日本語 DeBERTa モデルの構築

植田 暢大 @京都大学

言語処理学会第29回年次大会 併設ワークショップ
日本語言語資源の構築と利用性の向上 (JLR2023) 2023/3/17

DeBERTa [He+, ICLR 2021] とは

言語理解ベンチマーク SuperGLUE [Wang+, NeurIPS 2019] で高い性能を示す

SuperGLUE GLUE

2023年3月12日時点

	Rank	Name	Model	URL	Score	BoolQ	CB
	1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2
✚	2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0
	3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6
	4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2
	5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0
✚	6	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6
✚	7	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6
	8	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9
✚	9	T5 Team - Google	T5		89.3	91.2	93.9/96.8

パラメータ数

- 6B
- 269B
- 5.4B
- 10B
- 540B
- 11B
- 1.5B x N (ensemble)

https://super.gluebenchmark.com/leaderboard

構築したモデル

モデル	ステップ数	計算資源	学習時間	単語分割	PAS解析
単語 DeBERTaV2 base	500k	A100 40GB x 8	21日	-	0.7662
文字 DeBERTaV2 base	500k	A100 40GB x 8	20日	0.9857	-
単語 DeBERTaV2 large	300k	A100 40GB x 8	36日	-	0.7887
文字 DeBERTaV2 large	260k	A100 40GB x 16	26日	0.9877	-

https://huggingface.co/ku-nlp/

- 訓練コーパス（合計142GB）
  - 日本語 Wikpedia（3.2GB, 27M文, 1.3M文書）x 10
  - 日本語 CC-100（85GB, 619M文, 66M文書）
  - 日本語 OSCAR（54GB, 326M文, 25M文書）

over sampling

- 使用したライブラリ：transformers, DeepSpeed

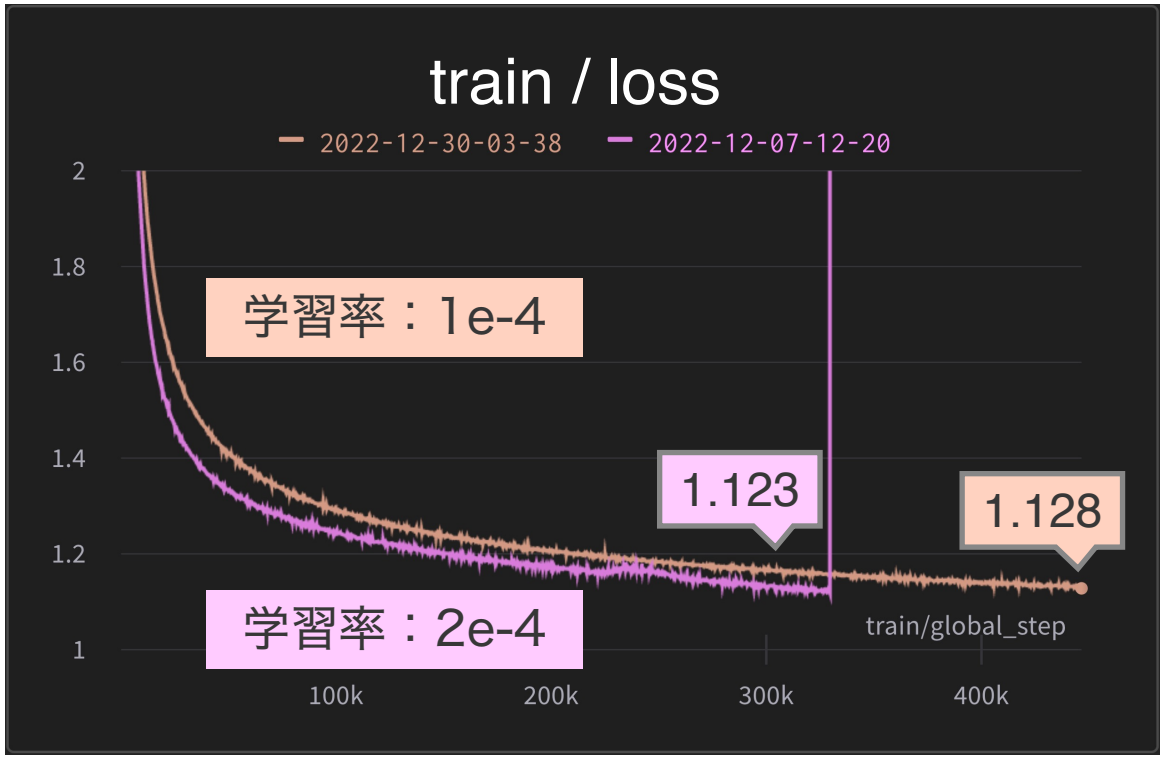
得られた知見1: DeepSpeed の効果

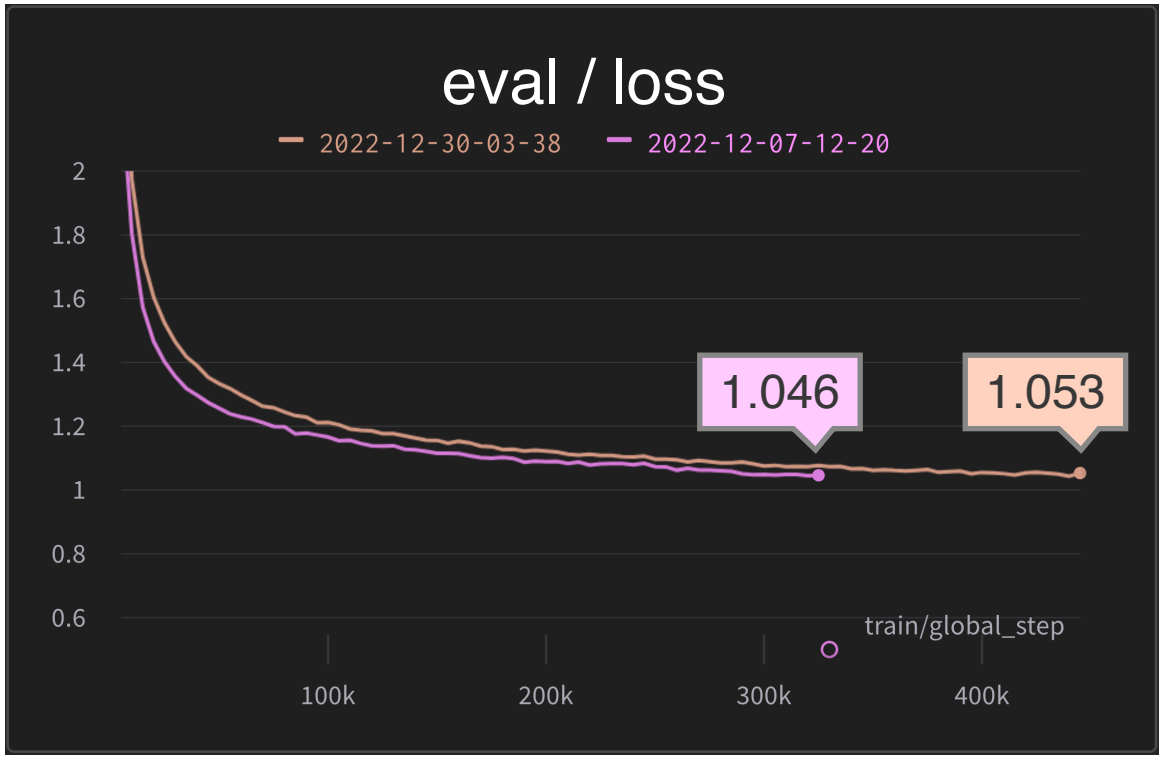
- 単語 DeBERTaV2 large で DeepSpeed のメモリ・時間効率を検証
- 設定：A100-SXM4-40GB 8枚, gradient accumulation steps = 8

設定	1ノード (8)		2ノード (4 + 4)	
	メモリ効率↑ (batch / device)	時間効率↓ (min / 1k steps)	メモリ効率↑ (batch / device)	時間効率↓ (min / 1k steps)
Naïve DDP	16	87.5	-	-
PyTorch FSDP (zero_dp_2)	18	96.2	-	-
DeepSpeed ZeRO stage 1	22	71.3	34	**90**
DeepSpeed ZeRO stage 2	22	**67.2**	36	110
DeepSpeed ZeRO stage 3	**38**	132	**38**	150

得られた知見2: 学習率について

- 学習率は損失が発散しない程度に大きいほうが良い
- 損失が発散しても学習終盤であれば良いモデルが得られる





文字DeBERTaV2 base の損失の変化