MoNMT: Modularly Leveraging Monolingual and Bilingual Knowledge for Neural Machine Translation

Jianhui Pang, Baosong Yang, Derek F. Wong, et al.

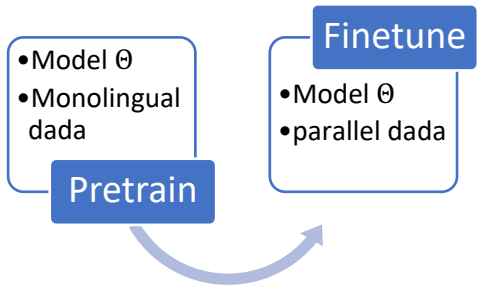
University of Macau

Open-Source: https://github.com/pangjh3/MoNMT

Background

Catastrophic Forgetting:

- Pretrain-and-finetune (PF) update the same model parameters.



- Background:
  - Pretrained monolingual knowledge can improve translation capabilities.
- Problems:
  - Traditional pretrain-and-finetune methods face catastrophic forgetting issues.
- Risk:
  - Finetuning may erase pretrained knowledge.
- Result:
  - Translation models don't fully utilize monolingual knowledge.

Motivation

- Motivation:
  - Make better use of monolingual data.
  - Imporve model generalization and robustness.
  - Synergize monolingual and bilingual knowledge to further enhance translation ability.

- Research Questions:
  - How to avoid the catastrophic forgetting problem.
  - How to balance the monolingual and bilingual knowledge in a translation model.

Methods

- Translation Formula:
  - ✓ Translation by tranferring the latent variables between tow languages, z_x and z_y.
  - ✓ The translation model is seperated into **three function**:
    - Encoding Module
    - Decoding Module
    - Transferring Module

1. The joint distribution:

$$\begin{aligned} & p(x, y, z_x, z_y) \\ & = p(y|z_y, z_x, x)p(z_y|z_x, x)p(z_x|x)p(x) \\ & \propto \underbrace{p(z_x|x)}_{\text{encode}} \underbrace{p(z_y|z_x)}_{\text{transfer}} \underbrace{p(y|z_y)}_{\text{decode}}, \end{aligned} \tag{1}$$

2. Denoise Auto-Encoding:

$$\begin{aligned} p(x, \hat{x}, z_x) &= p(x|z_x, \hat{x})p(z_x|\hat{x})p(\hat{x}) \\ &\propto \underbrace{p(z_x|\hat{x})}_{\text{encode}} \underbrace{p(x|z_x)}_{\text{decode}}, \end{aligned} \tag{2}$$

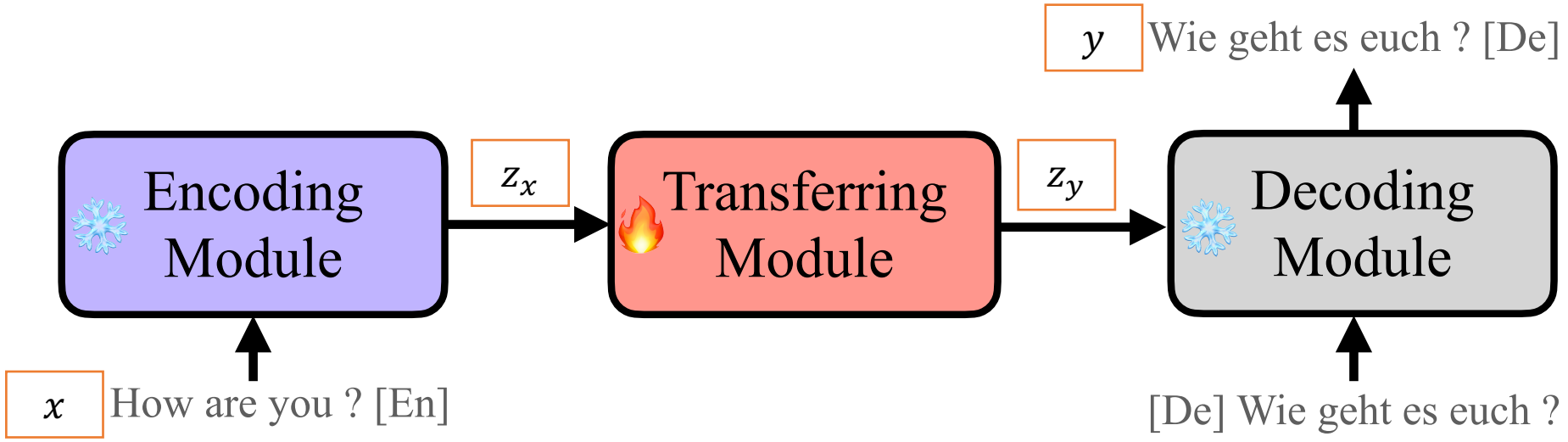
$$\begin{aligned} p(y, \hat{y}, z_y) &= p(y|z_y, \hat{y})p(z_y|\hat{y})p(\hat{y}) \\ &\propto \underbrace{p(z_y|\hat{y})}_{\text{encode}} \underbrace{p(y|z_y)}_{\text{decode}}, \end{aligned} \tag{3}$$

3. The joint distribution:

$$p(x, y, z_x, z_y) \propto \underbrace{p(z_y|z_x)}_{\text{transfer}}, \tag{4}$$

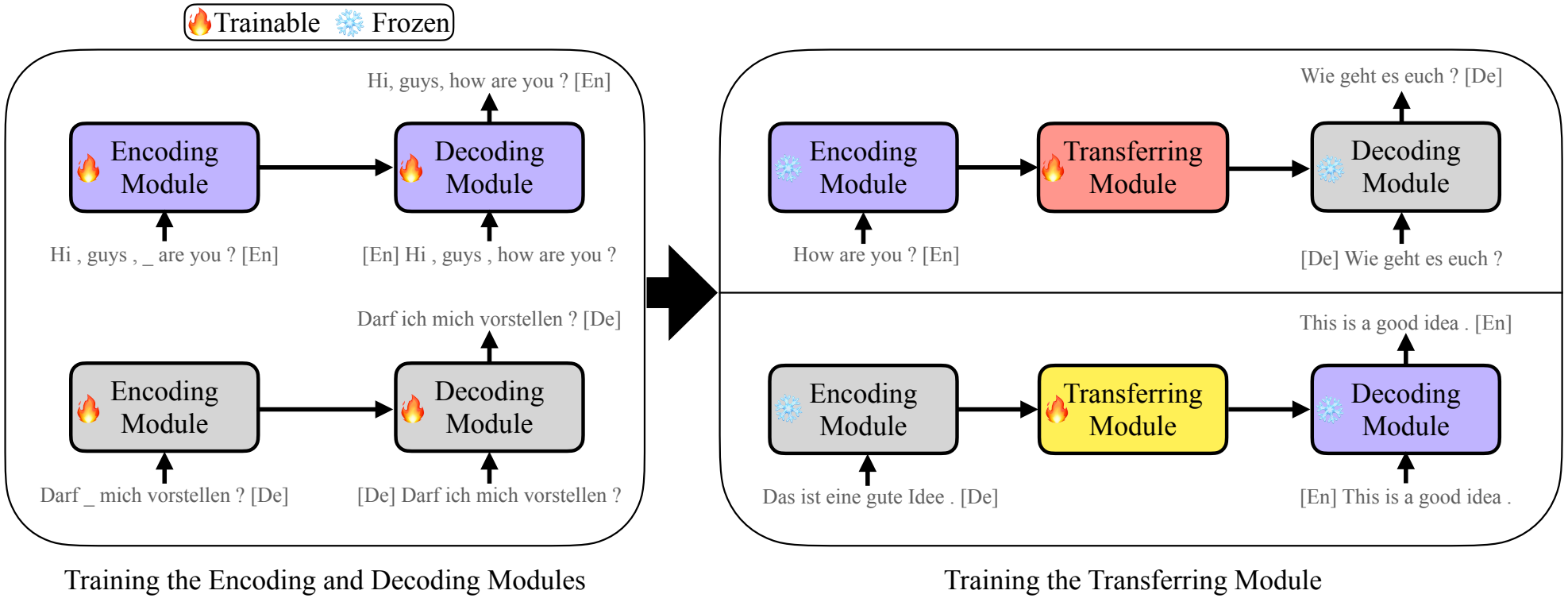
Our Model

MoNMT: separates the translation model into three modules.



- Encoding Module: trained on source monolingual data.
- Decoding Module: trained on target monolingual data.
- Transferring Module: trained on parallel data.

Training Strategy



Training the Encoding and Decoding Modules

Training the Transferring Module

- Encoding Module: trained on source monolingual data.
- Decoding Module: trained on target monolingual data.
- Transferring Module: trained on parallel data.

MoNMT: Experiment &Results

1. Multi-domain datasets
   - In-domain tests.
   - Out-of-domian tests.

	News			Medical			Law			Koran			IT			Subtitles		
	RD	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**
News	33.0	33.4	33.9	7.2	8.1	17.4	12.0	13.0	20.3	1.4	3.6	7.3	7.8	17.3	18.2	14.5	19.1	23.3
Medical	34.8	36.4	37.7	51.1	52.6	52.5	18.6	24.6	29.1	0.0	1.0	6.9	10.8	24.4	26.9	4.9	13.1	24.4
Law	39.9	41.1	41.4	18.6	24.6	29.1	57.3	58.2	57.2	0.6	1.4	6.7	7.2	17.4	18.8	4.2	7.9	18.4
Koran	12.5	12.7	15.1	2.7	2.7	6.5	3.2	3.5	6.9	13.7	20.9	21.3	3.4	9.0	9.4	6.7	8.6	11.1
IT	31.1	31.8	32.1	10.0	11.2	22.5	11.6	14.7	23.0	0.6	1.5	4.5	39.7	41.8	42.7	6.1	8.7	18.6
Subtitles	22.3	22.9	23.1	3.2	3.6	8.7	4.0	4.2	7.4	1.5	3.0	4.8	8.4	15.1	14.3	30.7	32.2	31.2
Average	28.9	29.7	**30.6**	14.8	16.0	**22.5**	17.8	19.7	**24.0**	3.0	5.2	**8.6**	12.9	20.9	**21.7**	11.2	13.9	**21.2**

(a) The BLEU scores for German-to-English on multi-domain translation tasks.

	News			Medical			Law			Ted		
	RD	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**	**RD**	**PF**	**Ours**
News	31.0	35.9	36.3	5.5	7.2	16.6	8.8	10.1	20.6	14.5	20.9	25.4
Medical	21.6	30.3	33.6	82.5	82.3	83.1	13.3	16.1	24.8	6.1	14.3	22.1
Law	33.2	39.3	40.6	12.6	17.0	26.3	61.2	63.4	62.5	7.3	7.4	19.2
Ted	22.6	28.5	29.0	4.1	4.5	11.8	6.3	7.5	13.5	19.1	41.9	42.5
Average	31.0	33.4	**34.9**	26.2	27.6	**34.5**	22.7	24.3	**30.4**	16.8	21.1	**27.3**

(b) The BLEU scores for Romanian-to-English on multi-domain translation tasks.

Table 1: Main Results, where the methods are the Transformer model with a random initialization (**RD**), the pretrain-and-finetune paradigm (**PF**), and the MoNMT model (**Ours**). Noted that All the models are trained on training sets in the first row and tested on the test sets in the first column.

MoNMT: Experiment &Results

2. Low-resource and high-resource datasets.

Model	WMT14 En ⇔ Fr		WMT14 En ⇔ De		WMT16 En ⇔ Ro		WMT18 En ⇔ Tr		#Trained Parameters
	En ⇒ Fr	Fr ⇒ En	En ⇒ De	De ⇒ En	En ⇒ Ro	Ro ⇒ En	En ⇒ Tr	Tr ⇒ En	
RD-Base	40.9	36.9	27.3	31.9	33.9	29.8	9.4	15.3	61M
PF-Base	41.3	37.4	27.9	32.5	35.4	34.5	11.1	17.6	61M
MoNMT-Base	39.7	35.9	27.9	32.2	36.2	35.3	12.7	19.3	19M
RD-Big	42.2	38.4	27.9	33.0	34.2	31.0	1.3	3.8	211M
PF-Big	**42.6**	38.7	29.1	33.4	37.4	35.9	13.0	20.7	211M
MoNMT-Big	42.3	**38.8**	**29.4**	**33.9**	**37.6**	**36.3**	**13.8**	**20.9**	76M

Table 2: Results on common-used translation tasks. "Base" and "Big" indicate that the model layer settings are the same as those of Transformer-Base and Transformer-Big (Vaswani et al., 2017). The high- and low-resource tasks are arranged in a left-to-right manner for ease of comparison.

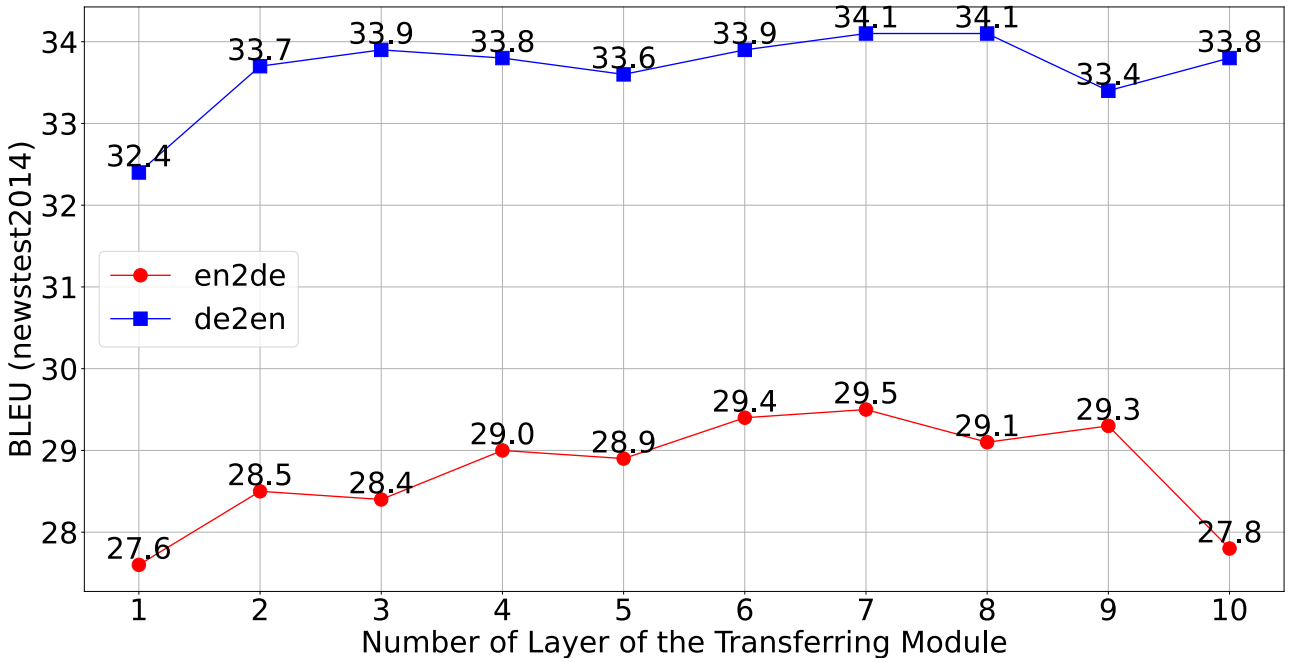
MoNMT: Analysis

1. The layer settings of the transferring module

- Findings:

✓ Only a single-layer performs comparably to the baseline. (13M parameters v.s. 211M< parameters)

✓ MoNMT with Seven layers achieves the best performance



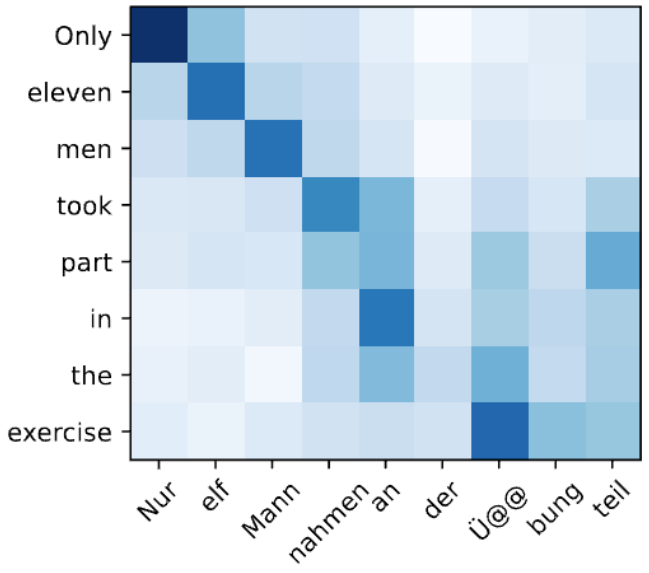
MoNMT: Analysis

2. Interpretability: how does the transferring module work?

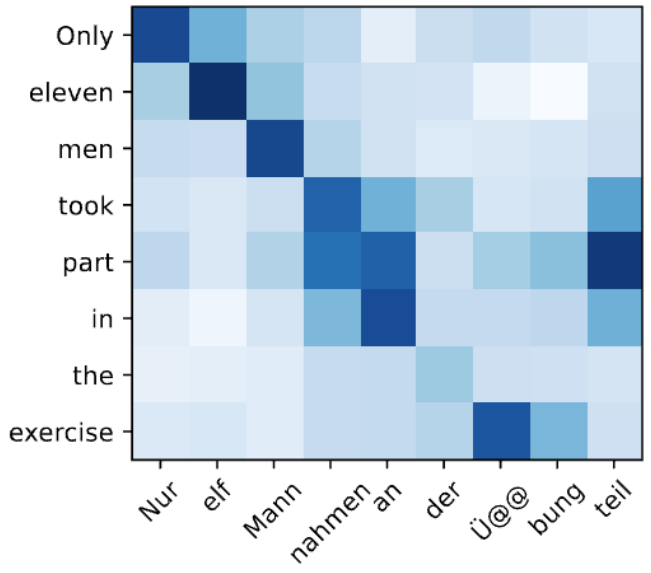
- Findings:

✓ Converts the source feature to the target feature.

✓ The pretrained model contains some alignment knowledge.



(a) Correlation of the Enc output source and target representations.



(b) Correlation between the Enc output source representations and Trans output target representations.

Figure 4: Heap maps of word-level correlation coefficient metrics of a German-to-English translation case.

Conclusion

- ✓ MoNMT sucessfully **synergizes monolingual and bilingual knowledge**, and avoids catastrophic forgetting.
- ✓ MoNMT show strong ability in **domain generalization and robustness**.
- ✓ MoNMT is effective in both **low- and high-resource** translation tasks.