DataFunSummit

自然语言处理

在线峰会

NLP基础技术 论坛

2021.07.10（周六） 09：00~17：30

2021

Microsoft Research 微软亚洲研究院 | DataFunSummit

多语言预训练模型在机器翻译中的应用

马树铭 微软亚洲研究院 研究员

Roadmap of Machine Translation

SMT
(Statistical Machine Translation)

- Translation model
- Language model
- Feature engineering

NMT
(Neural Machine Translation)

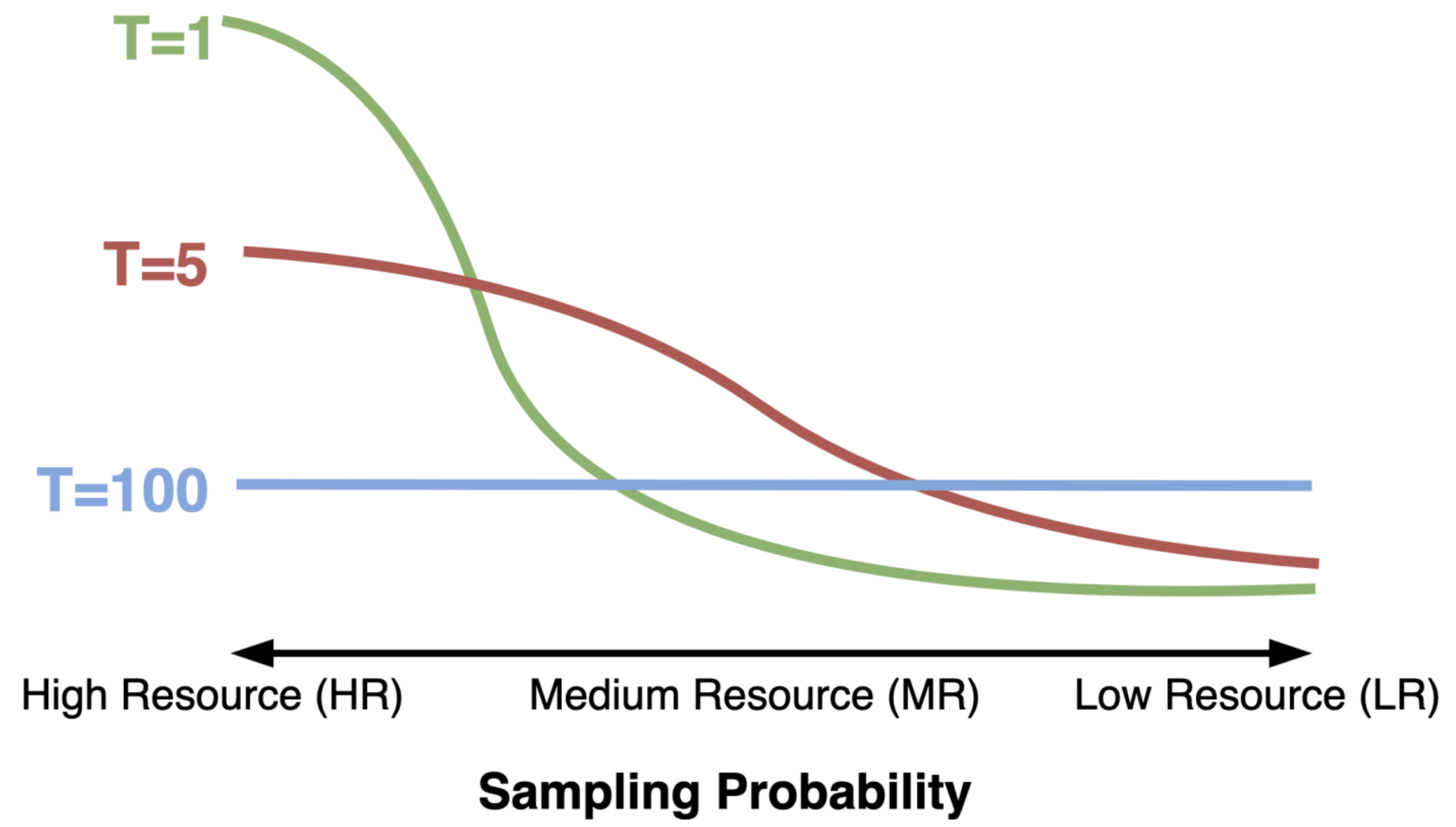
- End-to-end
- Multiple models for different languages

MNMT
(Multilingual Neural Machine Translation)

- One unified model
- Reduce training and serving cost
- Transfer among languages

Multilingual Neural Machine Translation

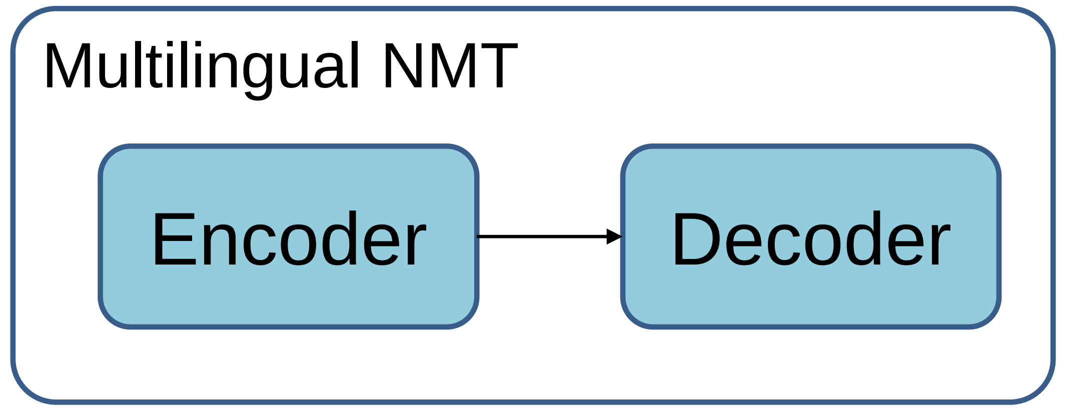
- Training
  - –Combination of multilingual language pairs
  - –Sampling the training data according to the data size
    - Sampling ratio $\alpha_L = \frac{{D_L}^{1/T}}{\sum {D_i}^{1/T}}$



- Modeling
  - One unified model
    - All languages share the same parameters
  - Cross-lingual Transferability
    - High-resource languages help low-resource ones
  - Prepend a language tag before the input
    - Indicate which target language to translate

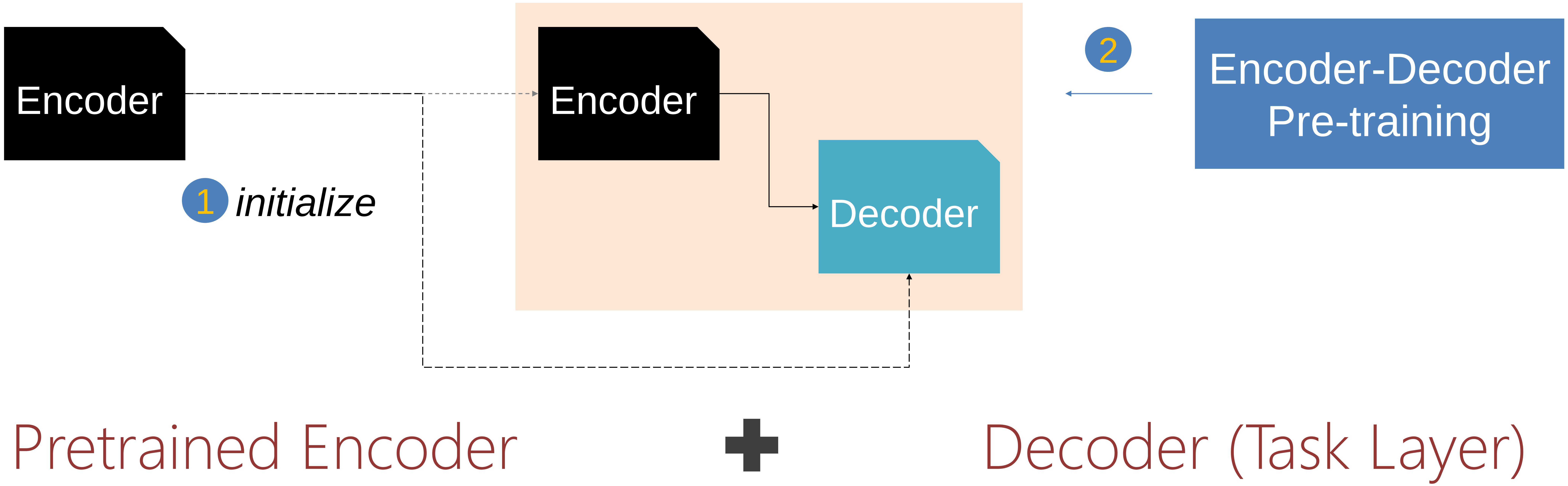
<zh> I want to read a book. → 我想读一本书。

<jp> I want to read a book. → 私は本を読みたい



Pretrained Model: DeltaLM[1]

- A pretrained encoder-decoder model for generation and translation.



[1] Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, Furu Wei. 2021. DeltaLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders. CoRR abs/2106.13736

DeltaLM: Decoder as the Task Layer

Pretrained Encoder

- Efficiency:
  - Speed up convergence & reduce training cost
- Effectiveness:
  - A strong encoder is important for MT
    - Our experiments verify this conclusion
  - Inheriting the cross-lingual transfer capability
    - Achieve SOTA results across NLU benchmarks

➕

Decoder (Task Layer)

- Decoupling enc. & dec. helps multilingual NLG
  - Hard to share space between languages
- More flexible in decoder’s architecture

We can unify two parts via encoder-decoder pre-training.

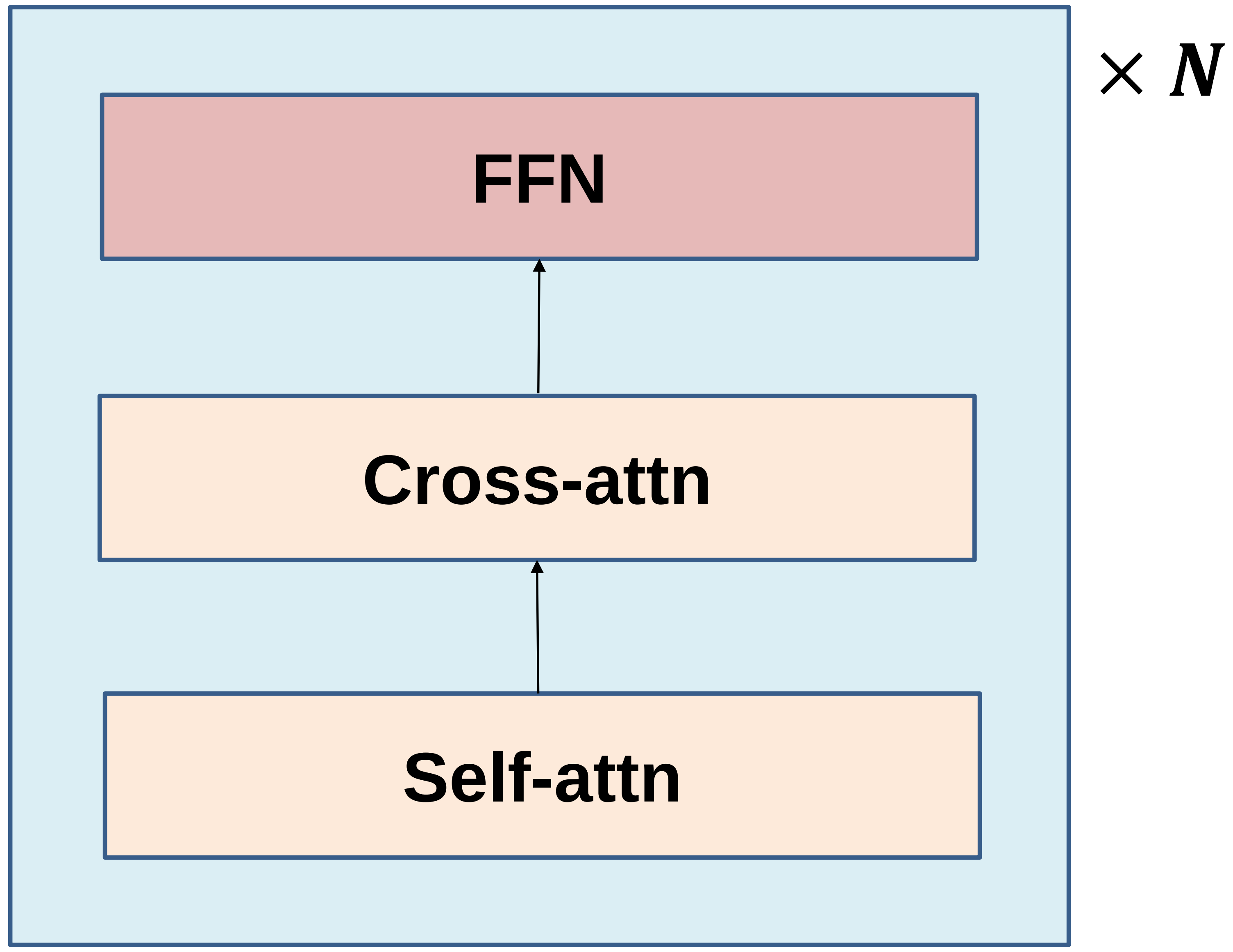
DeltaLM: Decoder as the Task Layer

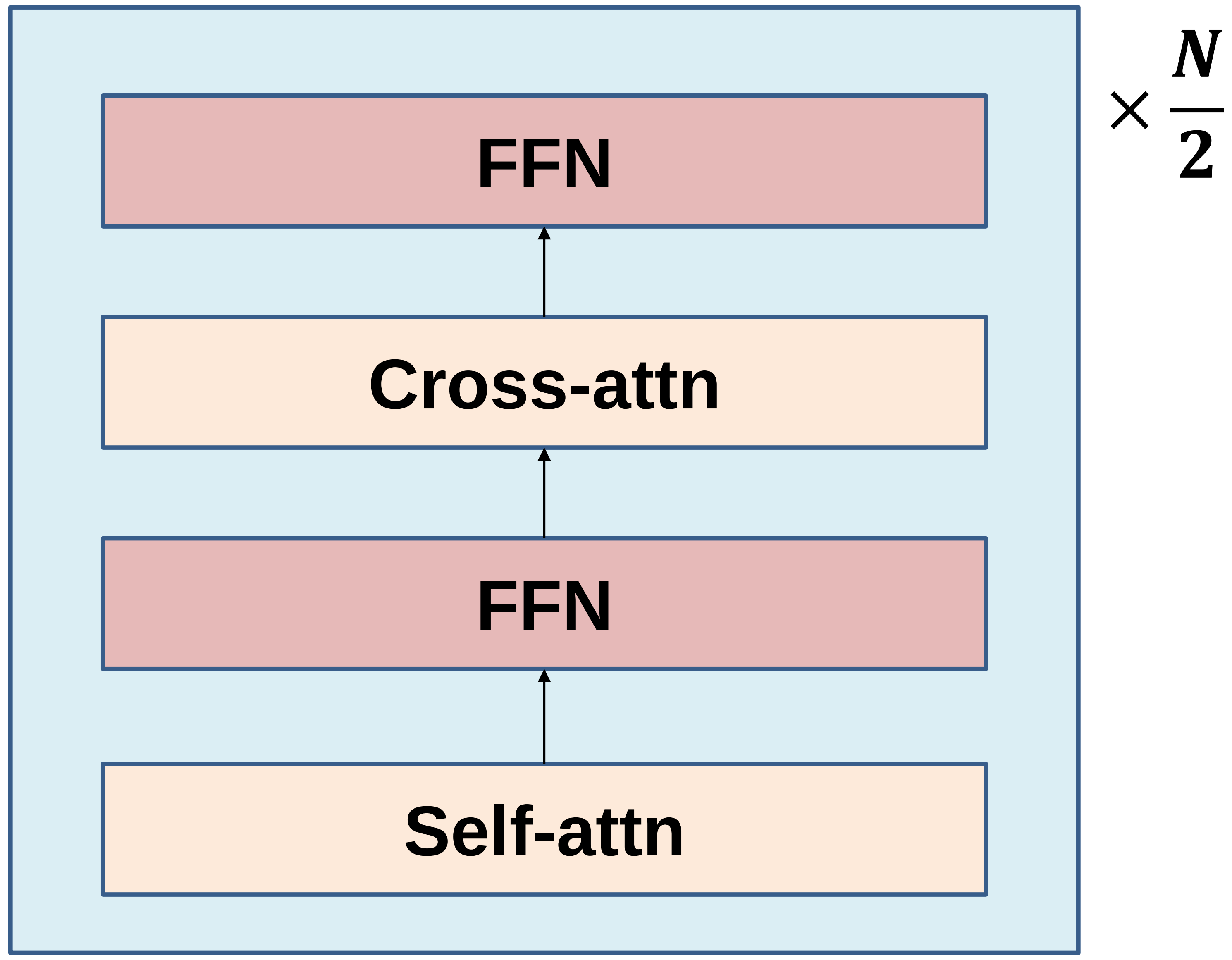
- How to initialize the decoder?
  - The structure of decoder is different from the encoder
  - Decoder initialization is understudied
- Which tasks to pre-train the encoder-decoder?
  - Mostly preserve the capability of pretrained encoder
  - Effectively leverage bilingual data

DeltaLM: Initialization

Microsoft Research 微软亚洲研究院 | DataFunSummit

- A novel interleaved decoder fully initialized by pretrained encoders





Vanilla decoder:

- One self-attn, one cross-attn, one FFN
- Initialization:
  - Pretrained encoder → Self-attn + FFN
  - Random initialized Cross-attn
- Cons:
  - Inconsistent with pretrained encoder (one FFN after one attention)

Our interleaved decoder:

- One FFN after one attention
- Initialize the self-attn/cross-attn in the interleaved way
  - Odd layers of pretrained encoder → Self-attn + FFN
  - Even layers of pretrained encoder → Cross-attn + FFN
- Fully use the weights of pretrained encoder

DeltaLM: Pre-training Task

- A novel pre-training task to leverage monolingual text + bilingual text
  - Span Corruption Task (T5):
    Reconstruct the text spans based on the input document

Input:

Thanks [Mask1] invitation [Mask2].

Target:

[Span1] for your [Span2] last week

  - Translation Pair Span Corruption Task[1]:
    Predict the text spans based on the input masked translation pair

Input:

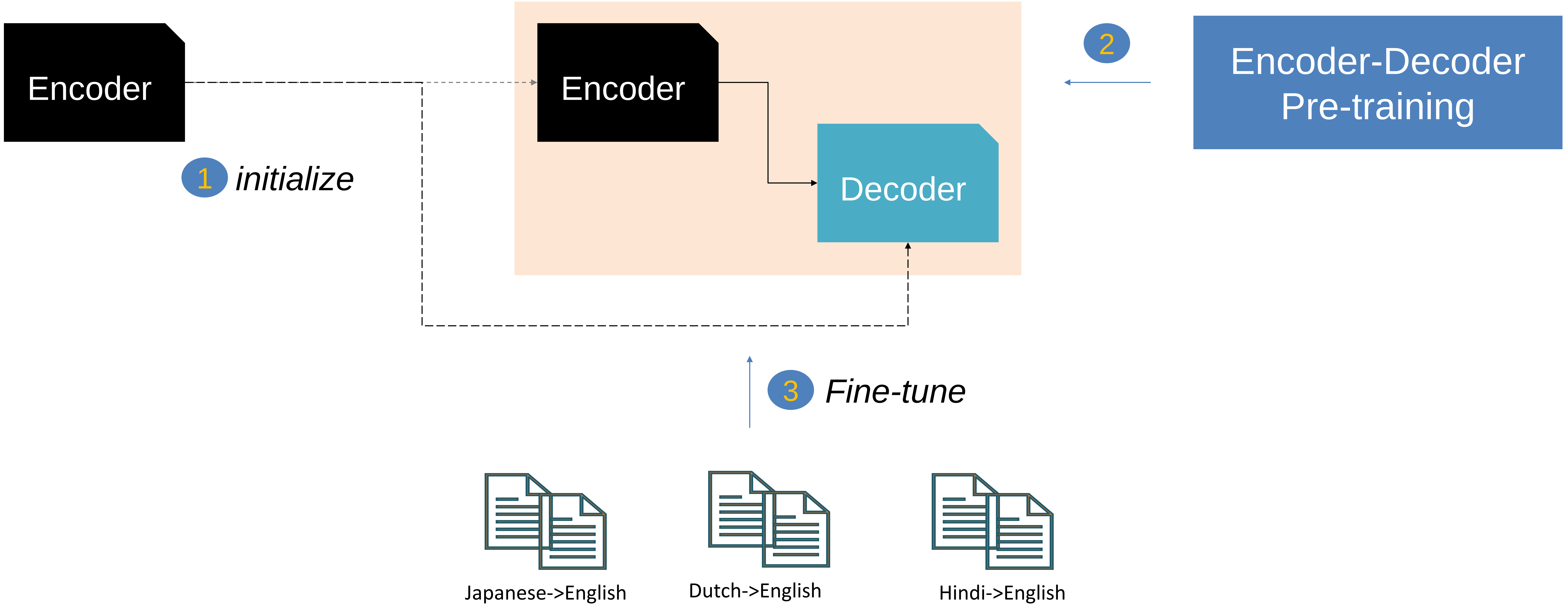
Thanks [Mask1] invitation [Mask2].
谢谢你上周的[Mask3]。

Target:

[Span1] for your [Span2] last week
[Span3] 邀请

DeltaLM for MNMT

- For MNMT, we can directly fine-tune DeltaLM.



Experiments: Multilingual Machine Translation

- DeltaLM reaches SOTA results on both X->E and E->X translation

X->E	#Params	fr -> en	cs -> en	de -> en	fi -> en	lv -> en	et -> en	ro -> en	hi -> en	tr -> en	gu -> en	Avg.
Transformer-big	240M	34.8	29.0	40.1	21.2	20.4	26.2	34.8	22.8	23.8	19.2	27.2
mBART	610M	36.2	29.9	40.0	22.2	20.6	27.2	37.2	23.3	25.7	21.7	28.4
FB m2m	420M	33.4	26.2	35.6	19.6	19.9	25.8	34.1	22.0	23.4	0.4	24.0
FB m2m	1.2B	35.8	29.6	40.7	22.8	23.0	30.6	38.2	24.6	26.1	0.5	27.2
DeltaLM	360M	36.5	30.9	42.2	23.0	22.3	29.2	37.7	27.0	27.3	22.7	**29.9**

E->X	#Params	en -> fr	en -> cs	en -> de	en -> fi	en -> lv	en -> et	en -> ro	en -> hi	en -> tr	en -> gu	Avg.
Transformer-big	240M	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
mBART	610M	33.7	20.8	38.9	14.5	18.2	20.5	26.0	15.3	16.8	12.9	21.8
FB m2m	420M	31.5	18.4	33.9	13.1	15.4	18.6	27.9	17.3	14.5	0.3	19.1
FB m2m	1.2B	35.5	22.1	42.2	16.6	19.2	22.9	32.0	17.9	15.5	1.3	22.5
DeltaLM	360M	35.8	22.4	40.9	15.7	18.8	20.6	26.9	17.3	18.5	16.2	**23.3**

* BLEU-4 is the evaluation metrics

** FB m2m supports 101 languages while DeltaLM is fine-tuned on a 11-language dataset

Experiments: Cross-lingual Summarization

- DeltaLM is competitive compared with mt5 large given only 30% parameters.

WikiLingua Dataset:

- Input: Spanish/Russian/Vietnamese/Turkish document
- Output: English summary

Models	#Params	es			ru			vi			tr			Avg.		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mBART	610M	38.3	15.4	32.4	33.1	11.9	27.8	32.0	11.1	26.4	34.4	13.0	28.1	34.5	12.9	28.7
mt5 small	300M	29.8	9.8	25.5	27.2	8.5	23.2	29.4	10.9	23.4	23.5	6.0	19.0	27.5	8.8	22.8
mt5 base	580M	36.3	13.7	30.6	32.5	11.1	26.9	32.5	13.6	26.0	26.0	7.5	20.5	31.8	11.5	26.0
mt5 large	1.2B	39.3	15.7	33.0	35.0	12.7	28.8	29.9	9.6	23.8	36.2	15.0	29.1	35.1	13.3	28.7
mt5 XL	3.7B	41.8	17.4	34.7	38.6	15.4	32.3	35.5	13.0	29.2	41.5	19.6	34.7	39.4	16.4	32.7
DeltaLM	360M	36.5	13.6	29.7	33.4	12.0	27.2	31.8	10.8	25.7	39.6	17.1	32.3	35.3	13.4	28.7

* R-1, R-2, R-3 denotes ROUGE-1, ROUGE-2, ROUGE-L

Experiments: Data-to-text Generation

- DeltaLM outperforms mt5 XL (3.7B) with only 360M parameters.

(JOHN E BLAHA BIRTHDATE 1942 08 26)
(JOHN E BLAHA BIRTHPLACE SAN ANTONIO)
(JOHN E BLAHA OCCUPATION FIGHTER PILOT)

→

John E Blaha, born in San Antonio on 1942-08-26, worked as a fighter pilot.

Data **Text**

Models	#Params	en			ru			Avg.		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mBART	610M	83.4	63.1	70.3	34.8	13.4	33.0	59.1	38.3	51.7
mt5 small	300M	78.8	59.2	67.2	29.7	10.5	28.4	54.3	34.9	47.8
mt5 base	580M	82.3	62.1	69.7	33.0	12.7	31.3	57.7	37.4	50.5
mt5 large	1.2B	83.8	64.4	71.6	33.4	13.4	32.1	58.6	38.9	51.9
mt5 XL	3.7B	83.5	63.6	71.0	34.3	13.7	32.8	58.9	38.7	51.9
DeltaLM	360M	83.4	63.9	71.1	35.0	15.0	33.3	**59.2**	**39.4**	**52.2**

* R-1, R-2, R-3 denotes ROUGE-1, ROUGE-2, ROUGE-L

Experiments: Multilingual Language Generation

- DeltaLM achieves consistent improvement across different tasks/languages.

Settings:
- Question generation (XQG)
  - Input: Chinese answer and the corresponding document
  - Output: Chinese question
- Abstractive summarization (XGiga)
  - Input: French document
  - Output: French summary

Models	#Params	XQG			XGiga		
		BLEU-4	METEOR	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
XLM	570M	23.41	23.32	47.20	56.27	39.20	52.84
XNLG[1]	480M	24.89	24.53	49.72	57.84	40.81	54.24
DeltaLM	360M	**25.80**	**24.87**	**52.05**	**58.39**	**42.02**	**54.94**

Experiments: Zero-shot Cross-lingual Transfer

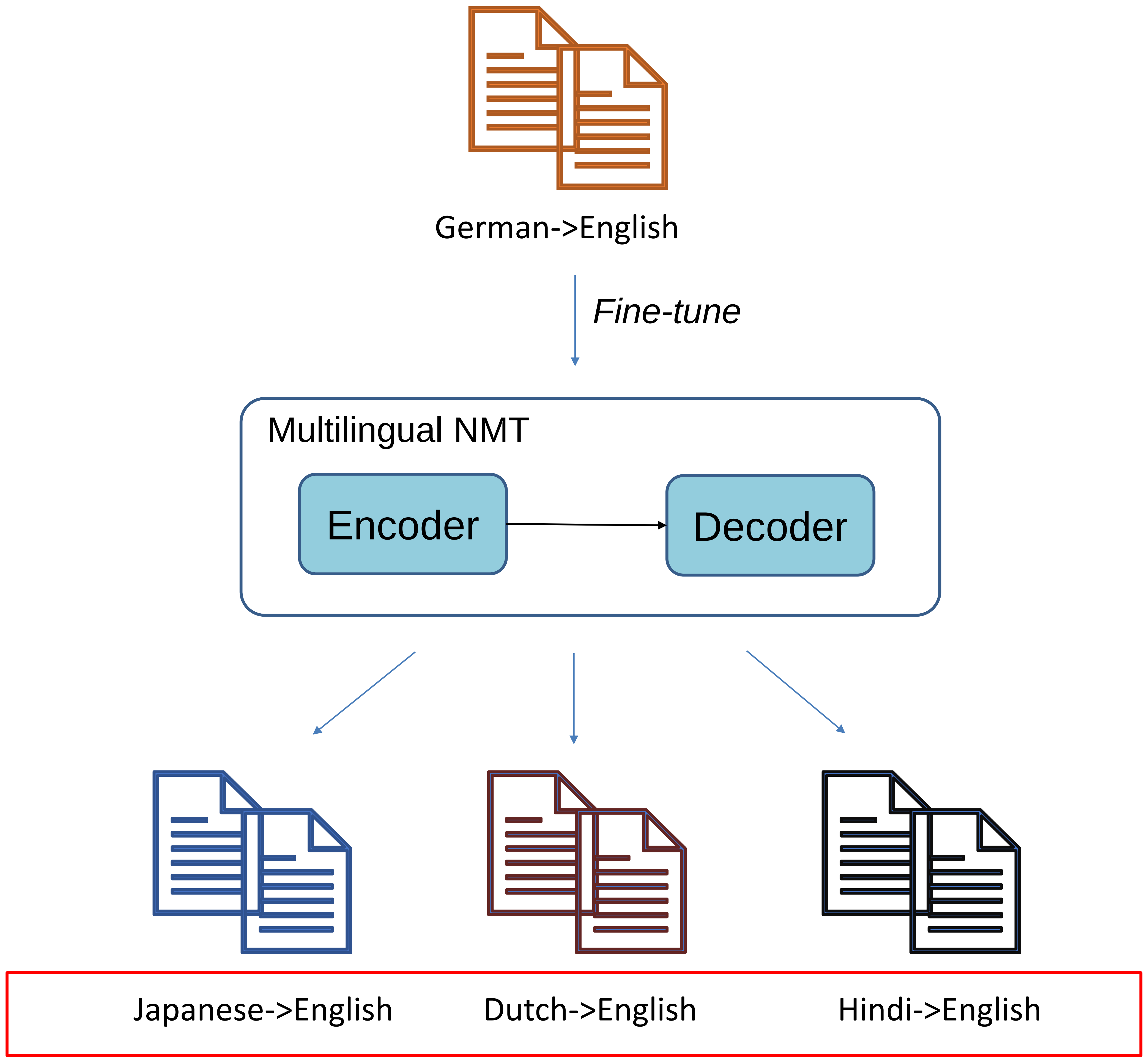
- DeltaLM has good capability of zero-shot transfer for NLG.

Settings:

- Abstractive summarization (XGiga)
  - Training:
    - English document → English summary
  - Testing:
    - French document → French summary
    - Chinese document → Chinese summary

Models	#Params	XGiga-fr			XGiga-zh		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
XLM	570M	14.53	1.80	13.43	0.71	0.28	0.70
XNLG	480M	39.98	20.31	36.31	41.66	28.70	38.91
DeltaLM	360M	**41.42**	**22.24**	**37.99**	**46.37**	**34.34**	**43.85**

Zero-shot Cross-lingual Transfer of NMT[1]



Do not have any training data

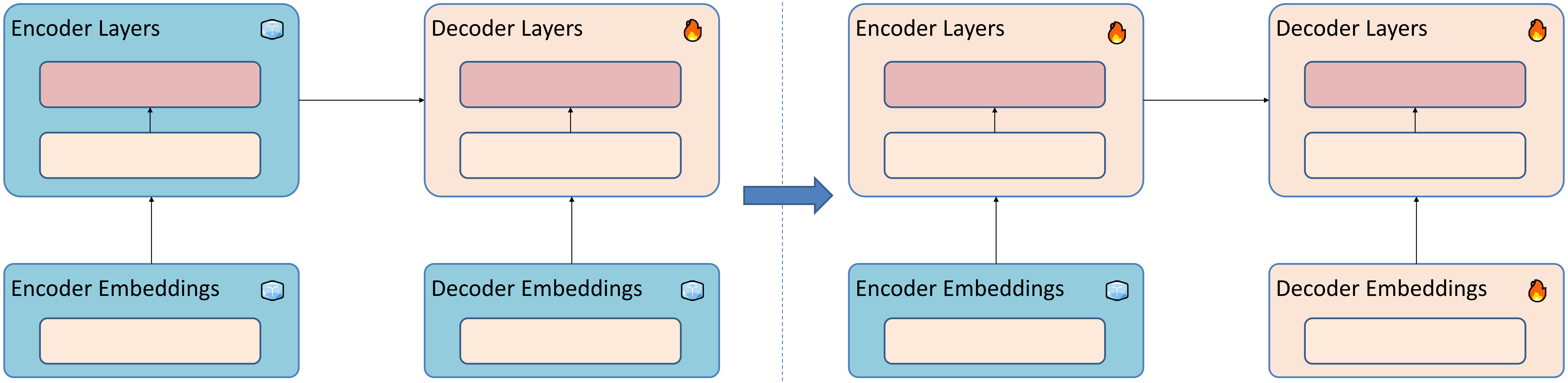
- Training
  - One language pair, e.g., German->English
- Modeling
  - One **unified** MT model
    - Cross-lingual Transferability
- Testing (zero-shot)
  - Unseen languages, e.g., Japanese->English

[1] Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, Furu Wei. 2021. Zero-shot Cross-lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders. CoRR abs/2104.08757

Two-stage Fine-tuning Method

Stage 1

Stage 2



- Freeze encoder & decoder embeddings
  - Preserve the cross-lingual transferability of the pretrained model
- Fine-tune decoder layers
  - Adapt the decoder to the pretrained encoders

- Fine-tune encoder layers & decoder
  - Improve the translation quality
  - Our preliminary experiments find this strategy is the best
- Remove the residual connection for self-attn
  - make the encoder outputs less position- and language-specific

Frozen

Fine-tuned

Experimental Details & Results

- Dataset
  - Trained on De-En parallel dataset
    - WMT19 43M parallel data
  - Tested on many-to-English language pairs
    - German group, Romance group, Slavic group, Uralic group, and Turkic group
    - German (De), Dutch (Nl), Spanish (Es), Romanian (Ro), Finnish (Fi), Latvian (Lv), Turkish (Tr), Russian (Ru), Polish (Pl)
- Results

Model	Data	German		Romance			Uralic			Indo-Aryan				East Asian			Avg.
		De	Nl	Es	Ro	It	Fi	Lv	Et	Hi	Ne	Si	Gu	Zh	Ja	Ko	
mBART	0.04B	27.4	43.3	24.7	28.2	29.8	18.8	14.2	15.7	12.3	9.6	7.2	10.3	8.3	6	21.1	18.4
CRISS	1.8B	28.8	47	32.2	35.4	48.9	23.9	18.6	23.5	23.1	14.7	14.4	19	13.4	7.9	24.8	25
M2M-100	7.5B	28	48.5	30	34.1	50	24.9	19.9	25.8	21.9	3.7	10.6	0.4	19.5	11.5	32.7	24.1
Ours	0.04B	33.8	54.7	30.1	33.9	43	26.3	17.7	25.7	17.5	14.4	12.2	17.3	13.4	10.7	31.2	**25.5**

Less data achieves better results

Transferability vs. Language Similarity

- Training with different languages
  - German (De), Spanish (Es), Hindi(Hi)
- Testing on different language families
  - German Family (De, Nl), Romance Family (Es, Ro, It), Indo-Aryan Family (Hi, Ne, Si, Gu)

Model	Train set	German Family		Romance Family			Indo-Aryan Family				Avg.
		De->En	Nl->En	Es->En	Ro->En	It->En	Hi->En	Ne->En	Si->En	Gu->En	
MNMT Baseline	De->En	33.7	3	3.6	3.4	1.7	0.1	0.1	0.2	0.2	3.5
	Es->En	6.8	5.5	32.5	6.4	17.3	0.3	0.1	0.2	0.2	5.2
	Hi->En	0.6	0.9	0.2	0.5	0.6	21.5	3.6	0.1	0.2	2
Ours	De->En	33.8	54.7	30.1	33.9	43	17.5	14.4	12.2	17.3	25.5
	Es->En	19.9	38.2	33	30.9	47	6.9	4.2	3.4	5.6	16.9
	Hi->En	19	38	20.1	20.7	34.3	24.3	16.7	9.6	17.8	18.7

- The transfer ability of NMT model benefits more on similar languages than distant languages.
- Promising results of transferring insides the language family with only one language pair.

Conclusions

- Pretrained language model benefits machine translation.
  - Supervised learning of multilingual neural machine translation
  - Zero-shot cross-lingual transfer

- DeltaLM has good capability of cross-lingual transfer and language generation to help machine translation.

Microsoft Research 微软亚洲研究院 | DataFunSummit

THANKS!

今天的分享就到这里...

Ending