

1. はじめに

CLIP は、画像とテキストを組み合わせた対照学習手法であり、大量の画像と画像に関連するテキストを用いて学習を行う。学習方法として、ミニバッチ内の画像とそれに対応するテキストを正例、それ以外を負例として対照学習を行う。学習済みの CLIP は、事前学習で使用していないデータに対する zero-shot の画像分類が可能である。一方で、CLIP の学習にはどのようなデータ拡張が精度向上に寄与するか分かっていない。そこで本研究では、CLIP のモデルを学習する時のデータ拡張が zero-shot 分類に有効であるかを調査する。

2. CLIP

CLIP は、画像とテキストのペアを用いたマルチモーダルな対照学習法の 1 つである。CLIP における損失関数を式 (1) に示す。

$$E = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^N \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \quad (1)$$

N はバッチサイズ、 $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ は正例ペアの特徴ベクトルの類似度、 $\text{sim}(\mathbf{z}_i, \mathbf{z}_k)$ は正例と負例ペアの特徴ベクトルの類似度を表す。CLIP の学習方法を図 1 に示す。CLIP では、画像と画像に対応するテキストを使用し、ミニバッチ内の画像に対応するテキストを正例、それ以外をテキストを負例として対照学習を行い、画像とテキストに共通する特徴表現を学習する。学習済みモデルを画像分類タスクに適用する場合、Image Encoder には画像を、Text Encoder には、「a photo of a [CLS]」のような文章をクラス数分入力し、画像と文章の類似度が最も高いものを分類したクラスとする。CLIP には、事前学習時にどのようなデータ拡張をすることで精度が向上するか明確ではないという問題がある。そこで、本研究では異なるデータ拡張を画像に対して適用し、効果的なデータ拡張方法について調査する。

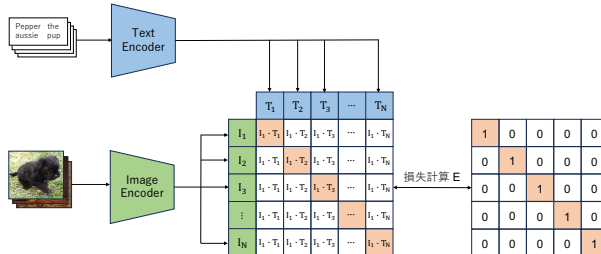
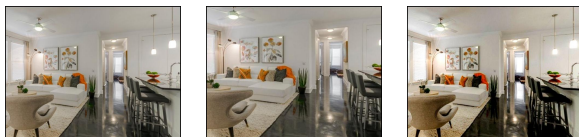


図 1: CLIP の学習方法

3. 実験概要

本研究では、データ拡張に RandomResizedCrop (RRC) とコントラスト変換を使用して CLIP の事前学習を行う。本実験で用いる RRC は画像の 90% から 100% の領域をランダムで切り抜き、コントラスト変換は画像のコントラスト値を -30% から 30% の間でランダムに変化させる。データ拡張を適用する前の画像を図 2(a)、RRC を元画像に適用した例を図 2(b)、コントラスト変換を元画像に適用した例を図 2(c) に示す。図 2 では、実験で用いるより強いデータ拡張を適用し、元画像との変化を明確にする。



(a) 元の画像 (b) RRC (c) コントラスト変換

図 2: データ拡張後の画像

モデルには ResNet-50、学習用データセットには、Conceptual 12M (CC12M)[2]、評価用データセットには ImageNet-1K (IN-1K)、CIFAR-10 (CI-10)、CIFAR-100 (CI-100)

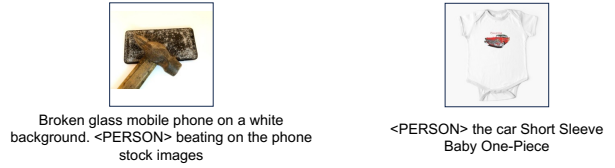


図 3: CC12M データセット

を使用する。CC12M データセットの画像とテキストの例を図 3 に示す。CC12M データセットでは、Google Cloud Natural Language API を用いてテキストに含まれる人物に関する情報を (PERSON) に置換する。評価タスクは zero-shot 分類である。学習時に加えるデータ拡張の種類とエポック数に応じてデータ拡張の種類を変化させた場合の精度変化を調査する。また、学習時のバッチサイズは 256、エポック数は 32、学習率は 0.0001 である。

3.1 評価結果

データ拡張の種類による zero-shot 画像分類の精度変化を表 1 に示す。表 1 より、RRC やコントラスト変換を適用した学習は、データ拡張なしと比べて精度が向上しており、CLIP においてデータ拡張が有効だとわかる。また、データ拡張の種類により精度が異なり、RRC が効果的であることがわかる。

表 1: データ拡張の種類による精度比較 [%]

データ拡張		評価用データセット		
RRC	コントラスト変換	IN-1K	C-10	C-100
✓		27.3	29.0	12.1
		31.1	48.5	15.5
	✓	29.3	40.6	12.0

次に、RRC とコントラスト変換の組み合わせによる zero-shot 画像分類の精度変化を表 2 に示す。表 2 より、RRC とコントラスト変換の両方を適用した学習は、RRC のみの適用と比べて精度が低い。一方で、コントラスト変換を 16epoch 目から導入すると、0epoch 目から導入した場合と比べて CIFAR-10 以外の精度が低下する。これは、複数のデータ拡張を組み合わせることで困難な学習となり、精度が低下したと考えられる。そのため、データ拡張の強度や適用のタイミングを学習状況に合わせて動的に変更する等の処理が必要である。

表 2: エポック数に応じてデータ拡張を変更した場合の精度

データ拡張		評価用データセット		
RRC	コントラスト変換	IN-1K	CI-10	CI-100
✓		31.1	48.5	15.5
✓	0epoch から導入	30.5	38.1	17.4
✓	16epoch から導入	29.9	44.6	12.9
✓	24epoch から導入	30.4	12.9	12.1

4. おわりに

本稿では、CLIP におけるデータ拡張による精度変化について調査した。データ拡張として RRC を用いることで、データ拡張を適用しない場合と比較して精度が向上し、CLIP においてデータ拡張が有効であると分かった。しかし、データ拡張を加えるタイミングにより精度が著しく変わるため、基準を設けてデータ拡張の強度を変えていく必要があると考えられる。今後は、他の種類のデータ拡張の検証や、パラメータ探索により決定したデータ拡張の強度での事前学習、学習時の誤差に応じてデータ拡張の強度や確率を変更するカリキュラム学習に取り組み、精度向上を目指す。

参考文献

- [1] Alec Radford, *et al.* "Learning Transferable Visual Models From Natural Language Supervision", ICML, 2021.
- [2] Soravit Changpinyo, *et al.* "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts", CVPR, 2021.