Zero-Shot Cross-Lingual Transfer with Meta-Learning

Farhad Nooralahzadeh[1]

Giannis Bekoulis[2] Johannes Bjerva[3] Isabelle Augenstein[3]

[1]University of Oslo

[2]Vrije Universiteit Brussel - imec

[3]University of Copenhagen

EMNLP 2020

Motivation

- Most of Machine Learning Methods (e.g. **Deep Neural networks**) need to have a **large** training set
- **Low-Resource** vs. **High-Resource** Tasks/Domains/Languages
- **Transfer Learning**
- **Strategic sharing** of knowledge has been shown to improve **downstream NLP** task performance

Meta-Learning

- Meta-Learning, or **learning to learn**, tackles the problem of **fast adaptation** on **new** and **few** training data
- Learns **structure** among multiple tasks, learning **new tasks** is **fast**
- **Repeatedly simulating** the learning process on **low-resource** domains/languages using many **high-resource** ones (Gu et al. 2018)

Multi-Task vs. Meta-Learning in NLP

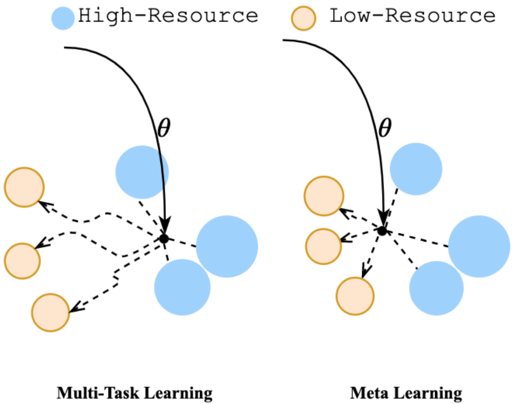


Figure is adapted from (Dou, K. Yu and Anastasopoulos 2019)

MAML: Model-Agnostic Meta-Learning (Finn et al. 2017)

- Learns a **good parameter** initialization for **fast adaptation** with only **small amount** of data and with a **few gradient** steps
- **Model**- and **Task Agnostic** (Any model trained with gradient descent)
- Can be applied in various setting: **Classification**, **Regression**, **Reinforcement Learning**

MAML in NLP

- Meta-Learning for Low-Resource **Neural Machine Translation** (Gu et al. 2018)
- Domain Adaptive **Dialog Generation** via Meta Learning (Qian and Z. Yu 2019)
- Model-Agnostic Meta-Learning for **Relation Classification** with Limited Supervision (Obamuyide and Vlachos 2019)

Method: MAML for Cross-Lingual NLU tasks

- Gu et al. (2018) exploit a set of **high-resource auxiliary languages** to improve the performance of a low-resource one in **Machine Translation** task
- **Cross-lingual setting in NLU** (`e.g., Natural Language Inference, Question Answering`)
- Only **English** dataset as a **high-resource** language and other languages are in a **low-resource** mode
- We introduce a `cross-lingual meta-learning` framework (**X-MAML**)

Method: Cross-lingual Meta-learning (X-MAML)

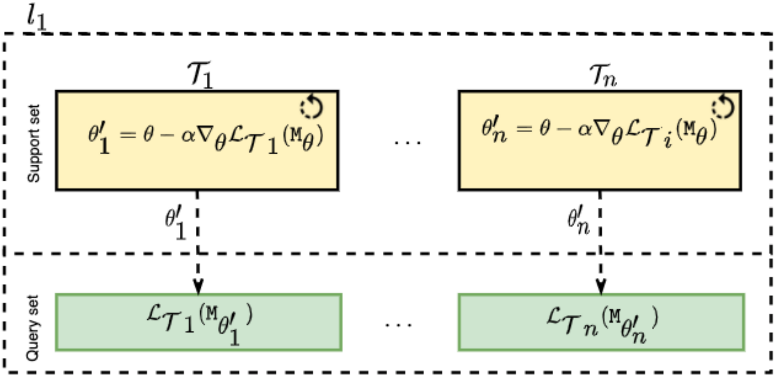
1. **Pre-training** stage on the high-resource language (i.e, English)
2. Meta-learning using **one or two low-resource languages** as auxiliary tasks
3. **Zero-shot** learning or **Fine-tuning** on the target languages.

Method: Cross-lingual Meta-learning (X-MAML)

h: High-resource language
L: Set of low-resource languages
M: Model pre-trained on h
$\mathcal{T}_i$: A batch from Development set of language i in L
α, β: Step size
θ: Initial parameter
θ'_i: Optimal parameter in $\mathcal{T}_i$, fast weight

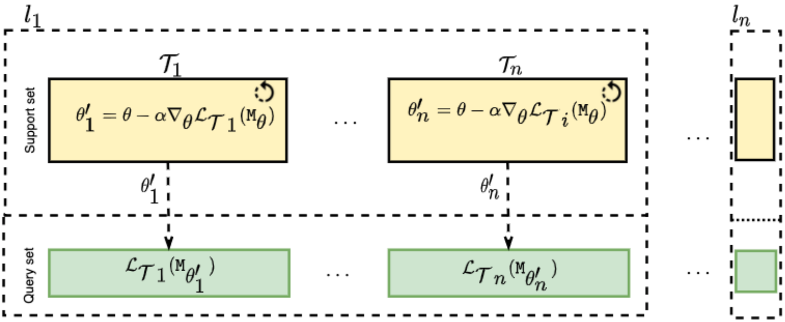
Method: Cross-lingual Meta-learning (X-MAML)

h: High-resource language
L: Set of low-resource languages
M: Model pre-trained on h
$\mathcal{T}_i$: A batch from Development set of language i in L
α, β: Step size
θ: Initial parameter
θ'_i: Optimal parameter in $\mathcal{T}_i$, fast weight



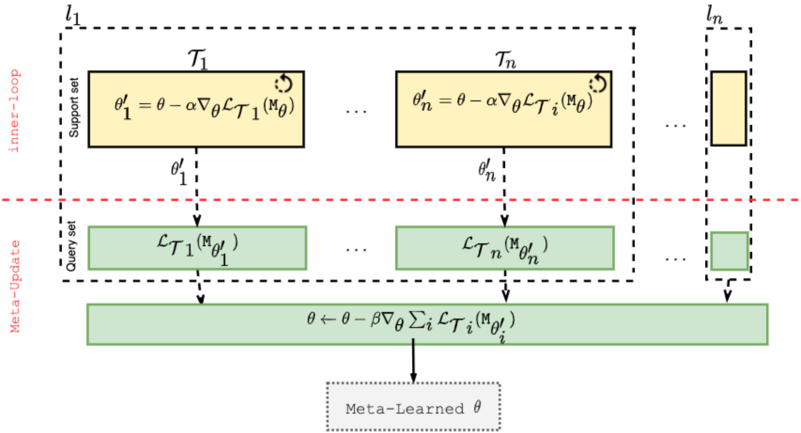
Method: Cross-lingual Meta-learning (X-MAML)

h: High-resource language
L: Set of low-resource languages
M: Model pre-trained on h
$\mathcal{T}_i$: A batch from Development set of language i in L
α, β: Step size
θ: Initial parameter
θ'_i: Optimal parameter in $\mathcal{T}_i$, fast weight



Method: Cross-lingual Meta-learning (X-MAML)

h: High-resource language
L: Set of low-resource languages
M: Model pre-trained on h
$\mathcal{T}_i$: A batch from Development set of language i in L
α, β: Step size
θ: Initial parameter
θ'_i: Optimal parameter in $\mathcal{T}_i$, fast weight



Method: Cross-lingual Meta-learning (X-MAML)

Algorithm 1: X-MAML

Input: high-resource language $\mathtt{h}$, set of low-resource languages $\mathtt{L}$,
Model $\mathtt{M}$, step size α and learning rate β

Pre-train $\mathtt{M}$ on $\mathtt{h}$ and provide initial model parameters θ

Select one or more languages from $\mathtt{L}$ as a set of auxiliary languages ($\mathtt{A}$)

while *not done* **do**

  for $\mathtt{l} \in \mathtt{A}$ **do**

    Sample batches of tasks $\mathcal{T}_i$ using development set of the auxiliary language l

    for *each* $\mathcal{T}_i$ **do**

      Sample k data-points to form $D_i^{train} = \{X^j, Y^j\} \in \mathcal{T}_i$

      Sample q data-points to form $D_i^{test} = \{X^j, Y^j\} \in \mathcal{T}_i$ for meta-update

      Compute $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta)$ on D^{train}

      Compute adapted parameters with gradient descent: $\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_\theta)$

      Compute $\mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta'})$ using D_i^{test}

  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_{\mathcal{T}_i}(\mathtt{M}_{\theta'})$

Perform either (i) zero-shot or (ii) few-shot learning on $\{\mathtt{L} \setminus \mathtt{A}\}$ using meta-learned parameters θ

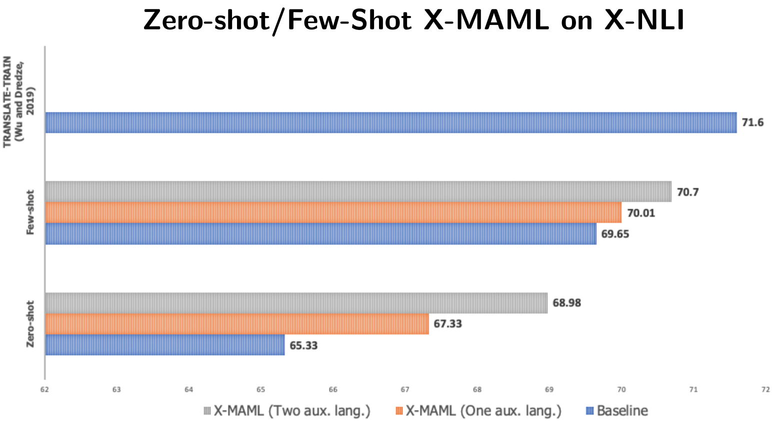
Experiments: Cross-Lingual NLI

- **XNLI**: Cross-Lingual NLI (Conneau et al. 2018)
  - An extension of the **SNLI/MultiNLI** corpus in **15** languages
  - **Train**: 392,702 pairs in **English**
  - The pairs are annotated with textual entailment and **translated** into 14 languages
  - **2,500** dev, **5,000** test pairs for 15 Languages

Experiments: Cross-Lingual NLI

- **Zero-shot/Few-Shot X-MAML on X-NLI**:
  - M: **Multi-BERT**
  - **Baseline**: Multi-BERT is trained on English
    - **Zero-shot**: Evaluated on the **test set** of each target language
    - **Few-shot**: Fine-tuned on **dev set** (2.5k) of target languages, then is **evaluated** on the test set
  - Apply X-MAML with **one or two auxiliary languages**
  - Report an average of **10** runs of X-MAML on XNLI

Experiments: Cross-Lingual NLI



- It **boosts** Multi-BERT performance on **cross-lingual NLI**
- **Alleviates** the **machine translating** step from the foreign language into English in the Multi-BERT setting.

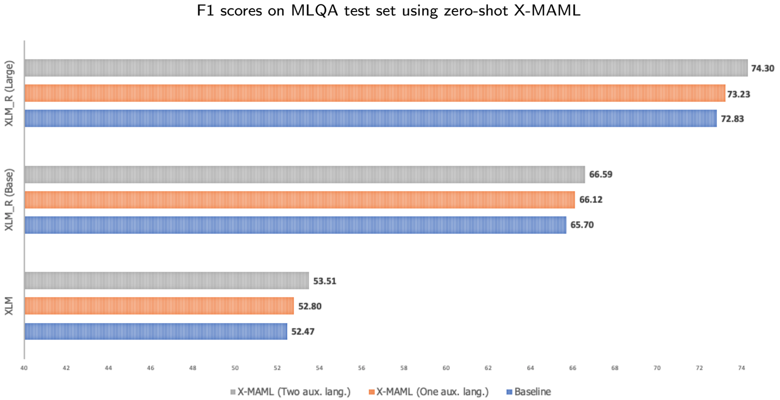
Experiments: Multilingual QA

MLQA: Multilingual Question Answering dataset (Lewis et al. 2019)

- **MLQA** Contains QA instances in **7 languages**: English (en), Arabic (ar), German(de), Spanish (es), Hindi (hi), Vietnamese (vi) and Simplified Chinese (zh).
- It includes over **12k QA** instances in English and **5k** for every other language
- This dataset has been used in many recent studies on **cross-lingual transfer learning** (e.g., Hu et al. 2020; Liang et al. 2020).

Experiments: Multilingual QA

Zero-Shot X-MAML on MLQA



- Overall, **zero-shot** learning models with X-MAML **outperform** the baselines
- Improvement: **+1.04%**(XLM), **+0.89%** (XLM-R_{base}) and **+1.47%** (XLM-R_{large}) in average F_1

Discussion and Analysis

- **Cross-lingual transfer** with meta-learning yields improved results even when languages strongly differ from one another
- **Zero-shot X-MAML on XNLI**, improved transfer performance is achieved for Russian (ru) $\rightarrow$ Hindi (hi)

Discussion and Analysis

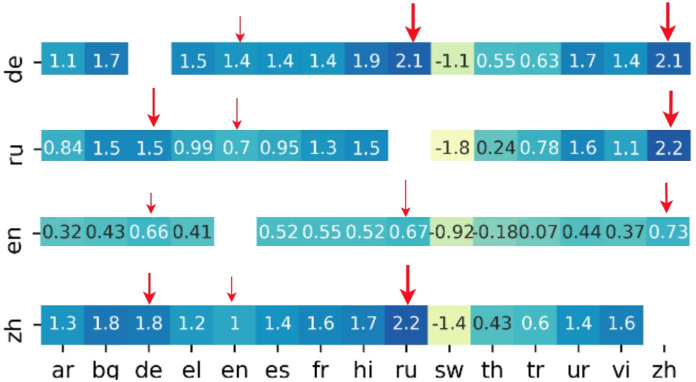
- **WALS**: World Atlas of Language Structure (Dryer and Haspelmath, 2013)
  - **Largest openly** available **typological database**
  - $\sim$ **200 linguistic features** (i.e, phonological, grammatical, lexical properties) with annotations for more than **2,500 languages**
- Investigate whether two languages **sharing** the **same typological feature** is beneficial for performance using **X-MAML**
- Languages with **similar morphosyntactic** properties can be **beneficial** to one another in X-MAML

Analysis: Typological Correlations (Zero-shot X-MAML)

25A Locus of Marking: Whole-language Typology

"whether the morphosyntactic marking in a language is on the syntactic heads or dependents of a phrase."

en,de,ru,zh: Dependent-marking



Conclusion

- We propose **X-MAML**, a cross-lingual meta-learning architecture, and study it for two natural language understanding tasks (Natural Language Inference and Question Answering)
- We test X-MAML on cross-lingual few-shot as well as zero-shot learning, across a total of 15 languages
- We observe consistent improvements over strong models including Multilingual BERT and XLM-RoBERTa
- Languages which **share certain morphosyntactic** features tend to benefit from this type of **transfer**

Thanks.

farhad.nooralahzadeh@uzh.ch