Rethinking the Value of Pre-training for NMT

Wenxuan Wang

2020-07

How Much Information is the Machine Given during Learning?

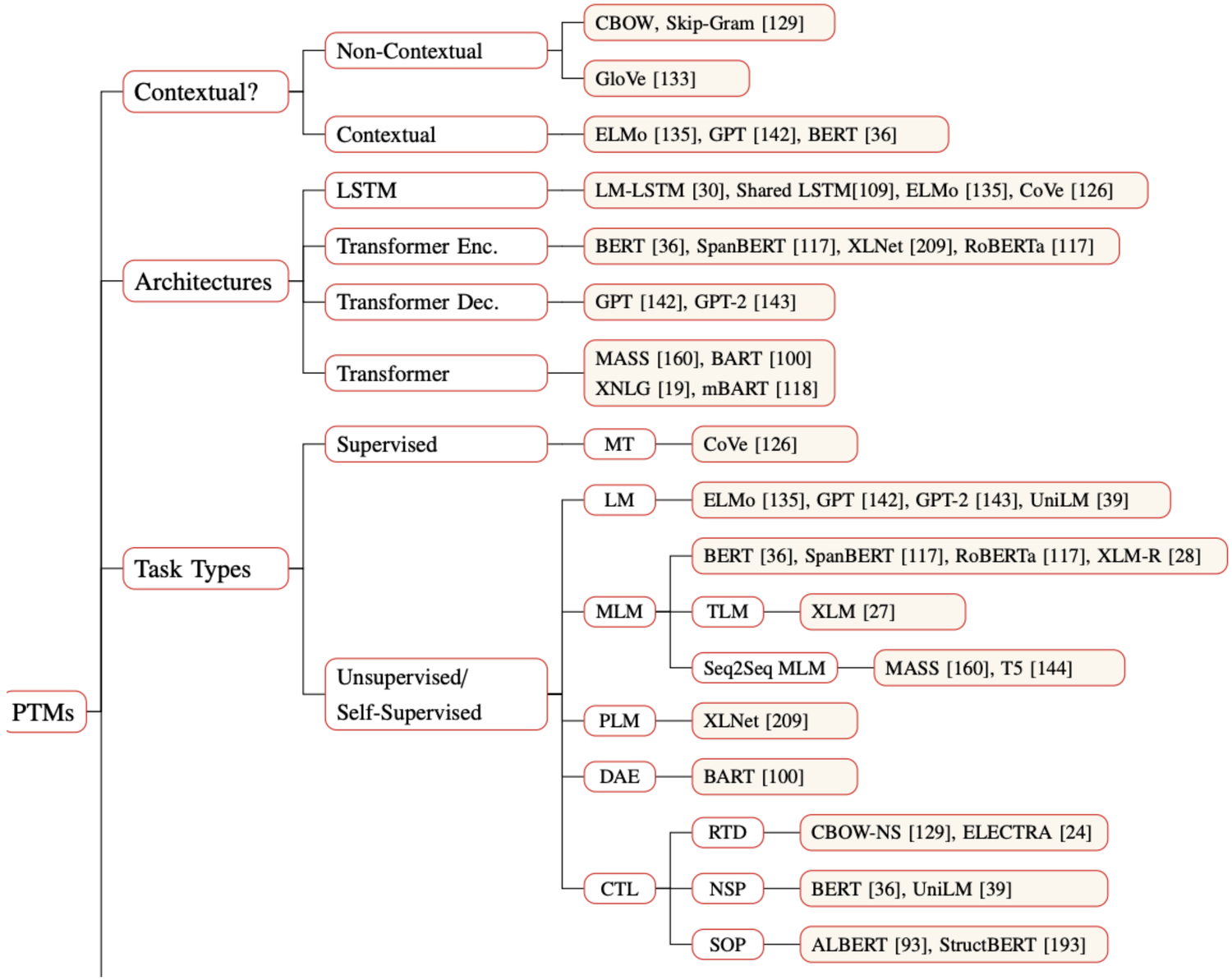
- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Lifelong Learning

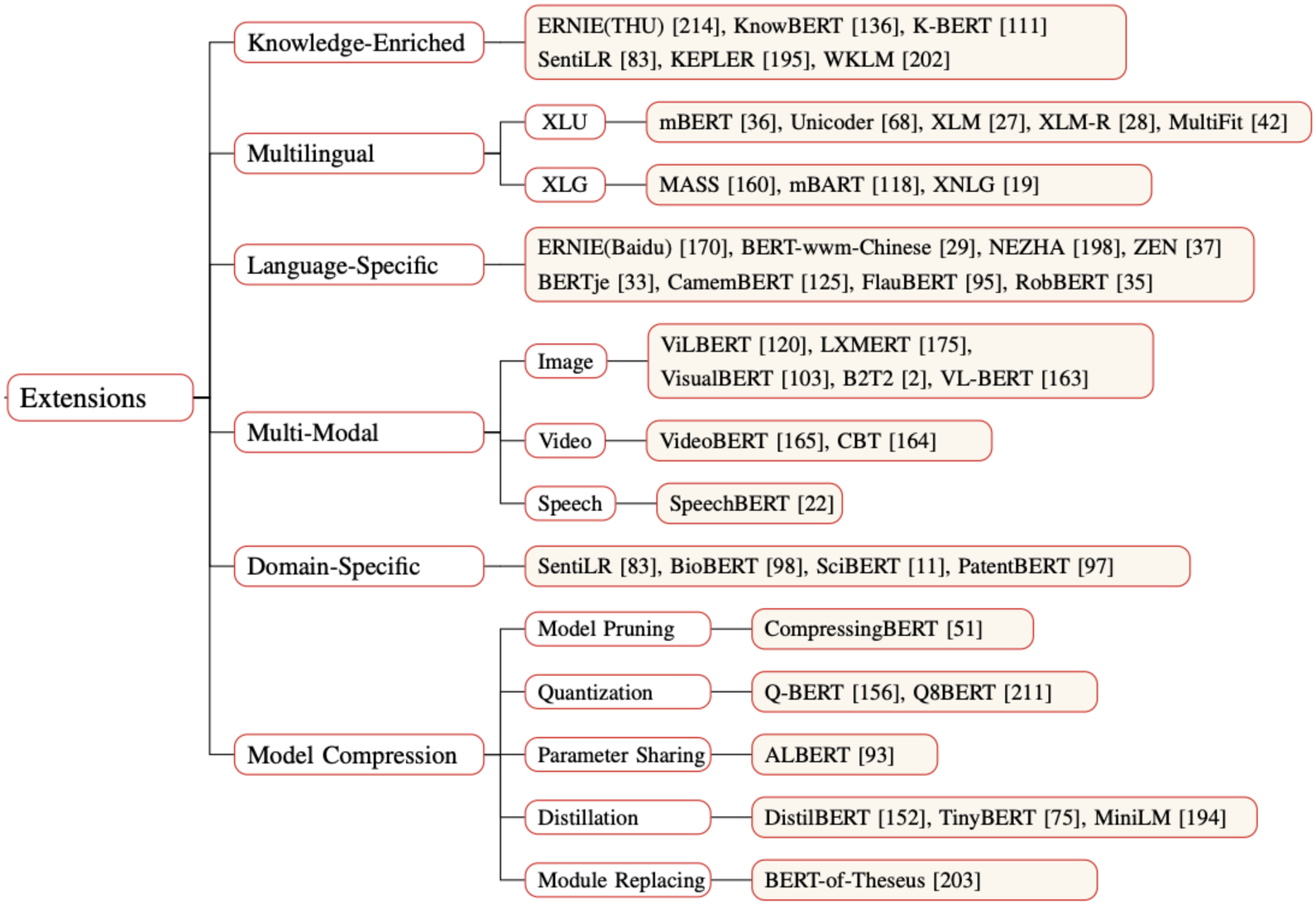
Motivation

- Pre-training models have made impact in CV and NLU.
  - Resnet
  - BERT, XLNET…
- How to utilize pre-train models for NMT(NLG)?

Pre-trained Models for Natural Language Processing: A Survey



Pre-trained Models for Natural Language Processing: A Survey



Motivation

- Transfer Learning in NMT
  - Pre-training + fine-tuning
    - Mono-lingual data pre-training + translation data fine-tuning.
    - High resource pre-training + low resource fine-tuning.
  - Multi-task learning
    - Multi-lingual Translation.

Research Directions

- What type of pre-training can help?
- When can pre-training help?
- How can pre-training help?
- Why does pre-training help?

What type of pre-training can help?

- Pre-training Model
  - Architecture (LSTM, Transformer)
  - Type (single direction autoregressive, bi-direction autoregressive, autoencoder)
  - Task
- Inherit Type
  - Feature Extractor
  - Weights Initialization

What?

- Pre-training Model

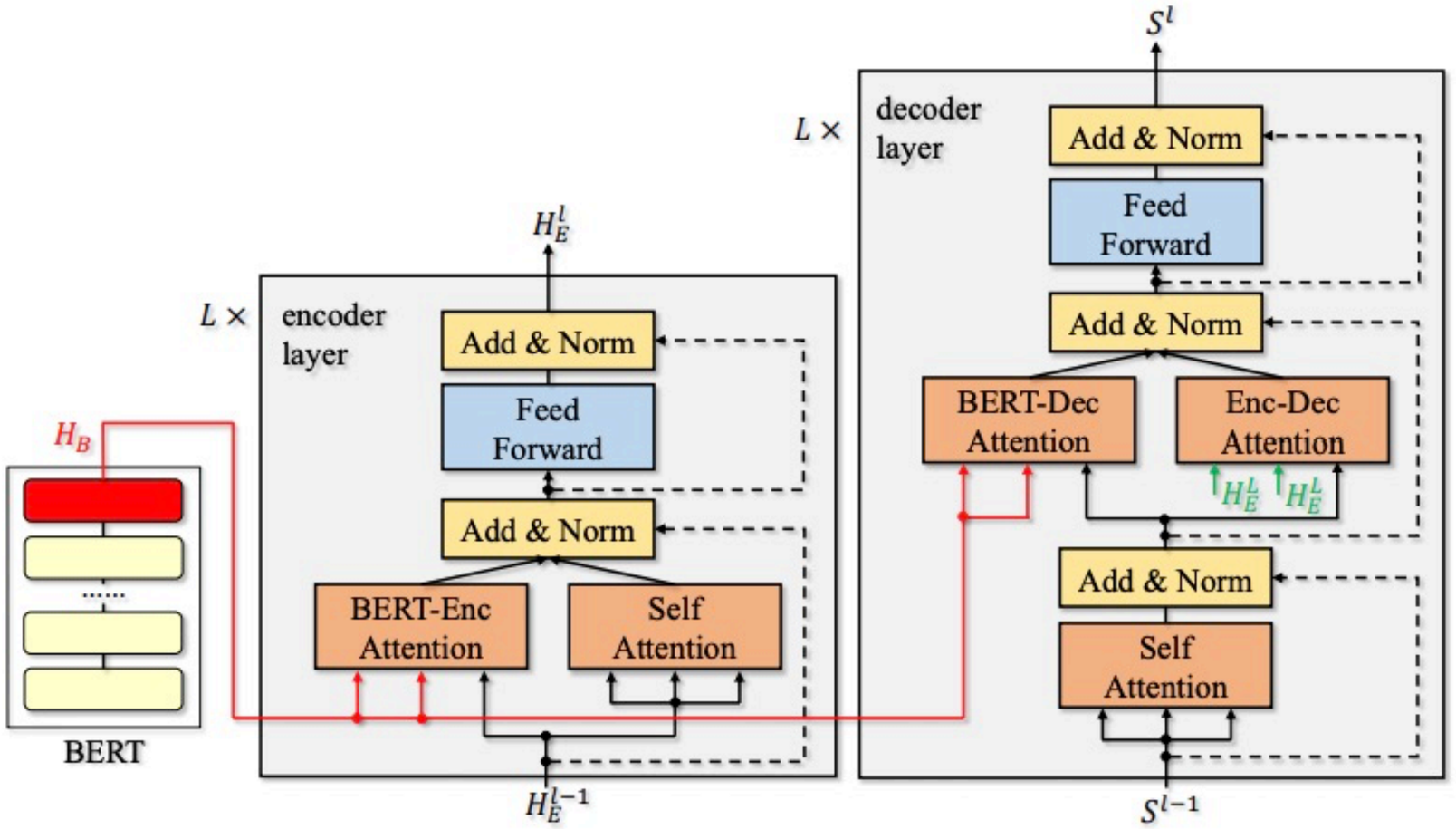
模型	语言模型	特征抽取	上下文表征	最大亮点
ELMO	BiLM	BiLSTM	单向	2个单向语言模型拼接；
ULMFiT	LM	AWD-LSTM	单向	引入逐层解冻解决finetune中的灾难性问题；
SiATL	LM	LSTM	单向	引入逐层解冻+辅助LM解决finetune中的灾难性问题；
GPT1.0	LM	Transformer	单向	统一下游任务框架，验证Transformer在LM中的强大；
GPT2.0	LM	Transformer	单向	没有特定模型的精调流程，生成任务取得很好效果；
BERT	MLM	Transformer	双向	MLM获取上下文相关的双向特征表示；
MASS	LM+MLM	Transformer	单向/双向	改进BERT生成任务：统一为类似Seq2Seq的预训练框架；
UNILM	LM+MLM+S2SLM	Transformer	单向/双向	改进BERT生成任务：直接从mask矩阵的角度出发；
ENRIE1.0	MLM(BPE)	Transformer	双向	引入知识：3种[MASK]策略(BPE)预测短语和实体；
ENRIE	MLM+DEA	Transformer	双向	引入知识：将实体向量与文本表示融合；
MTDNN	MLM	Transformer	双向	引入多任务学习：在下游阶段；
ENRIE2.0	MLM+Multi-Task	Transformer	双向	引入多任务学习：在预训练阶段，连续增量学习；
SpanBERT	MLM+SPO	Transformer	双向	不需要按照边界信息进行mask；
RoBERTa	MLM	Transformer	双向	精细调参，舍弃NSP；
XLNet	PLM	Transformer-XL	双向	排列语言模型+双注意力流+Transformer

What?

- Pre-training Model
- Inherit Type
  - Feature Extractor
    - Incorporating BERT into Neural Machine Translation (ICLR 2020)
  - Weights Initialization
    - Multilingual Denoising Pre-training for Neural Machine Translation (Arxiv 2020)

What?

- Incorporating BERT into Neural Machine Translation (ICLR 2020)



What?

- Incorporating BERT into Neural Machine Translation (ICLR 2020)

Table 2: BLEU of all IWSLT tasks.

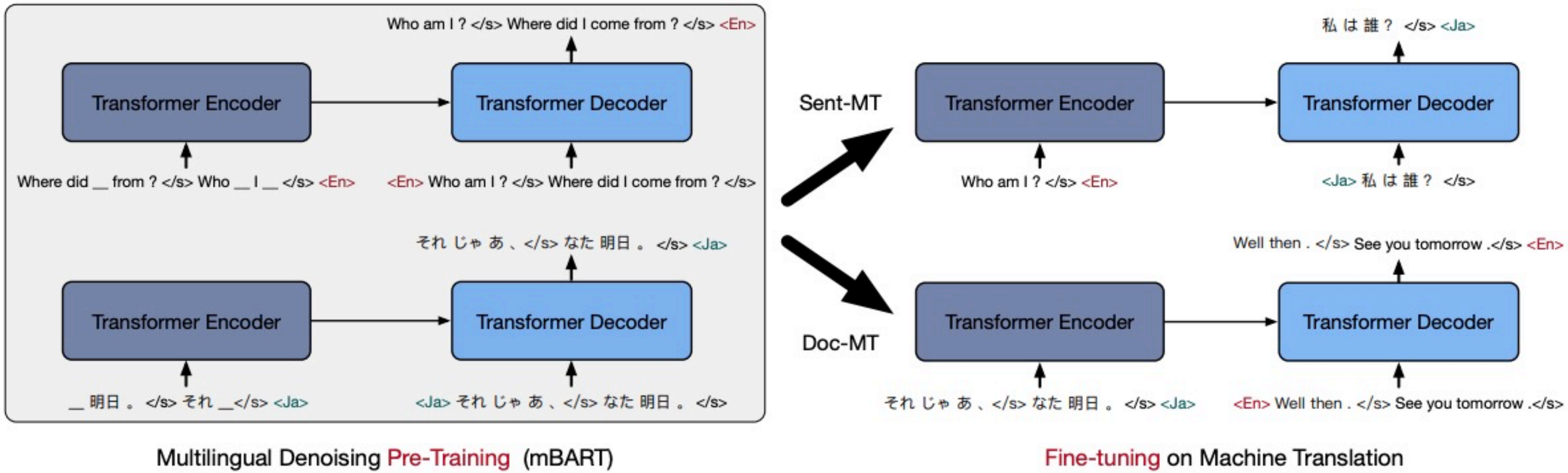
	Transformer	BERT-fused
En→De	28.57	30.45
De→En	34.64	36.11
En→Es	39.0	41.4
En→Zh	26.3	28.2
En→Fr	35.9	38.7

Table 3: BLEU scores of WMT'14 translation.

Algorithm	En→De	En→Fr
DynamicConv (Wu et al., 2019)	29.7	43.2
Evolved Transformer (So et al., 2019)	29.8	41.3
Transformer + Large Batch (Ott et al., 2018)	29.3	43.0
Our Reproduced Transformer	29.12	42.96
Our BERT-fused model	30.75	43.78

What?

- Multilingual Denoising Pre-training for Neural Machine Translation (Arxiv 2020)



What?

- Multilingual Denoising Pre-training for Neural Machine Translation (Arxiv 2020)

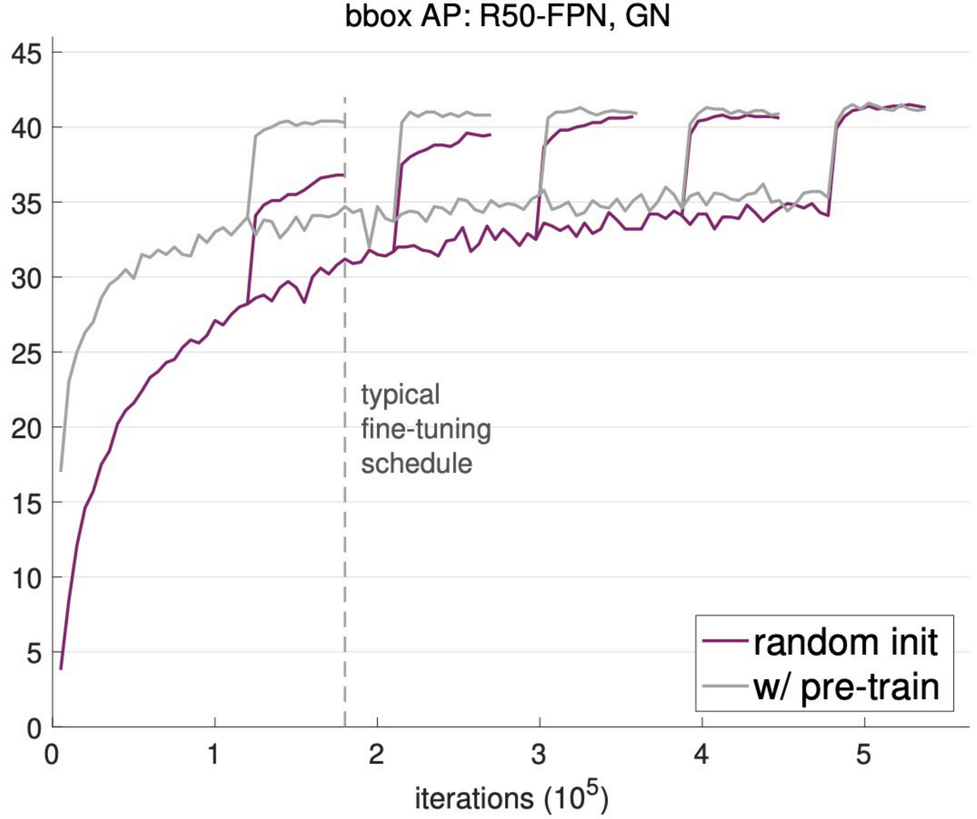
  - mBART initialization leads to significant gains (up to 12 BLEU points) across low/medium-resource pairs (<10M bi-text pairs), without sacrificing performance in high-resource settings.

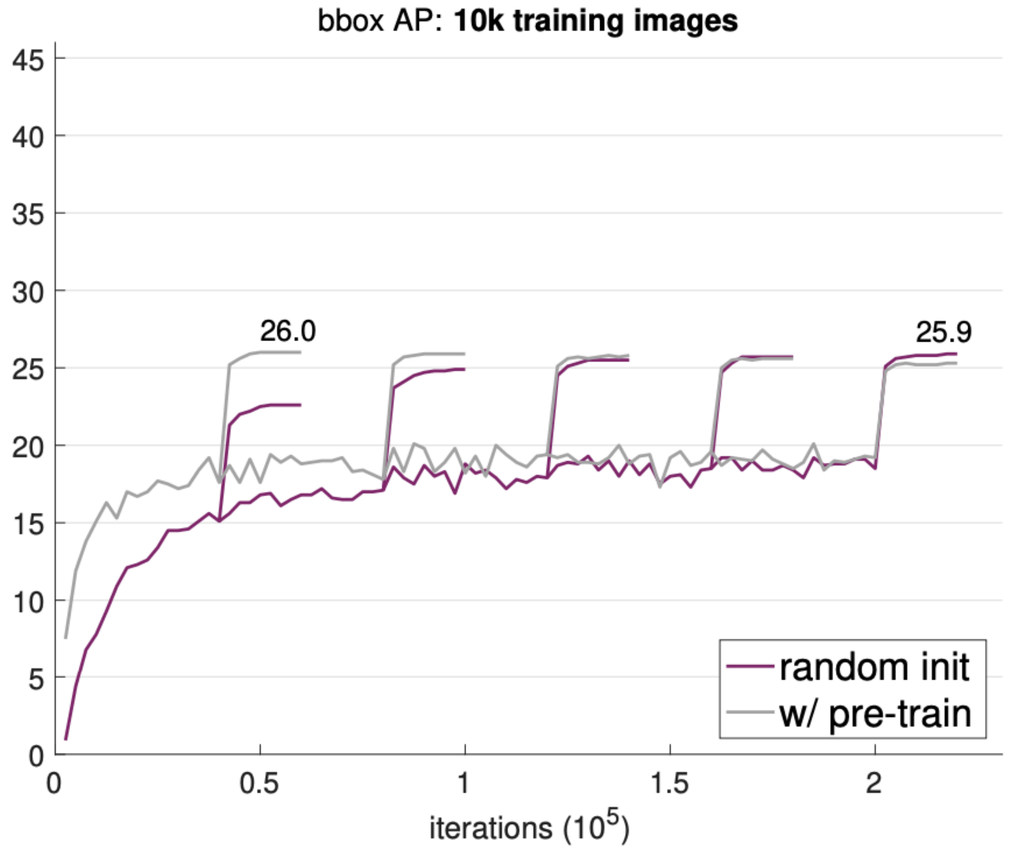
When?

- When can pre-training help?
- Do we always need per-training?
  - Rethinking ImageNet Pre-Training (ICCV 2019)
  - Rethinking Pre-training and Self-training (Arxiv 2020)
  - To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks (ACL 2020)

When?

- Rethinking ImageNet Pre-Training (ICCV 2019)
  - 不用imagenet预训练，随机初始化的模型和pretrain+finetune的模型性能一样，只是需要充足的训练的时间（标配pretrain+finitene的时间）。 用预训练的模型只是加速下游任务的收敛。





When?

- Rethinking Pre-training and Self-training (Arxiv 2020)

- 对比了pretrain和self train的区别。在cv的coco检测数据集上做实验，发现了:

  - 1）pre-train的帮助随着finetune的数据集大小的增大而减小，最终finetune的数据集够大时或者用了很强的数据增强方法的时候，用pretrain模型反而会降低最终的效果。但是self training并没有这个现象，数据集够大的时候也会有帮助。

  - 2）用了pretrain提升性能之后，再用self-training能继续提升性能。

Setup	Augment-S1	Augment-S2	Augment-S3	Augment-S4
Rand Init	39.2	41.5	43.9	44.3
ImageNet Init	(+0.3) 39.5	(-0.7) 40.7	(-0.8) 43.2	(-1.0) 43.3
Rand Init w/ ImageNet Self-training	(+1.7) 40.9	(+1.5) 43.0	(+1.5) 45.4	(+1.3) 45.6

When?

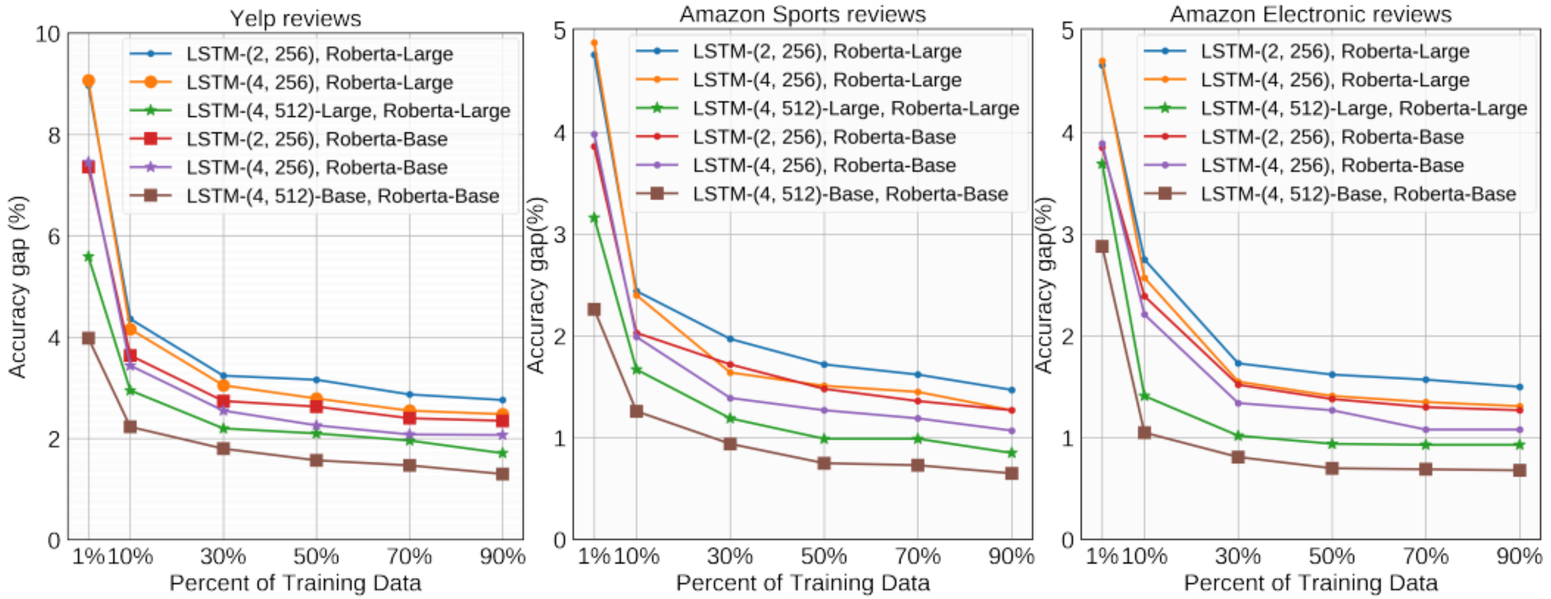
- To Pretrain or Not to Pretrain: Examining the Benefits of Pretrainng on Resource Rich Tasks (ACL 2020)

- 探究对于 pretraining model，下游任务数据量与预训练的帮助的关系

- 在roberta和lstm上做实验

- 选了三个下游分类任务来测试

- 下游任务数据越大，预训练的帮助越不明显



How can pre-training help?

- The Gap between Pre-training Task and Fine-tuning Task.
- Intermediate-Task:
  - Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? (ACL 2020)
  - Don't Stop Pretraining: Adapt Language Models to Domains and Tasks （ACL 2020 Honorable Mention Paper)
- Multi-task
  - Multi-Task Deep Neural Networks for Natural Language Understanding （ACL 2019)
  - Understanding and Improving Information Transfer in Multi-Task Learning （ICLR 2020)

How can pre-training help?

- Early-stop
  - Do Better ImageNet Models Transfer Better? (CVPR 2019)
- Fine-tuning setting is important.
  - Rethinking the Hyperparameters for Fine-tuning (ICLR 2020)
- How to align weights between pre-training model and fine-tuning model.
  - In Neural Machine Translation, What Does Transfer Learning Transfer? (ACL 2020)

How: Intermediate-Task

- Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? (ACL 2020)

- 预训练模型pretrain+下游任务finetune之间，会加入intermediate的任务来过渡训练。

- 什么类型的intermediate的任务会帮助下游任务，以及为什么会有帮助。

- 作者对roberta的预训练模型尝试了110种中间task-下游task的组合

	Name	|Train|	|Dev|	task	metrics	genre/source
Intermediate Tasks	CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
	SciTail	23,596	1,304	natural language inference	acc.	science exams
	Cosmos QA	25,588	3,000	question answering	acc.	blogs
	SocialIQA	33,410	1,954	question answering	acc.	crowdsourcing
	CCG	38,015	5,484	tagging	acc.	Wall Street Journal
	HellaSwag	39,905	10,042	sentence completion	acc.	video captions & Wikihow
	QA-SRL	44,837	7,895	question answering	F1/EM	Wikipedia
	SST-2	67,349	872	sentiment classification	acc.	movie reviews
	QAMR	73,561	27,535	question answering	F1/EM	Wikipedia
	QQP	363,846	40,430	paraphrase detection	acc./F1	Quora questions
	MNLI	392,702	20,000	natural language inference	acc.	fiction, letters, telephone speech
Target Tasks	CB	250	57	natural language inference	acc./F1	Wall Street Journal, fiction, dialogue
	COPA	400	100	question answering	acc.	blogs, photography encyclopedia
	WSC	554	104	coreference resolution	acc.	hand-crafted
	RTE	2,490	278	natural language inference	acc.	news, Wikipedia
	MultiRC	5,100	953	question answering	$F1_\alpha$/EM	crowd-sourced
	WiC	5,428	638	word sense disambiguation	acc.	WordNet, VerbNet, Wiktionary
	BoolQ	9,427	3,270	question answering	acc.	Google queries, Wikipedia
	CommonsenseQA	9,741	1,221	question answering	acc.	ConceptNet
	Cosmos QA	25,588	3,000	question answering	acc.	blogs
	ReCoRD	100,730	10,000	question answering	F1/EM	news (CNN, Daily Mail)

How: Intermediate-Task

- Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? (ACL 2020)

- 发现HellaSWAG, Cosmos QA和CommonsenseQA这种需要high level inference和resoning的task作为中间任务效果最好。

- 发现probing task并不能反映/预测下游任务的性能，说明现有的probing task还不太行，需要新的更有效的probing task出现。

Target	QAMR	CSenseQA	SciTail	CosmosQA	SocialIQA	CCG	HellaSwag	QA-SRL	SST-2	QQP	MNLI	Baseline Performance
CB	-4.0	-0.4	-6.2	-0.4	-21.7	-12.2	-3.1	-7.2	-1.2	-31.0	-0.4	99.1
COPA	-4.0	8.7	4.3	6.0	-3.7	-20.7	6.7	-3.7	-2.0	0.7	-0.7	86.0
WSC	-0.3	0.0	1.3	2.9	-4.8	-3.2	3.6	4.8	2.6	-3.8	0.3	67.3
RTE	0.6	3.4	3.4	5.1	-4.3	-18.2	4.8	1.1	2.6	-2.4	3.1	83.5
MultiRC	2.4	7.9	2.6	10.1	-10.6	-8.1	6.8	2.6	1.1	-4.2	6.5	47.4
WiC	-1.3	0.1	2.5	1.7	-2.0	-1.1	0.1	2.1	-6.4	1.4	0.9	70.5
BoolQ	-0.1	0.9	0.1	1.1	-2.8	-10.6	0.7	0.0	0.9	-4.2	1.4	86.6
CSenseQA	-4.7	-1.6	-2.6	0.1	-7.8	-12.0	0.4	-5.1	-0.9	-7.6	-2.6	74.0
CosmosQA	-2.5	-0.1	-2.1	-0.4	-9.1	-6.9	-0.0	-3.0	-0.0	-8.4	-0.5	81.9
ReCoRD	-4.0	-0.0	-1.5	-0.1	-12.4	-6.1	0.2	-4.7	-0.5	-11.9	-1.6	86.0
Avg. Target	-1.8	1.9	0.2	2.6	-7.9	-9.9	2.0	-1.3	-0.4	-7.1	0.7	78.2

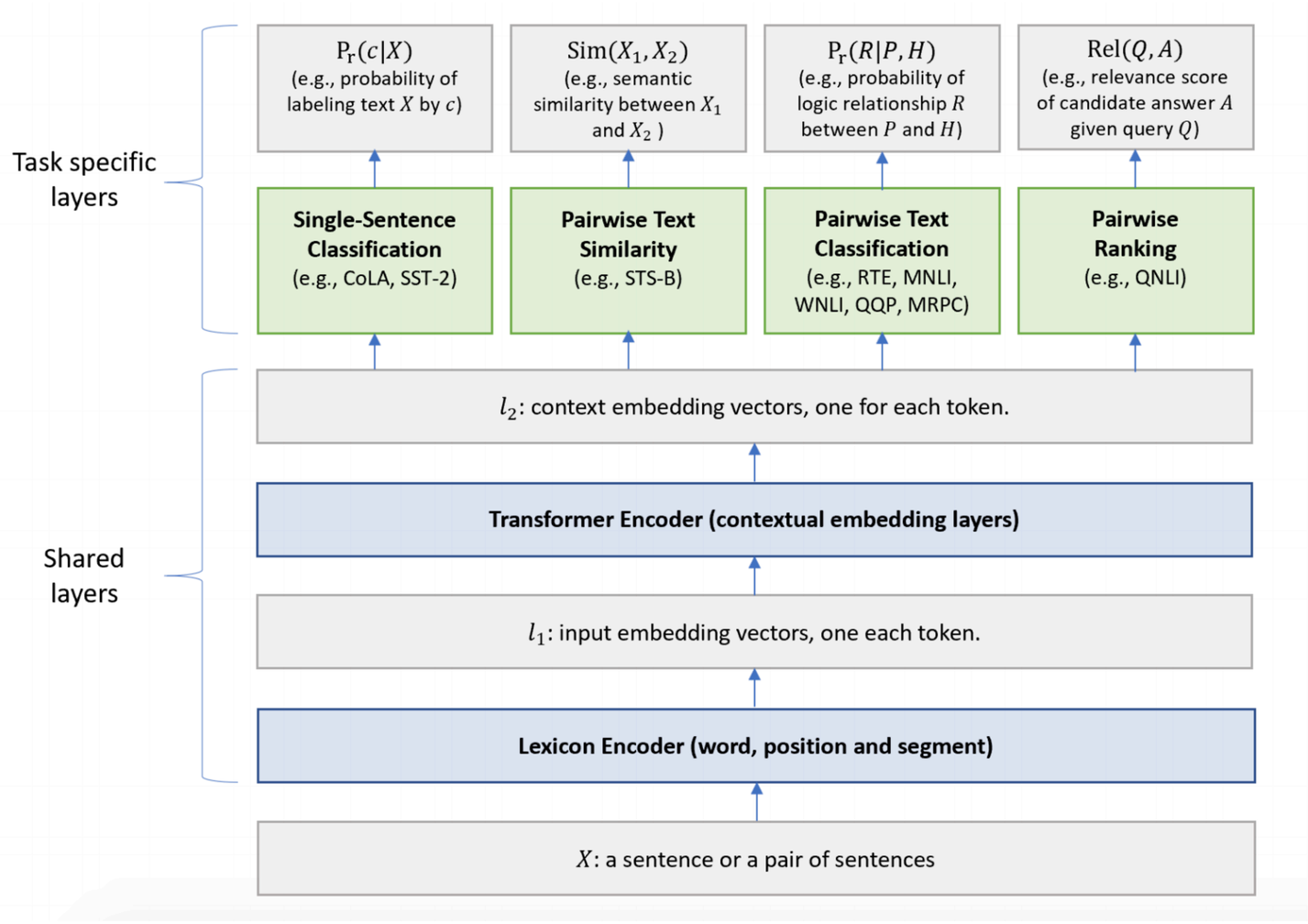
How: Intermediate-Task

- Don't Stop Pretraining: Adapt Language Models to Domains and Tasks （ACL 2020 Honorable Mention Paper)
  - Pretrain和Finetune的gap，包括data domain的gap和task的gap。
  - 作者在general预训练的基础上，加入了针对target domain的预训练和target task的预训练

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

How: Multi Tasks

- Multi-Task Deep Neural Networks for Natural Language Understanding (ACL 2019)



How: Multi Tasks

- Multi-Task Deep Neural Networks for Natural Language Understanding (ACL 2019)
  - Obtains new state-of-the-art results on ten NLU tasks

Model	CoLA 8.5k	SST-2 67k	MRPC 3.7k	STS-B 7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	**Score**
BiLSTM+ELMo+Attn [1]	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4/76.1	-	56.8	65.1	26.5	70.5
Singletask Pretrain Transformer [2]	45.4	91.3	82.3/75.7	82.0/80.0	70.3/88.5	82.1/81.4	-	56.0	53.4	29.8	72.8
GPT on STILTs [3]	47.2	93.1	87.7/83.7	85.3/84.8	70.1/88.1	80.8/80.6	-	69.1	65.1	29.4	76.9
$\text{BERT}^{4}_{\text{LARGE}}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	65.1	39.6	80.5
$\text{MT-DNN}_{\text{no-fine-tune}}$	58.9	94.6	**90.1/86.4**	89.5/88.8	**72.7/89.6**	86.5/85.8	**93.1**	79.1	65.1	39.4	81.7
MT-DNN	**62.5**	**95.6**	**91.1/88.2**	**89.5/88.8**	**72.7/89.6**	**86.7/86.0**	**93.1**	**81.4**	65.1	**40.3**	**82.7**
Human Performance	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0/92.8	91.2	93.6	95.9	-	87.1

How: Multi Tasks

- Understanding and Improving Information Transfer in Multi-Task Learning （ICLR 2020)

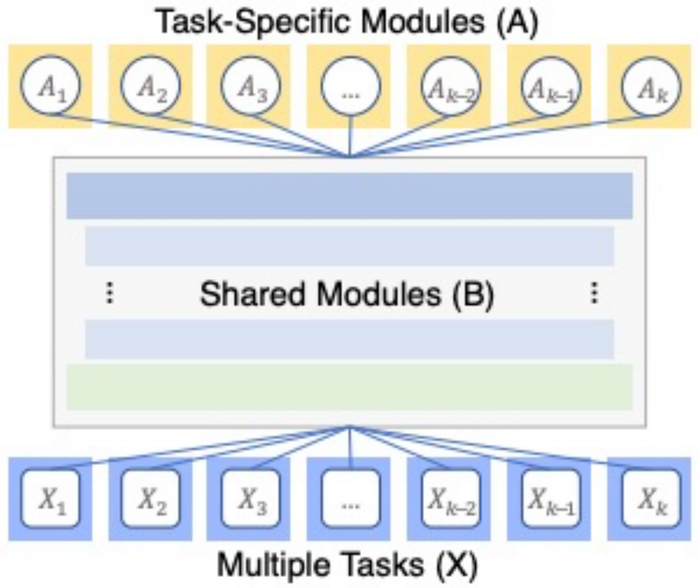
- 以通用语言理解评估GLEU和情感分类模型为benchmark

- 理论性的分析了上文的share 底层模块的multi task learning的效果的三个影响因素：

  - share模块的capacity: 如果太大，不同task就没有相互关系，就不能强迫模型学习共通可transfer的知识。如果太小，不同task就会互相干扰影响性能。

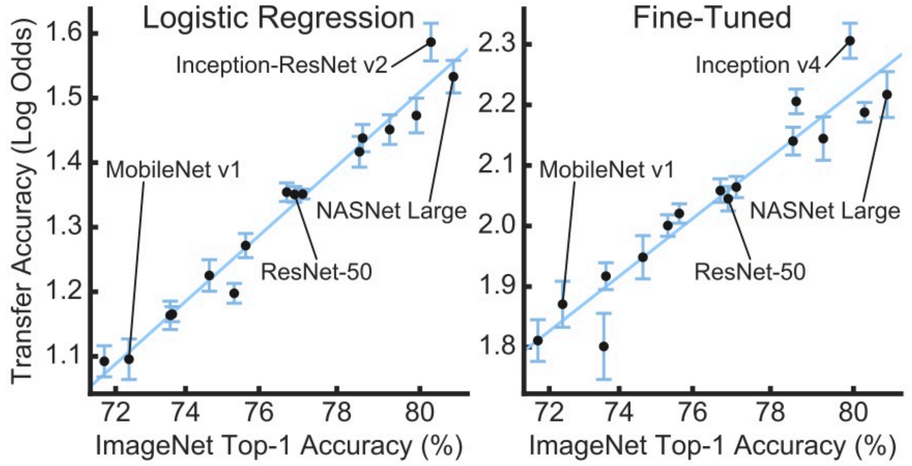
  - 任务的covarience: 任务的covariance是衡量任务数据的alignment的，
    不同的task如果太不similarity，就对迁移有副作用。

  - 训练过程中每个task的权重。



How: Early Stop

- Do Better ImageNet Models Transfer Better? (CVPR 2019)
  - 预训练任务的准确率和迁移到下游任务的性能有什么关系。
  - 在16个常用的cv模型和12个常用的数据集上做了系统的测试，
  - 1）imagenet预训练准确率和迁移任务的准确率高度正相关，无论是fine-tune还是特征提取+小分类器；
  - 2）但是有一些正则方法，如batch norm,dropoutm和lable smoothing，会影响特征提取的效果



How ?

- Rethinking the Hyperparameters for Fine-tuning (ICLR 2020)

- pretrain+finetune这个pipeline中，finetune的hyperparameter很重要。如果超参数没选好导致最终下游任务效果不好，人们可能会错误的怪罪到pretrain上。这篇文章在cv任务上检验了finetune中的超参数，发现：

  - 1）最佳超参数和数据，特别是domain similarity有关；

  - 2）learning rate，尤其是momentum，很影响性能(不能只设成0.9)，而且和domain的similarity有关,越近的domain应该用越小的momentum；

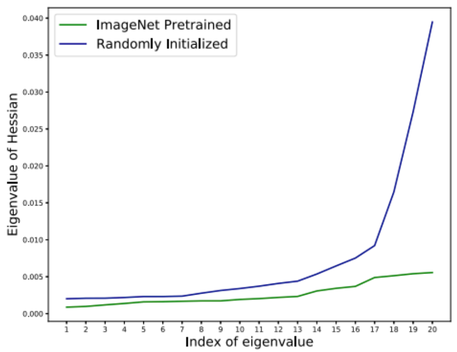
m	η	λ	Dogs	Caltech	Indoor	Birds	Cars	Aircrafts	Flowers
0.9	0.01	0.0001	17.20	14.85	23.76	18.10	9.10	17.55	**3.12**
0.9	0.01	0	17.41	14.51	24.59	18.42	9.60	**17.40**	3.33
0.9	0.005	0.0001	14.14	13.42	24.59	17.24	**9.08**	18.21	3.50
0.9	0.005	0	14.80	13.67	22.79	17.54	9.31	17.82	3.53
0	0.01	0.0001	11.00	12.11	21.14	17.41	11.07	20.58	5.48
0	0.01	0	10.87	12.16	21.29	**17.21**	10.65	20.46	5.25
0	0.005	0.0001	10.21	11.86	21.96	18.24	13.22	24.39	7.03
0	0.005	0	**10.12**	**11.61**	**20.76**	18.40	13.11	23.91	6.78

Why?

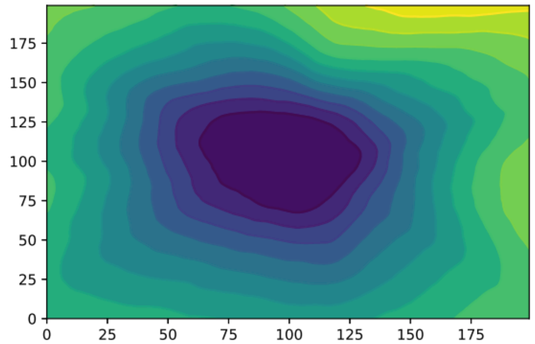
- Why Does Pre-train Help?
  - Good Initialization / Better Landscape:
    - Towards Understanding the Transferability of Deep Representations (Arxiv 2020)
    - In Neural Machine Translation, What Does Transfer Learning Transfer? (ACL 2020)
  - Better Generalization ability:
    - Pretrained Transformers Improve Out-of-Distribution Robustness (ACL 2020)

Why ?

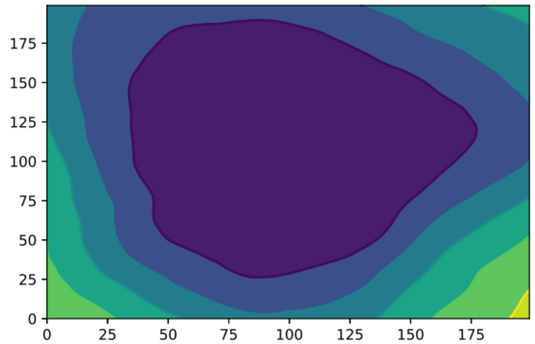
- Towards Understanding the Transferability of Deep Representations (Arxiv 2020)
  - 解释为什么pretrain model对下游任务有帮助。
    - 1）通过测train-test的差距，论证了用pre-train的模型泛化能力更强；
    - 2）通过对landscape的visualization，证明pre-train后模型的loss的landscape更加的flat；



(a) Eigenvalues of Hessian.



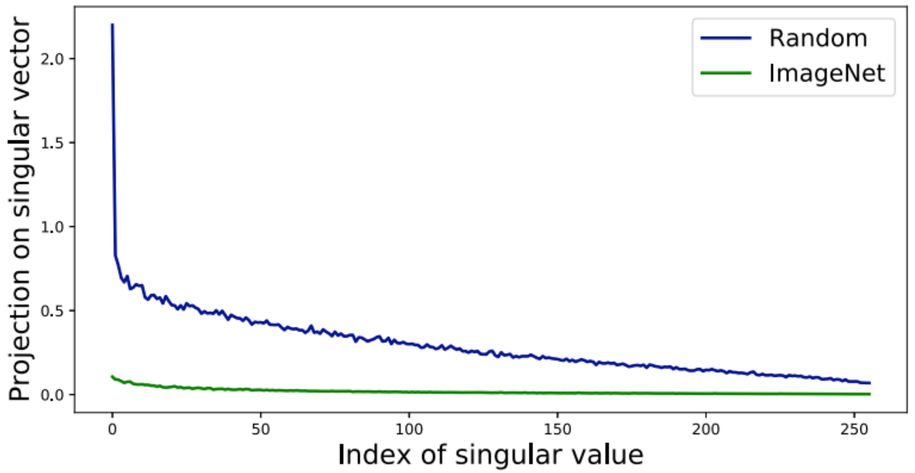
(b) Randomly initialized.



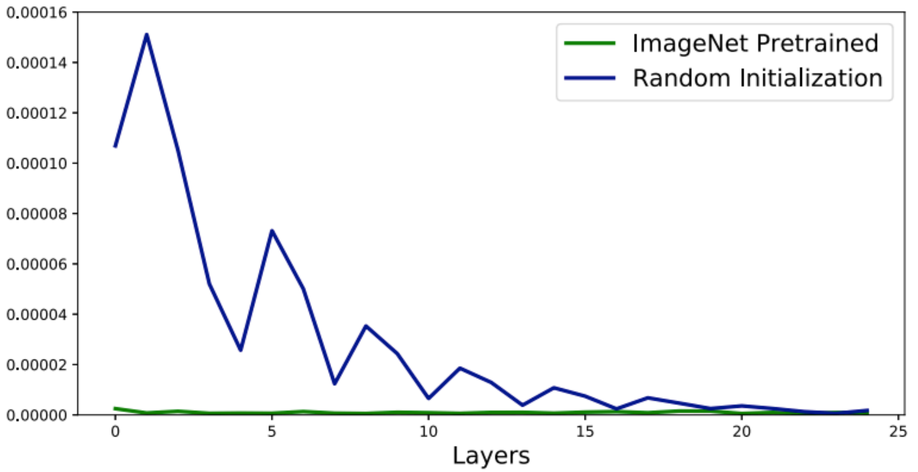
(c) ImageNet pretrained.

Why ?

- Towards Understanding the Transferability of Deep Representations (Arxiv 2020)
  - 解释为什么pretrain model对下游任务有帮助。
    - 3）沿着梯度传播的角度，发现了pre-train的模型的梯度保持的很好（layer isometry



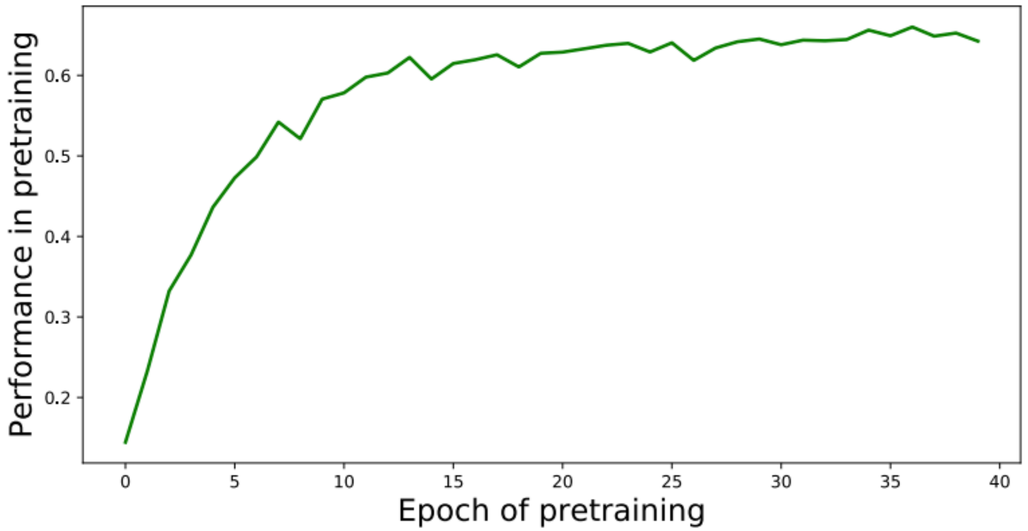
(c) Projection of weight on components of gradient.



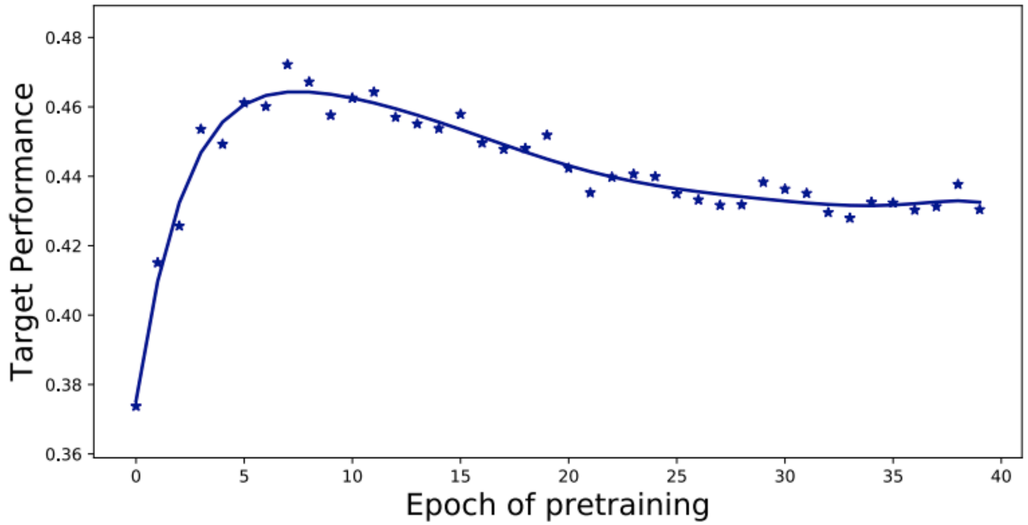
(d) Scale of gradient in different layers.

Why ?

- Towards Understanding the Transferability of Deep Representations (Arxiv 2020)
  - 解释为什么pretrain model对下游任务有帮助。
    - 4）看learning dynamic，发现特征的迁移性随epoch先增大，后减小。



(a) Accuracy of pretraining.



(b) Accuracy on the target dataset.

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)
  - 首先作者对比了不同种类和方向的parents预训练下的low resource翻译任务，发现并没有显著的区别。

	BLEU		
Parent	My→En	Id→En	Tr→En
-	4.0	20.6	19.0
En→De	17.5	27.5	20.2
En→Ru	17.8	27.4	20.3
De→En	17.3	26.3	20.1
Ru→En	17.1	26.8	20.6

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)

  - 然后作者探讨了embedding的迁移问题（因为词表不同，这是一个迁移任务的难点）。作者先对比了迁移embedding和迁移内部网络的效果，发现只迁移embedding反而会降低翻译效果，而只迁移内部网络能提升效果，都迁移能进一步的提升效果。

Transferring		BLEU						
		De→En parent			En→De parent			
Emb.	Inner	My→En	Id→En	Tr→En	My→En	Id→En	Tr→En	avg.
Y	Y	17.8	27.4	20.3	17.5	27.5	20.2	21.7
N	Y	13.6	25.3	19.4	10.8	24.9	19.3	18.3
Y	N	3.0	18.2	19.1	3.4	18.8	18.9	13.7
N	N	4.0	20.6	19.0	4.0	20.6	19.0	14.5

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)

  - 然后作者实验了几种迁移embedding方法，包括按照频率分配，按照token匹配，随机分配和共享词表。实验发现token匹配和共享词表能提升迁移效果，说明如果match对了，迁移embedding也很有用。

	BLEU						
	De→En parent			En→De parent			
Embedding	My→En	Id→En	Tr→En	My→En	Id→En	Tr→En	avg.
-	4.0	20.6	19	4.0	20.6	19	14.5
Exclude embedding	13.6	25.3	19.4	10.8	24.9	19.3	18.3
Frequency assign	14.2	24.4	19.4	13.9	24.3	19.4	19.2
Random assign	13.9	24.6	19.2	13.8	23.9	19.3	19.0
Token matching	17.8	27.4	20.3	17.5	27.5	20.2	21.7
Joint vocabulary	18.5	27.5	20.9	18.5	28.0	19.6	22.0

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)

  - 接着作者探究了预训练的模型到底迁移了什么，方法是直接用parents model不经过任何微调，直接做子语言对的翻译。发现模型在做copying。

Parent	Shared	Example
En→De	Id→En	src: Bank Mandiri bisa masuk dari mikro hingga korporasi . out: Bank Mandiri bisa memperingatkan dari cen@@ hingga korporasi . alignment: 0-0 1-1 3-3 5-5 6-6 7-7 9-2 9-4 9-8 9-9
De→En	Id→En	src: Bank Mandiri bisa masuk dari mikro hingga korporasi . out: seperti Mandiri bisa masuk a mikro hingga korporasi . alignment: 2-2 3-0 3-1 3-3 3-9 5-5 6-6 7-7 7-8 9-4

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)
  - 然后作者设计了很多花式parent 预训练集，包括copying，随即替换句子，随即输入等等。实验竟然发现所有预训练效果都比不预训练好。作者认为这是因为很多时候预训练只是为了一个更好的initialzation。

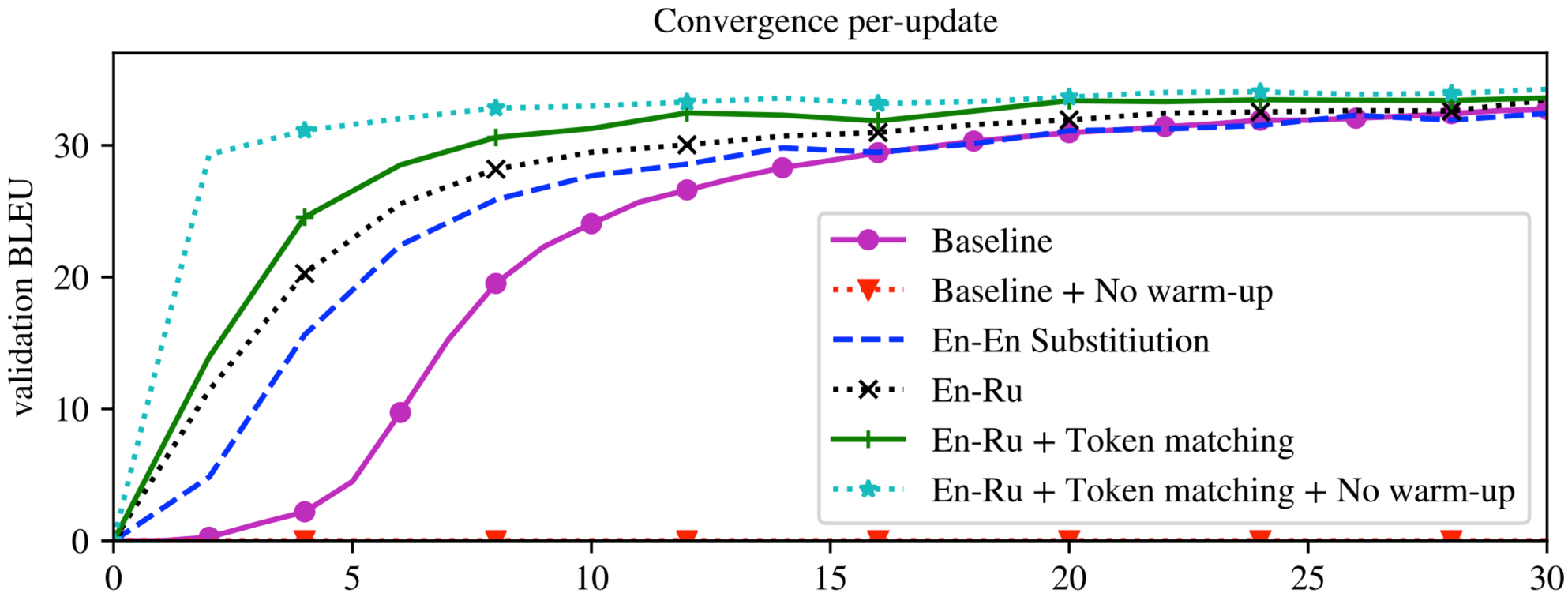
Parent	**Type**
Mono copy sequence (En→En)	**src:** Madam President , on a point of order . **tgt:** Madam President , on a point of order .
Mono substitution sequence (En_S →En)	**src:** Click write , ideologies rotate sful ECHO recommended struggle **tgt:** Madam President , on a point of order .
Mono copy sequence (Zh→Zh)	**src:** 保持点神秘感。 **tgt:** 保持点神秘感。
Mono substitution sequence (Zh_S →Zh)	**src:**比赛漂亮家宝1503 知识产权 **tgt:** 保持点神秘感。
Random copy sequence (Rand→Rand)	**src:** 1 3 2 1 1 **tgt:** 1 3 2 1 1
Random substitution sequence ($Rand_S$ →Rand)	**src:** 2 4 3 2 2 **tgt:** 1 3 2 1 1

	BLEU			
Parent	My→En	Id→En	Tr→En	Tr(10k)
-	4.0	20.6	19.0	14.3
De→En	17.8	27.4	20.3	20.2
En→En	10.4	23.3	18.5	16.0
En_S →En	12.3	23.8	19.0	16.5
Zh→Zh	8.3	22.5	18.8	16.3
Zh_S →Zh	11.2	23.5	19.0	16.3
Rnd→Rnd	6.2	21.9	19.0	15.2
Rnd_S →Rnd	7.9	22.0	19.3	15.1

Why ?

- In Neural Machine Translation, What Does Transfer Learning Transfer? （ACL 2020)

  - 最后作者为了验证这个，在高资源翻译任务上测试，发现预训练过的模型收敛更快，而且不需要learning rate warm up，但是最终能达到的效果和不预训练一样。



Why ?

- Pretrained Transformers Improve Out-of-Distribution Robustness (ACL 2020)
  - Systematically measure out-of-distribution (OOD) generalization for seven NLP tasks.
  - Measure the generalization of previous models including bag-of-words models, ConvNets, and LSTMs.
  - Pretrained Transformers' generalized better.