Deep Learning 基礎講座 2024

最終課題 工夫レポート

氏　　名：松下 剛士

アカウント ID：37984

アカウント名：Mattsun_05

1　データ前処理工程における工夫点

1.1　新たなラベルの定義

訓練データを確認したところ，最頻回答（以降，mode_answer）は大きく５つのグループに分けられると考えた．”unanswerable”, “yes”, “no”, “color”, “others”である．特に，訓練データにおいて，全体の約 38 % に当たる 7,559 件が”unanswerable”であった．本タスクでは，回答の選択肢が非常に多いため，効率的に学習を進めることが重要である．そのため，後述するマルチタスク学習のためいくつくかの新たなラベルを作成した．

表 1. 新たに作成したデータラベル一覧

ラベル名（実際のカラム名）	内容
answer type ラベル(answer_label)	”unanswerable”, “yes”, “no”, “color”, “others”のいずれであるかを示すラベル．(0 ~ 4)
answerable ラベル(unans_flg)	回答が”unanswerable”かどうか. (0 or 1)
yes/no ラベル(yes_no)	回答が”yes”, “no”, その他 のいずれかを示す．(0 ~ 2)
color ラベル(color_flg)	出現頻度の多い色にラベルを付与．0 は色表現以外の回答を意味する．(0 ~ 6) white：1，grey：2，black：3，blue：4，red：5， other color (pink, green, purple, yellow, orange, gold)：6

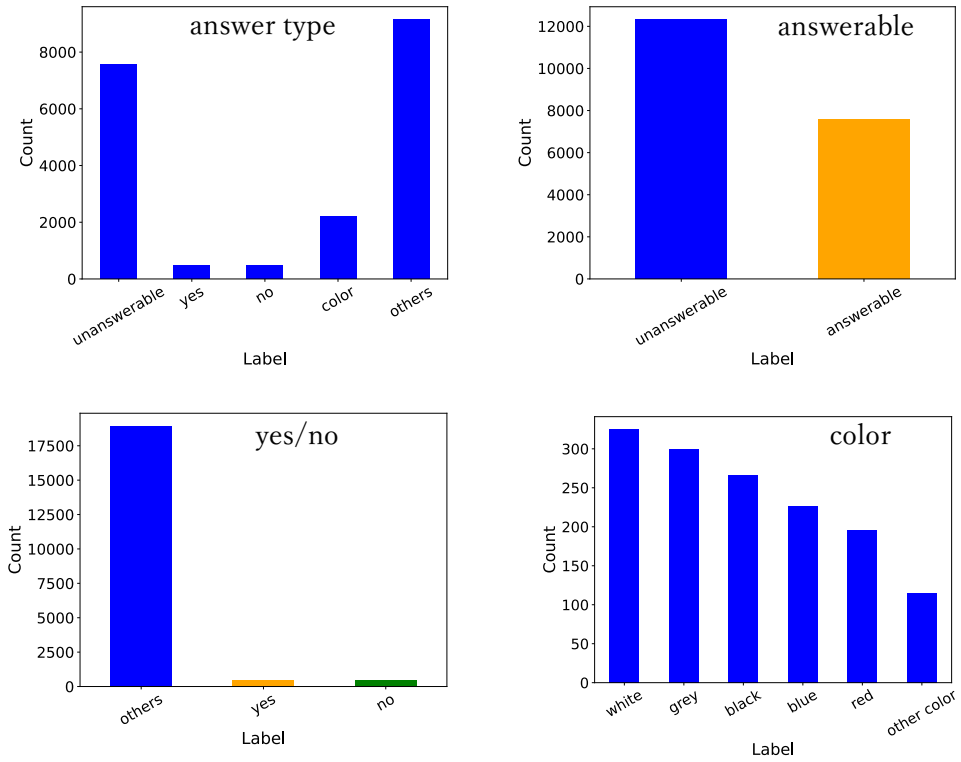

図 1. 各ラベルの出現数

1.2　質問文の前処理

ベースラインコードにおいては，質問文について前処理がなされていなかった．そこで，モデルに入力する文章の均質性を担保するために以下の処理を行った．ベースラインコードにおいて，回答文に適用されている処理 process_text と併せて，変換例を示す．

- 大文字，小文字を区別し，変換前の文章のまま保持.
- カンマ，ピリオド，クエスチョンマークを変換前の文章のまま保持し，前後に空白スペースを追加することで，1 つの単語として区別.
- 小数点は小数点のまま保持.

・ 短縮表現のカンマの追加.

・ 冠詞を削除せず，そのまま保持.

表 2. 質問文の変換例

変換前	変換後	
質問文	process_text(answer 用)	preprocess_question(question 用)
Hi, this is test sentence.	hi this is test sen10ce	Hi , this is test sentence .
Pi is 3.14.	pi is 3 14	Pi is 3.14 .
We have ten apples.	we have 10 apples	We have 10 apples .
Can you count the number of apples?	can you count number of apples	Can you count the number of apples ?
I dont know where he is.	i don't know where he is	I don't know where he is .
What is this? Foreign Language	what is this foreign language	What is this ? Foreign Language

1.3 data augmentation

本タスクの画像には，上下反転しているものや横向きになっていると考えられる画像がいくつか存在した．そのような画像を適切な向きに変換することで，より適切に学習が行われると考えた．そこで，元のデータに対して，右に 90 度回転させた画像，270 度回転させた画像を訓練データに追加した．180 度回転も実施したが，スコア向上に結び付かなかったため除外した．具体的には，transform クラスの定義において，基準となる transform_resize，90 度及び 270 度の回転を加えた transform_resize_90, transform_resize_270 を定義し，元のデータフレーム（train.json）を 3 倍に拡張した．そして，1 つのデータに対して transform_resize，transform_resize_90, transform_resize_270 を適用させ，学習を行っている．そのため，モデル学習に利用したデータサイズは 19,873 × 3 = 59,619 件である.

また，各画像データを統一したデータサイズにリサイズする際に，バイキュービック補完[1]を行い，データを高品質に保ったままリサイズを実施している.

1.4 回答選択肢の絞り込み

訓練データに登場する回答すべてをクラス分類数，つまりモデルが出力する回答のコーパスとして利用すると約 4 万通りと膨大である．この回答選択肢の中には，訓練データ中に 1 回しか登場しないような非常にまれな回答も含まれている．加えて，訓練データには存在しない回答も含まれる class_mapping を追加するとコーパスはさらに膨大になる．モデルが出力する回答の選択肢を絞ることを考えた.

class_mapping の内容を確認すると，訓練データにおける mode_answer の約 94.7 %が含まれていることが確認できた．この class_mapping には，VQA タスクにおける一般的な回答が含まれていると仮定し，class_mapping を主として，class_mapping に含まれていない訓練データの mode_answer を追加したコーパスを回答の選択肢として用いることとした．この作業により，回答選択肢は 40,835 通り → 6,392 通りと大幅に縮小させることができた.

2 モデル構築における工夫点

2.1 複数のマルチモーダルモデルの構築及びアンサンブル

本タスクにおいては，画像と文章の異なる情報源から情報を収集し回答を予測するマルチモーダルなモデルを構築した．使用したベースモデルの概要及び学習内容は以下のとおりである．なお，各モデルにおける損失関数については，2.2 節で詳説している．そして，推論時において valid データに対して TTA（Test Time Augmentation）を実施した．TTA においては訓練データと同様，回転無し，90 度回転，270 度回転である．図 2 に示すとおり，3 種の異なるマルチモーダルモデルそれぞれについて TTA を適用した計 9 つの

推論結果のスコアを平均し，最もスコアが高かったインデックスを最終予測値として出力した.

表 3. 各モデルの概要

モデル名	ベースモデル	モデルの概要
CLIP モデル	CLIP[2, 3] 【ViT-B/32】	・ epochs = 20 ・ うち，はじめ 10epoch はメインタスクの損失関数を CrossEntropyLoss, 後半 10epoch は Influence-Balanced Focal Loss[7]を利用. ・ CLIP によるエンコーディング部分の再学習は行わず，エンコーディング以降のレイヤーのみ学習 ・ 学習パラメータは約 336 万. ・ 初期値依存性を低減するため，Xe の初期値を利用. ・ 過学習抑制のため weight decay を利用.
MMBT モデル	BERT[4] transformers よりインストール ResNet 152 [5] torchvision よりインストール	・ BERT 及び ResNet152 をベースに MMBT モデルをスクラッチ ・ epochs = 20 ・ うち，はじめ 10epoch はメインタスクの損失関数を CrossEntropyLoss, 後半 10epoch は Influence-Balanced Focal Loss[7]を利用. ・ BERT・ResNet152 を含め，すべてのパラメータを再学習. ・ 学習パラメータは約 2 億 3,180 万. ・ 初期値依存性を低減するため，Xe の初期値を利用. ・ 過学習抑制のため weight decay を利用.
ViLT モデル	ViLT[6] 【vilt-b32-finetuned-vqa】	・ epochs = 5 ・ VQA タスク用にファインチューイングされたモデルを利用. ・ 左記モデルをそのまま用いるのではなく，classifier レイヤーに入力する特徴量を別途抽出し，サブタスクへの入力としても利用できるよう工夫. ・ ViLT モデルのファインチューニング部分とサブタスクレイヤーの新規学習部分を同時に学習するため，各々に個別の学習率を適用し，訓練データに過学習しないよう工夫. ・ 学習パラメータは約 1 億 2,300 万. ・ メインタスクにおける損失関数は KL タイバージェンスを利用.

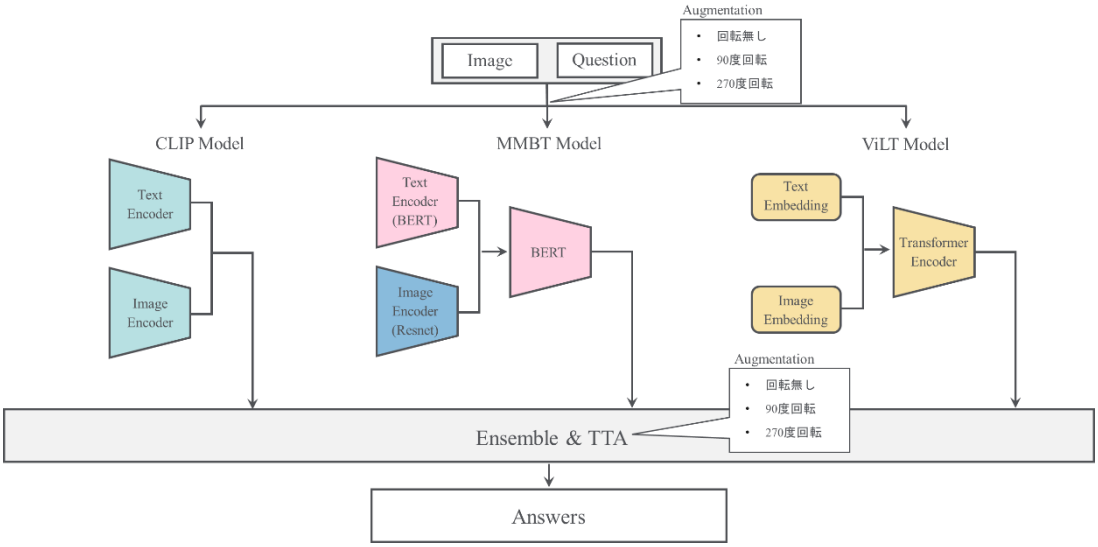

図 2. 本モデルのイメージ図

2.2 マルチタスクモデル

1.1 節で取り上げたように，本タスクでは，”unanswerable”, “yes”, “no”, color が mode_answer となる質問が多い．そこで本モデルでは，サブタスクとして，以下の 4 つの項目に関して同時に推論を行い，パラメータの学習に反映させるマルチタスクモデルを構築した．

- answer type タスク　回答が”unanswerable”, “yes”, “no”, “color”, “others”のいずれであるかを推論する 5 分類タスク．
- “unanswerable”タスク　回答が”unanswerable”かどうかの 2 値分類タスク．
- “yes” / “no” タスク　回答が”yes”, “no”, その他の 3 分類タスク．
- color タスク　回答が 1.1 節で定義したどの色に分類されるか．7 分類タスク

また， 1.1 節で作成した 5 つの answer type に反応する Mask を作成し，推論値とアダマール積をとった上で，最終的な推論値とした．

損失関数については，2 値分類タスクである，“unanswerable”タスクには BinaryCrossEntropyLoss を適用し，その他のタスクは CrossEntropyLoss を適用した．各タスクの推論による損失をメインタスクの損失に加算し，逆伝播の計算を行った．

メインタスクの損失については，CLIP モデル及び MMBT モデルについては，学習途中で CrossEntropyLoss から後述する Influence-Balanced Focal Loss に変換を行った．ViLT モデルについては，最も精度が高かった KL ダイバージェンスを用いた．

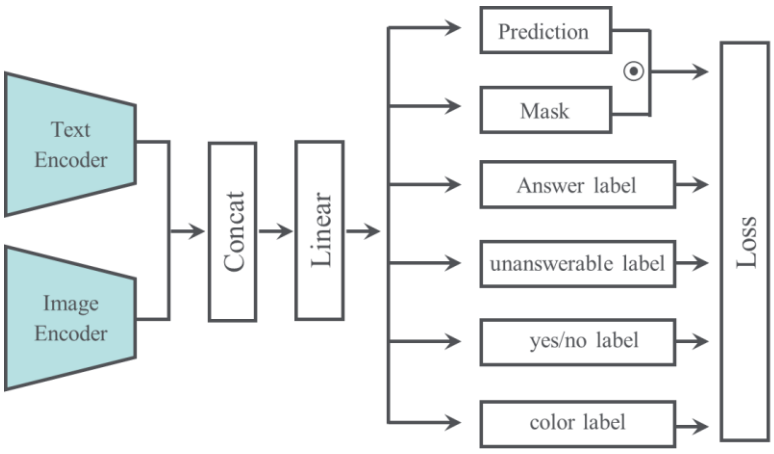

図 3. マルチタスクモデルのイメージ図

2.3 評価指標：Influence-Balanced Focal Loss

1.1 節で述べたように今回のデータでは，”unansumble”や”yes”, “no”, 色に関する回答などが多く，その他の回答は出現頻度が少ないという偏ったデータであると考えた．そのような偏ったデータに対するクラス分類タスクにおける損失関数として良い評価を得ているのが Influence-Balanced Focal Loss[7]である．通常の CrossEntropyLoss では，局所的な決定境界を過学習してしまうという問題があるのに対し，Influence-Balanced Focal Loss では，局所的な決定境界の平滑化を行うことによって，過学習を抑制し，汎化性能向上が確認されている．先に述べたように，本モデルでは，CLIP モデル及び MMBT モデルに Influence-Balanced Focal Loss を適用している．ViLT モデルについては，Influence-Balanced Focal Loss を適用したもののスコア向上に寄与しなかったので，不採用とした．

3 参考文献

[1] バイキュービック補完法，https://qiita.com/sasshi_i/items/c2356e533d1834811037.

[2] Alec Radford, Jong Wook Kim, et al., Learning transferable visual models from natural language supervision, *Proceedings of the 38th International Conference on Machine Learning*, 139:8748-8763, 2021.

[3] Fabian Deuser, Konrad Habel, et al., Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model, arXiv preprint arXiv:2206.05281 2022.

[4] Jacob Devlin, Ming-Wei Chang, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *North American Chapter of the Association for Computational Linguistics*, 4171–4186, 2018.

[5] Kaiming He, Xiangyu Zhang, et al., Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778, 2015.

[6] Wonjae Kim, Bokyung Son, Ildoo Kim, ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, *Proceedings of the 38th International Conference on Machine Learning*, 139:5583-5594, 2021.

[7] Seulki Park, Jongin Lim, et al.; Influence-Balanced Loss for Imbalanced Visual Classification, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 735-744, 2021.