Johns Hopkins University

Center for Language and Speech Processing

Por Qué Não Utiliser Alla Språk? Mixed Training with Gradient Optimization in Few-Shot Cross-Lingual Transfer

Haoran Xu, Kenton Murray

OverView

The current ubiquitous paradigm of few-shot cross-lingual transfer first trains on source language and fine-tunes with a few target shots (target-adapting).

We show some deficiencies of this approach and propose a one-step mixed training method that trains on both source and target data with stochastic gradient surgery, a novel gradient-level optimization.

Deficiencies of Target-Adatping

- **Deficiency 1: Unrealistic Development Set**
  Previous studies utilize a large amount of dev sets for each target language for model selection, e.g., even around 10K dev examples for Arabic in the NER task. However, it is unlikely that such a dev set would be available in reality, especially for the extreme low-resource training.

Solution 1:

ord-FS+dev*: ordinary Few-Shot method (target adapting)* ***with*** *unrealistically dev set.*

ord-FS*: ordinary Few-Shot method (target-adapting)* ***without*** *unrealistically dev set.*

- **Deficiency 2: One Model for Each Language**
  we do not need to fine-tune specialized models for every target language, which is of particular interest when scaling to dozens or even hundreds of languages.

Solution 2:

mix-FT: mixed fine-tuning on concatenated target examples together.

- **Deficiency 3: Language Domain Gap**
  Abruptly shifting the source domain to the target domain leads to very poor performance.
- **Deficiency 4: Quick Overfitting**
  the model performs best on the dev set at the beginning of training at a small number of shots, e.g. 1-shot, 5-shot.

Solution 3:

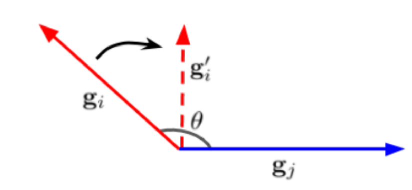
naïve-mix-train: naively training both source and all target examples together.

Scan me for more details!

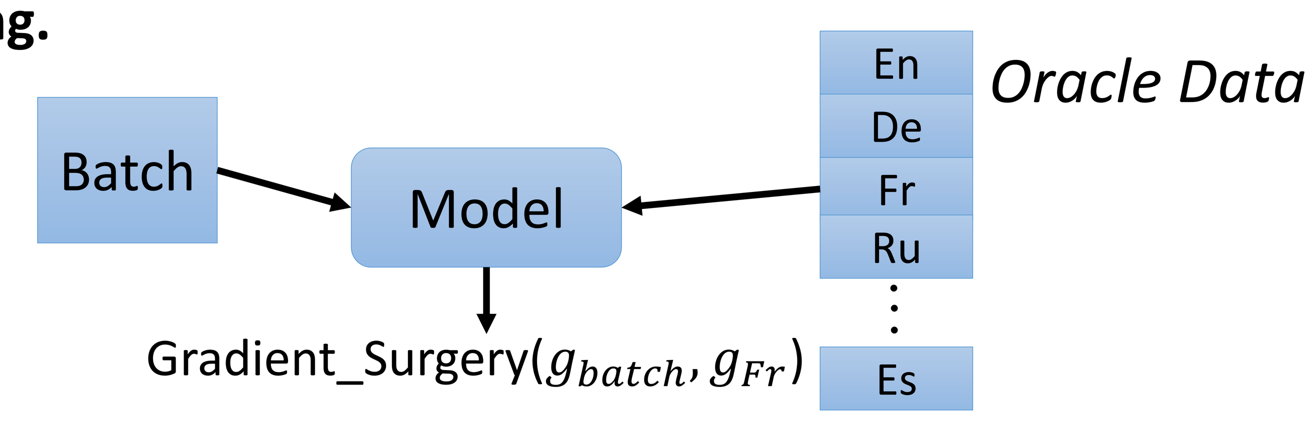
Mixed Training with Stochastic Gradient Surgery

One issue of **naive-mix-train** is conflicting gradients among languages. The main idea is using gradient surgery (Yu et al., 2020). However, it is extremely computationally expensive to de-conflict gradients between every pair of languages, especially when it comes to large-scale languages for training.

$$g_s' = g_s - \frac{g_s \cdot g_t}{\| g_t \|^2} g_t$$



gradient-mix-train: We randomly choose a target language to conduct gradient surgery **in each batch training.**



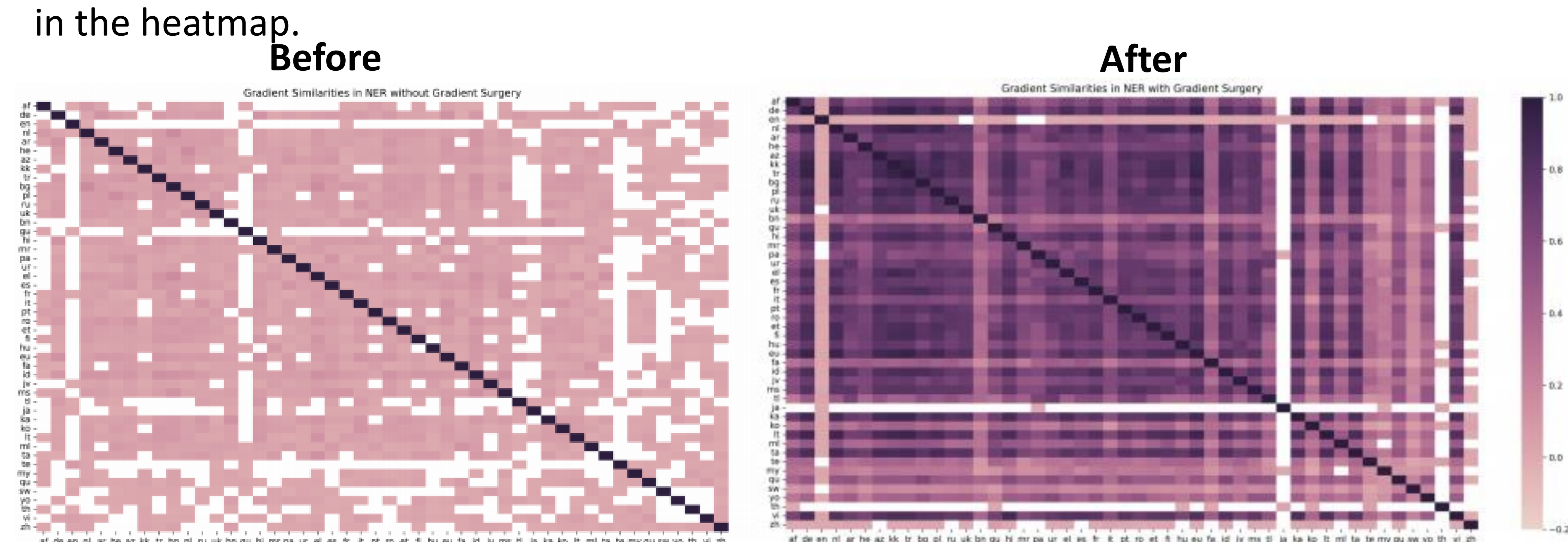
Main Results

We conduct experiments on 4 tasks, NER (48 langs), POS (35 langs), TyDiQA (9 langs), XNLI (15 langs). We repeat every experiment 5 times with 5 different random seeds.

K	Methods	NER Avg. F1 (%)	sd.
$K=0$	Zero-Shot	64.56	-
$K=1$	ord-FS+dev (Zhao et al., 2021)	65.92	0.84
	ord-FS (Zhao et al., 2021)	64.11	0.98
	mix-FT (Ours)	65.71	0.90
	naive-mix-train (Ours)	67.31	0.58
	gradient-mix-train (Ours)	**69.58**	0.99
$K=5$	ord-FS+dev (Zhao et al., 2021)	68.22	0.69
	ord-FS (Zhao et al., 2021)	65.91	0.91
	mix-FT (Ours)	70.60	0.85
	naive-mix-train (Ours)	72.06	0.68
	gradient-mix-train (Ours)	**73.27**	0.60
$K=10$	ord-FS+dev (Zhao et al., 2021)	69.85	0.60
	ord-FS (Zhao et al., 2021)	68.75	0.67
	mix-FT (Ours)	73.89	0.56
	naive-mix-train (Ours)	74.13	0.45
	gradient-mix-train (Ours)	**75.92**	0.61

Analysis

Visualization of Gradient De-Conflicting: Gradient similarities across 48 languages in the NER task with 5 shots before and after **Stochastic Gradient Surgery.** Deeper colors represent higher cosine similarities. Conflicting gradients are directly marked as white cells in the heatmap.



Which Language Benefits Most?

We retrieve Top-5 languages that achieve the highest improvement by using gradient-mix-train methods compared to **ord-FS** on all tasks in 5-shot learning.

NER		POS		TyDiQA		XNLI	
lang.	Δ F1 (%)	lang.	Δ F1 (%)	lang.	Δ F1 (%)	lang.	Δ Acc. (%)
pa	17.60	wo	3.82	bn	12.27	sw	2.36
zh	15.24	mr	3.51	te	11.14	ur	1.95
ar	14.14	hi	2.60	sw	10.58	ru	1.68
vi	13.22	tr	2.18	ar	9.45	fr	0.91
hi	12.68	fi	1.55	fi	9.05	zh	0.78