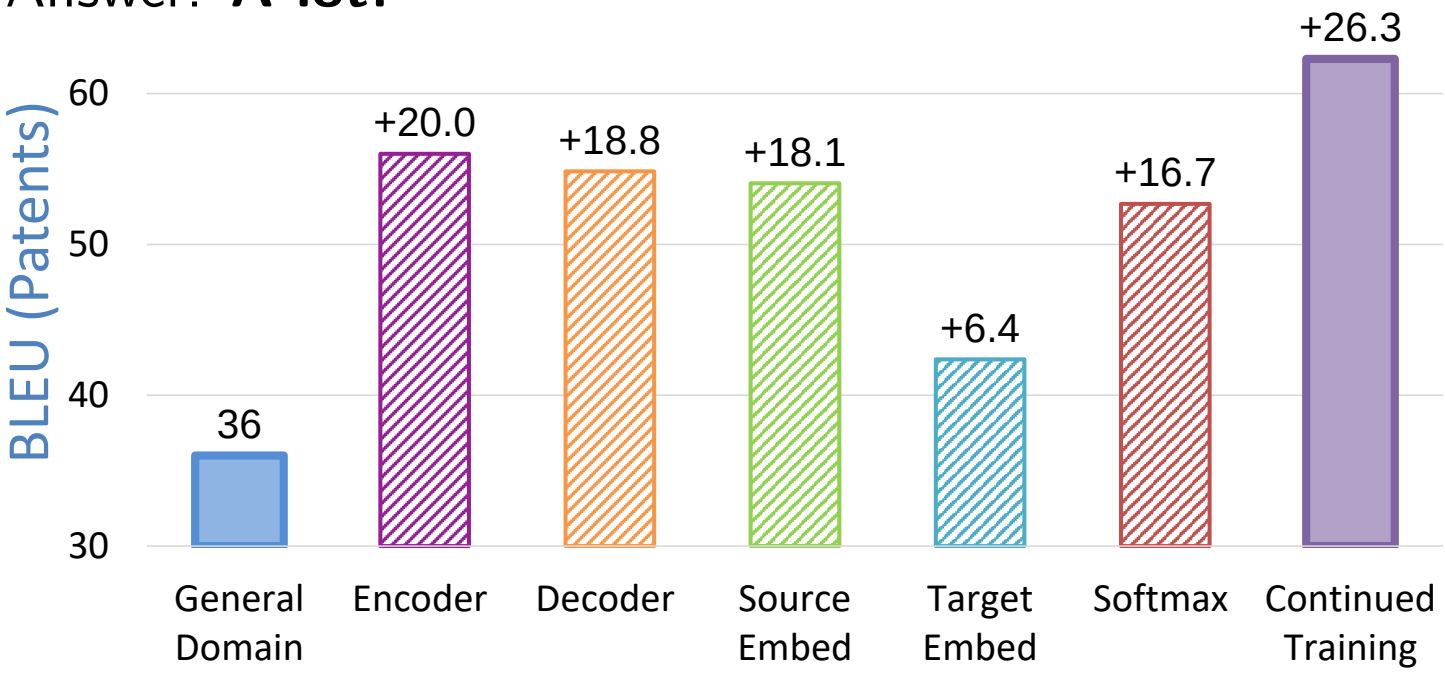
[1] When initial general-domain model is reasonably good

[2] Except for the target embeddings

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]



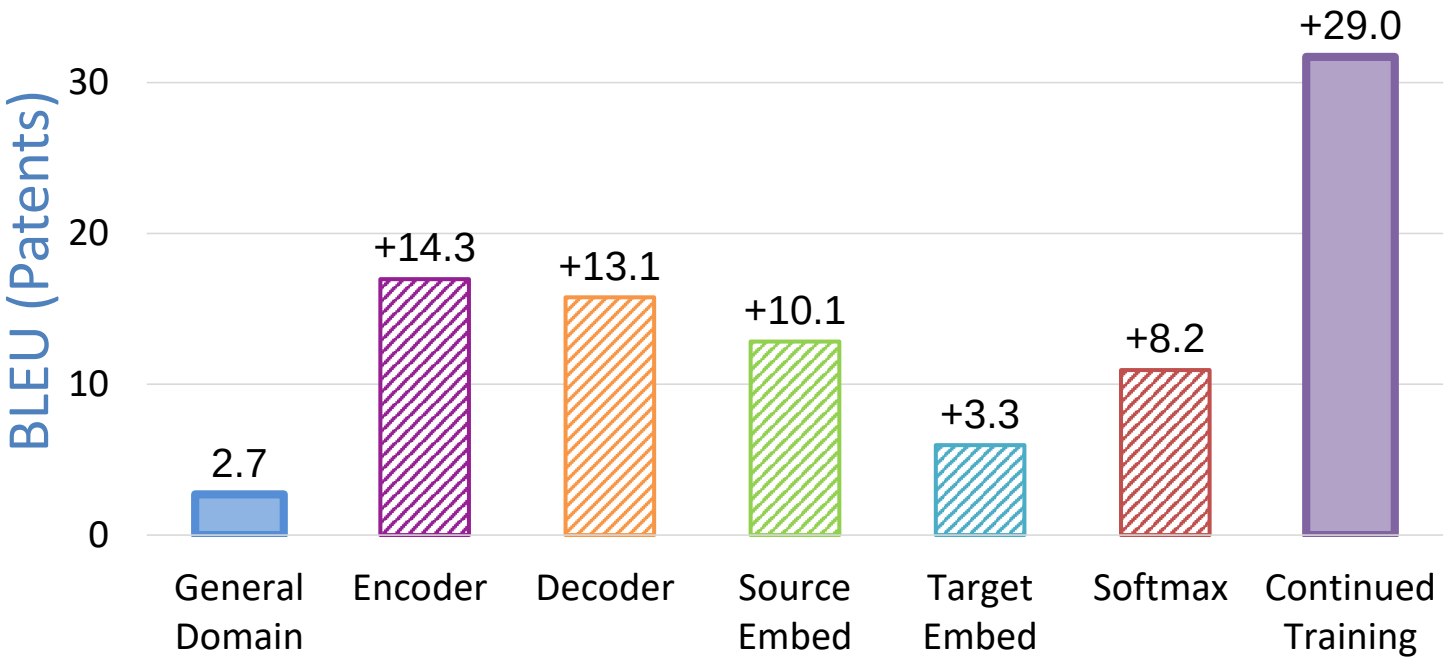
[1] When initial general-domain model is reasonably good

[2] Except for the target embeddings

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]



(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