Is Modularity Transferable? A Case Study through the Lens of Knowledge Distillation

Mateusz Klimaszewski[1] Piotr Andruszkiewicz[1]
Alexandra Birch[2]

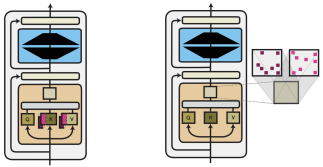
[1]Warsaw University of Technology
[2]University of Edinburgh

LREC-COLING 2024

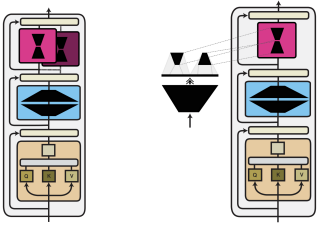
Modular Deep Learning

Modular Deep Learning advantages:

- positive transfer
- compositionality
- parameter efficiency



(a) Input Composition (b) Param. Composition



(c) Function Composition (d) Hypernetwork

Source: Pfeiffer et al., 2023. Modular Deep Learning.

Modularity properties

Modularity properties:

- autonomous
  - single responsibility
    (e.g. a specific task/language)
- parameter-efficient
  - cheaper than full fine-tuning
  - fraction of model parameters
- attached to a base model

Source: adapterhub.ml

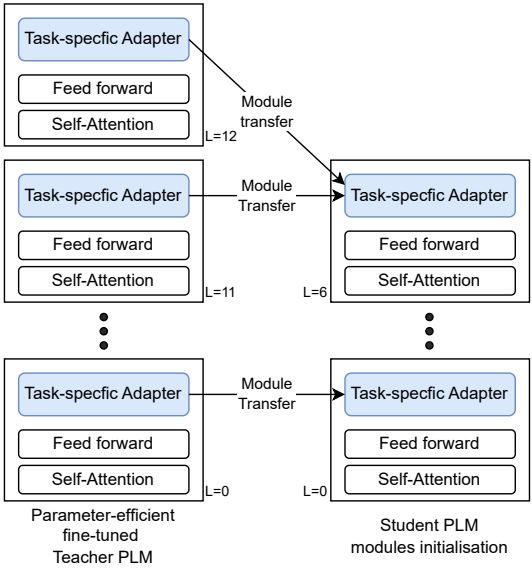
Modularity properties

Modularity properties:

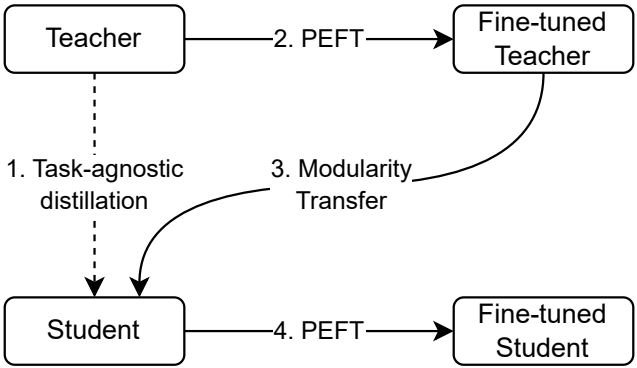
- autonomous
  - single responsibility
    (e.g. a specific task/language)
- parameter-efficient
  - cheaper than full fine-tuning
  - fraction of model parameters
- ~~attached to a base model~~
- **transferable (?)**

Source: adapterhub.ml

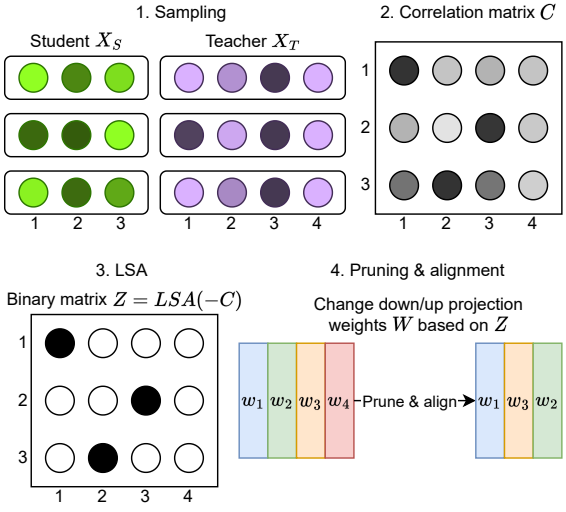
Case study - overview



Experimental schema



Incompatible PLMs - pruning & alignment



Experiment setup

Modular methods:

- Adapter (Pfeiffer et al., 2021)
- LoRA (Hu et al., 2022)

Tasks:

- Named Entity Recognition (NER, WikiNeural - Tedeschi et al., 2021)
- Natural Language Inference (NLI, XNLI - Conneau et al., 2020)
- Paraphrase Identification (PI, PAWS-X - Yang et al., 2019)

Experiment setup - models

Table: Parameters, layer count and hidden dimension size of the evaluated models.

Model	Params	Layers	Hidden dim
D'mBERT	$135M$	6	768
mBERT	$178M$	12	768
XLM-R_{BASE}	$278M$	12	768
XLM-R_{LARGE}	$560M$	24	1024

As the models have mismatched layer counts, we test two approaches: skip modules (denoted SKIP, i.e. transfer every second module) or average them (AVG).

Matching PLMs experiment

	NER (F1)		PI (Acc)		NLI (Acc)	
	AVG	REL	AVG	REL	AVG	REL
			Adapter			
Teacher	95,35		82,60		67,98	
Student	92,94	−2,41	71,32	−11,28	62,12	−5,86
TM-Student$_{\text{AVG}}$	93,02	−2,32	72,96	−9,64	62,33	−5,65
TM-Student$_{\text{SKIP}}$	93,45	−1,90	75,11	−7,49	63,01	−4,97
			LoRA			
Teacher	93,27		74,68		63,00	
Student	90,09	−3,18	65,80	−8,88	60,56	−2,43
TM-Student$_{\text{AVG}}$	90,63	−2,64	68,52	−6,16	60,53	−2,47
TM-Student$_{\text{SKIP}}$	90,80	−2,47	70,69	−3,99	60,52	−2,47

Incompatible PLMs experiment

	NER (F1)		PI (Acc)	
	AVG	REL	AVG	REL
		Adapter		
Teacher	95,34		88,81	
Student	93,30	−2,04	84,12	−4,69
TM-Student$_{SKIP}$	93,34	−2,00	84,27	−4,54
		LoRA		
Teacher	93,64		87,03	
Student	90,83	−2,82	78,72	−8,31
TM-Student$_{SKIP}$	90,84	−2,80	78,64	−8,39

Conclusions

- We present a case study of transferable modularity property.
- We evaluate current modular techniques in two scenarios: *matching* and *incompatible* PLMs.
- The results show that for *matching* PLMs, the modularity transfer provides gains with current MDL approaches, while *incompatible* PLMs might require more robust alignment techniques.