Student Cup 2022
3rd place solution

Team : UZIA

チーム紹介

チーム名 ： UZIA (ユージア)

ユーザー名 ： ktask(ケイタスク)、jt(ジェイティー)、gregley(グレグリー)

所属 ： 会津大学大学院コンピュータ理工学研究科、会津大学コンピュータ理工学部

研究 ： ヘルスケア分野の統計解析

解法

① KaggleのNLPコンペの解法を参考にBERT系のモデルをいろいろ試す
② 大量アンサンブル

- 前処理

  htmlタグの除去

  trainデータの重複データの除去

  encoding errorの除去 (フォーラムの706番目のデータについてを解決) → KaggleのFeedback Prize 4th place solutionを参考

- モデル

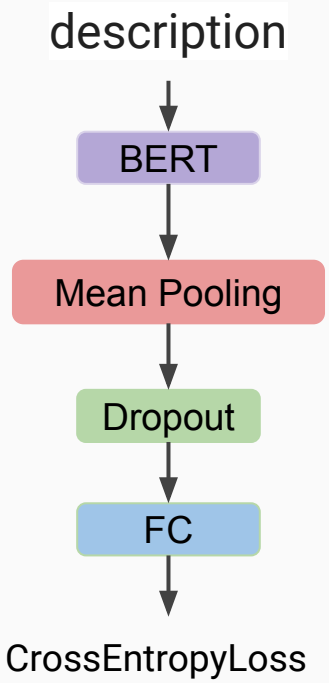
  deberta-v3-base

  deberta-v3-large

  deberta-large

  roberta-large

モデルの構築



検証方法 ： Stratified KFold (jobflag) fold=5

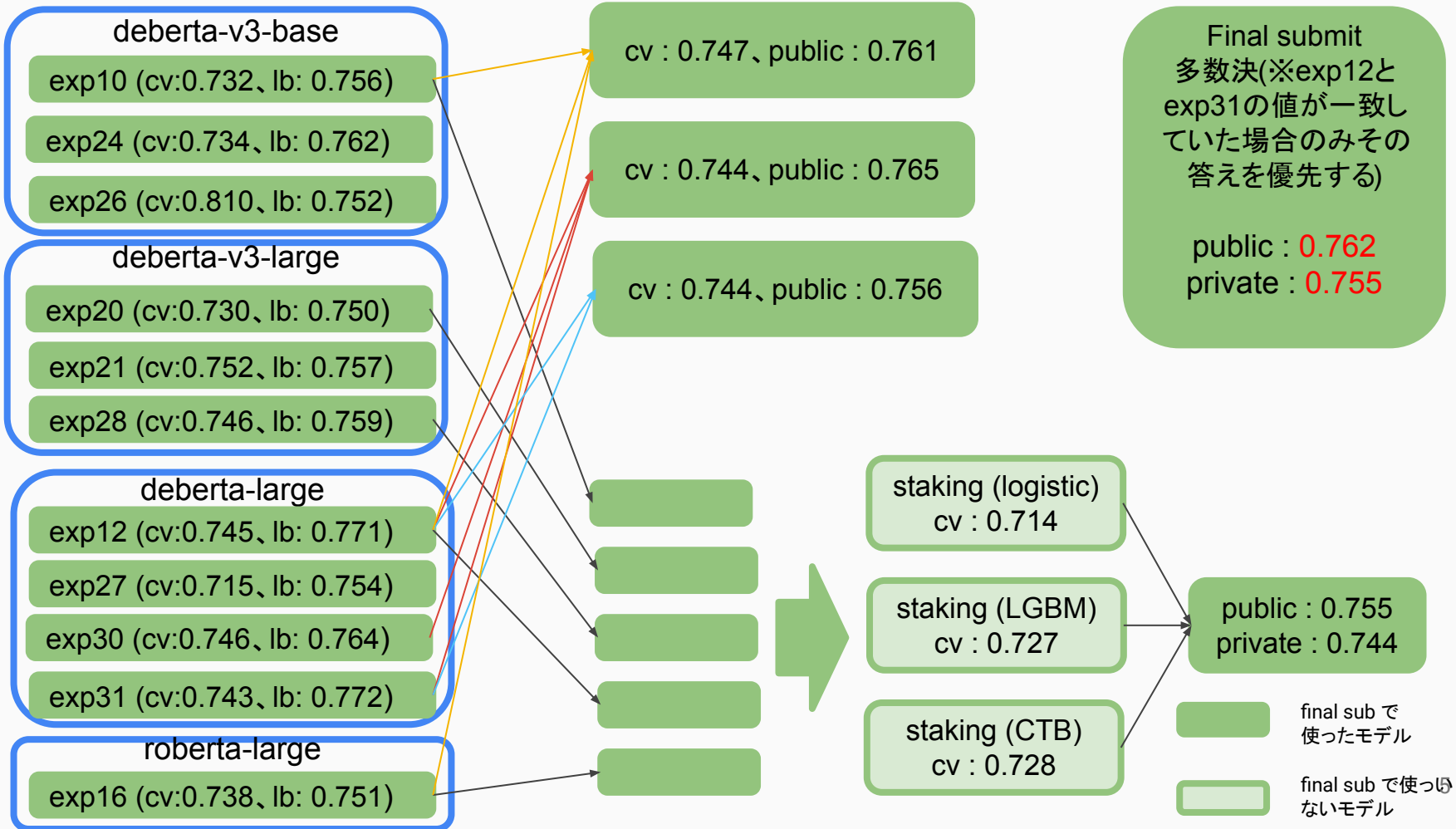
損失関数 ： CrossEntropyLoss

optimizer ： AdamW

epoch ： 10

max_len ： 1024

全体像


deberta-v3-base
exp10 (cv:0.732、lb: 0.756)
exp24 (cv:0.734、lb: 0.762)
exp26 (cv:0.810、lb: 0.752)
deberta-v3-large
exp20 (cv:0.730、lb: 0.750)
exp21 (cv:0.752、lb: 0.757)
exp28 (cv:0.746、lb: 0.759)
deberta-large
exp12 (cv:0.745、lb: 0.771)
exp27 (cv:0.715、lb: 0.754)
exp30 (cv:0.746、lb: 0.764)
exp31 (cv:0.743、lb: 0.772)
roberta-large
exp16 (cv:0.738、lb: 0.751)
cv : 0.747、public : 0.761
cv : 0.744、public : 0.765
cv : 0.744、public : 0.756
Final submit
多数決(※exp12とexp31の値が一致していた場合のみその答えを優先する)
public : 0.762
private : 0.755
staking (logistic)
cv : 0.714
staking (LGBM)
cv : 0.727
staking (CTB)
cv : 0.728
public : 0.755
private : 0.744
final sub で使ったモデル
final sub で使ったモデル

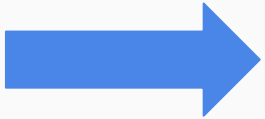
重複データを削除することで精度は変わるのか？

今回のデータセットにはtrainデータ内に12件、trainデータとtestデータに40件の重複データが含まれていた

全データ

	cv	public	private
roberta-large	0.738	0.751	0.739
deberta-v3-base	0.732	0.756	0.731
deberta-v3-large	0.752	0.757	0.743

deberta-v3-baseのみ精度が上がった



重複データを削除してもcvとlbの乖離は改善されなかった

重複削除

	cv	public	private
roberta-large	0.722	0.728	0.713
deberta-v3-base	0.734	0.762	0.747
deberta-v3-large	0.746	0.759	0.712

その他取り組んだこと

- testデータによるデータの水増し

  予測した各jobflagが0.9を超えるものを追加して学習を行う

  →あまり精度の向上はしなかったがlbが高かったものは最後のアンサンブルに使用
- 逆翻訳によるデータの水増し

  trainデータに対して英語→フランス語→英語の逆翻訳を行ったが精度が悪化した

運営の皆様、参加者の皆様

ありがとうございました！