CH-002

詳細画像分類における Contrastive Learning の活用
Improving accuracy of Fine-grained Classification using Contrastive Learning

大城 慶知[1)]　遠藤 聡志[2)]
Yoshitomo Oshiro　Satoshi Endo

1　はじめに

画像分類は一般画像分類と詳細画像分類に大別される．詳細画像分類とはエントリーカテゴリの中で，クラス間の細かい違いを持つサブカテゴリを分類する問題のことを指す．エントリーカテゴリが車の場合に，サブカテゴリのメーカーや車種を推定するタスクであり，こういったサブカテゴリの推定は多数存在する．これらの問題は専門家にしか扱うことができず，機械学習モデルによる代替が期待される．また，高コストな Part アノテーションを利用しない手法が求められる．詳細画像分類のデータセットの一つである CUB-200-2011[4] では嘴・羽・胸に Part アノテーションが付与されている．Part アノテーションのような Part を検出するモデルを作成し，分類に必要な特徴量として抽出する Part-base は主流な手法の一つである．Part-base の例である，Korsch ら [2] の DeepFVE や Ge ら [7] の LSTM モデルでは高い分類精度が確認されている．一方，これらの手法ではアーキテクチャが複雑になりモデルごとにチューニングを行う必要がある．そこで比較的シンプルなアーキテクチャで，細かな特徴量を取得できる手法として Khosla ら [3] の Supervised Contrastive Learning(SupCon) に着目した．SupCon の研究では，一般画像分類でラベルを用いて対照学習を行うことで分類に必要な表現を獲得し分類精度が向上することを示している．詳細画像分類においても，同じラベルを持つ画像が類似した Part を持つよう特徴表現を獲得することは有効であると考えた．しかし，詳細画像分類の CUB-200-2011 データセットでは，解像度を維持するため入力の画像サイズが 448x448 で学習する．そのため SupCon は十分な計算リソースがない場合に，batch size をクラス数より小さく設定する必要がある，よって，サンプル数を確保できずに適切な学習が行えない．本稿ではこの問題を解決するために SupCon のアーキテクチャを詳細画像分類に応用する．応用にあたって，損失関数の変更が必要となる．また，疑似特徴表現の Proxy を追加することを提案しサンプル数確保の問題に対応する．これらを通して詳細画像分類での対照学習の有効性を示す．

2　詳細画像分類の動向

詳細画像分類はクラス間の細かな特徴量の違いからラベルを推定する必要があるため，分類に重要な部分の特徴量をモデルに与える必要がある．この問題に対処すべく，分類の判断材料となる部分 (Part) を抽出し，Part の特徴量を用いて分類を行う Part-base の手法が挙げられる．初期の研究では Part アノテーションを用いて Part 抽出モデルを作成し分類に利用した [8]．しかし，Part アノテーションはコストが高く，データセットによっては Part アノテーションが存在しないこともあり，画像ラベルのみを用いた半教師ありや教師なしでの Part 推定が注目されていった．Korsch ら [2] は，逆伝播の勾配計算を用いて重要性の高い画像領域を Part として推定し，Part を Fisher Vector Encoding で固定長の特徴量に変換するアーキテクチャを提案した．Ge ら [7] は Mask R-CNN と CRF に基づく Segmentation を相互に学習して，識別対象を複数枚抽出して LSTM で分類するアーキテクチャを提案した．しかし，これらの研究は潜在的な領域を抽出するモジュールを必要とするため，パイプラインが複雑になり学習コストが高くなる．他のアプローチとして，Chang ら [6] はチャネル間で相互情報量を向上するように特徴表現に損失を与えた．Zhuang ら [9] はペアによる入力バッチを構築し，ペア間で手がかりとなる識別情報を特徴表現に学習させるアーキテクチャを提案した．最近の He ら [1] による研究では，Vision Transformer[10](ViT) を用いて分類に有効なパッチを選択し，対照損失を利用して特徴表現間の距離を遠ざけることで SoTA を獲得している．He ら [1] は ViT の事前学習済みモデルで学習した結果，90.3%のベースラインとなる精度が得られた．次に，分類に重要なパッチを選択するモジュールを追加することで 91.0%の精度が得られた．最後に，対照損失を追加することで 91.7%まで精度向上した．このように He らの研究では対照損失の有効性を示しているが，特徴表現間の距離を遠ざける役割で利用されている．本稿では近年注目を挙げている対照学習を用いて，特徴表現間の距離を近づける役割も考慮して学習を行う．

3　要素技術

3.1　対照学習

対照学習とはモデルの特徴表現同士の距離を計算してラベルに対応する類似度を獲得する表現学習手法の一種である．距離の計算はコサイン類似度で行われ，モデルの特徴表現に対し L2 正則化を行い，内積を計算することで実装されている．コサイン類似度は $1 <= CosSim(A, B) <= -1$ の値を取るため，ソフトマックス関数を通して，Cross-Entropy を計算することで，類似度による多クラス分類問題として扱うことができる．アーキテクチャは Encoder Network と呼ばれる画像の特徴抽出を目的とした関数 $E(\cdot)$ と Projection Network と呼ばれる全結合層のニューラルネットワーク $P(\cdot)$ で構成される．Projection Network は，距離学習を行うために指定した次元数をもつ特徴表現に変換するためのネットワークとして機能する．上記の設定で距離学習を行ったあと，Projection を破棄し Encoder の出力に Logit 層を追

1）琉球大学大学院理工学研究科情報工学専攻,Graduate School of Engineering and Science, University of the Ryukyus
2）琉球大学工学部工学科知能情報コース,Computer Science and Intelligent Systems, University of the Ryukyus

加することで本タスクを解くアーキテクチャである．

4 先行研究

4.1 Supervised Contrastive Learning

Supervised Contrastive Learning[3](SupCon) とは教師ラベルを用いて対照学習を行ったモデルである．既存の対照学習の損失に変更を加えてバッチ内の特徴表現の距離計算に対応し，Cross-Entropy よりも優れた表現を獲得した．

$$
\begin{aligned}
L^{sup} &= \frac{1}{2N}\sum_{i=1}^{2N}\frac{-1}{2N_{y_i}-1}\sum_{j=1}^{2N}L_{i,j}^{sup} \\
L_{i,j}^{sup} &= \mathbb{1}_{i\neq j}\cdot\mathbb{1}_{y_i=y_j}log\frac{exp(z_i\cdot z_j/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{i\neq k}\cdot exp(z_i\cdot z_k/\tau)}
\end{aligned}
\tag{1}
$$

式.1 に示した損失を適応することで，全てのラベルの組み合わせで距離の計算を可能にしている．式.1 について，τ は temperature と呼ばれ，コサイン類似度のスケーリングを行うハイパーパラメータである．コサイン類似度は 1 <= $CosSim(A,B)$ <= −1 と小さい値を取るため，τ で割ることで値を大きくし学習しやすくする働きがある．また，N は batch size を表しており，別々の Augmentation を掛けた N を 2 つ用意して $N+N=2N$ とすることで，必ずポジティブ (同じラベルを持つサンプル) が存在し計算可能にしている．

Dataset	Cross-Entropy	SupCon
CIFAR-10	95.0	**96.0**
CIFAR-100	75.3	**76.5**
ImageNet	78.2	**78.7**

表 1 SupCon:Top-1 Accuracy(論文:[3] からの引用)

この損失を適応した結果，表 1 のような結果が得られた．また，Augmentation・Optimizer や Learning Rate のハイパーパラメータをそれぞれ変更した実験を行い，Cross-Entropy と比較してハイパーパラメータに対するロバスト性が高いことも示している．

5 提案手法

SupCon は分類に重要な表現を得るために，同じラベルを持つ画像間の特徴表現を近づけ，違うラベルを持つ場合は遠ざけるように学習した．詳細画像分類においても，バッチ内で特徴表現を比較することで有効な特徴量が抽出できると仮定して SupCon を活用する．しかし，SupCon を適用する場合に以下 2 つの問題点があると考える．

1. 損失がポジティブサンプル数で変動しやすい
2. 小さい batch size 下で動作しない

以下のセクションで，それぞれの問題点を詳しく説明し対応した改善案を提案する．

5.1 損失関数の変更

Prannay らによって提案された SupCon では式.1 のように損失関数を定義している．しかし，この損失はマルチラベル分類に対してマルチクラス分類の損失をそのまま適応しているため．ポジティブの個数による損失の影響が大きくなってしまい学習が安定しないことがある．式.2 に示した教師あり対照学習の損失関数を提案する．式.2 では分母にネガティブサンプル (異なるラベルを持つサンプル) の合計とアンカー (距離を計算するときに基準となるサンプル) のみにすることで，ポジティブが複数存在していてもマルチクラス分類として処理することができる．

$$
\begin{aligned}
L^{supv2} &= \frac{1}{2N}\sum_{i=1}^{2N}\frac{-1}{2N_{y_i}-1}\sum_{j=1}^{2N}L_{i,j}^{supv2} \\
L_{i,j}^{supv2} &= \mathbb{1}_{i\neq j}\cdot\mathbb{1}_{y_i=y_j}log\frac{exp(z_i\cdot z_j/\tau)}{exp(z_i\cdot z_j/\tau)+\sum_{k=1}^{2N}\mathbb{1}_{i\neq j\neq k}\cdot exp(z_i\cdot z_k/\tau)}
\end{aligned}
\tag{2}
$$

この変更により損失にどのような違いが生まれるかを式.3 に例示する．

$$
\begin{aligned}
Label(i) &= [0,0,1,1] \\
z(i) &= [[0.6,0.4],[0.6,0.4],[-0.6,0.4],[-0.6,0.4]] \\
Sim(z(0),z(2)) &= 0.3846 \\
L^{sup} &= 0.0061 \\
L^{supv2} &= 0.0061
\end{aligned}
\tag{3}
$$

式.3 では 2 つのベクトルとラベルを用意し，それぞれに適切なラベルを振られているため極小の損失を取る．

この状態に対して $Label$ と z にそれぞれ $Label(0)$ と $z(0)$ を追加していくと表 2 のような損失が得られる．表 2 からは，同じベクトルとラベルを追加しているにも関わらず損失が大きくなっている．このことから，従来の損失ではバッチごとにポジティブの個数が変わると学習が安定しなくなる，バッチ内にポジティブの個数が多いと学習が遅くなる，などの悪影響がが考えられる．

追加した個数	SupConLoss	SupConLossv2
0 個	0.0061	0.0061
1 個	0.5996	0.0073
2 個	1.0517	0.0081
3 個	1.4200	0.0086

表 2 ポジティブの個数と損失

5.2 Embedding Layer による Proxy の追加

詳細画像分類は細かい特徴量を失わないために，448x448 と比較的大きな入力画像サイズを取る．そのためモデルやリソースにも影響されるが，batch size がクラス数より小さい計算例が多くある．SupCon では batch size が小さくなると，比較サンプルが少なくなるため精度が下がることが示されている．そこで batch size をクラス数より小さく設定した際に，Embedding Layer によって疑似的なサンプルを設けることを提案する．図 1 に，Cross-Entropy，SupCon と提案手法のタスクとしての違いをトランプの suit をクラスとして用いて例示した．一般的な Cross-Entropy（左）では，入力された画像に対してラベルを推定するというタスクである．それと比較して，Khosla らが提案した Supervised Contrastive Learning(中央左) はバッチ間で同じラベルが付いている

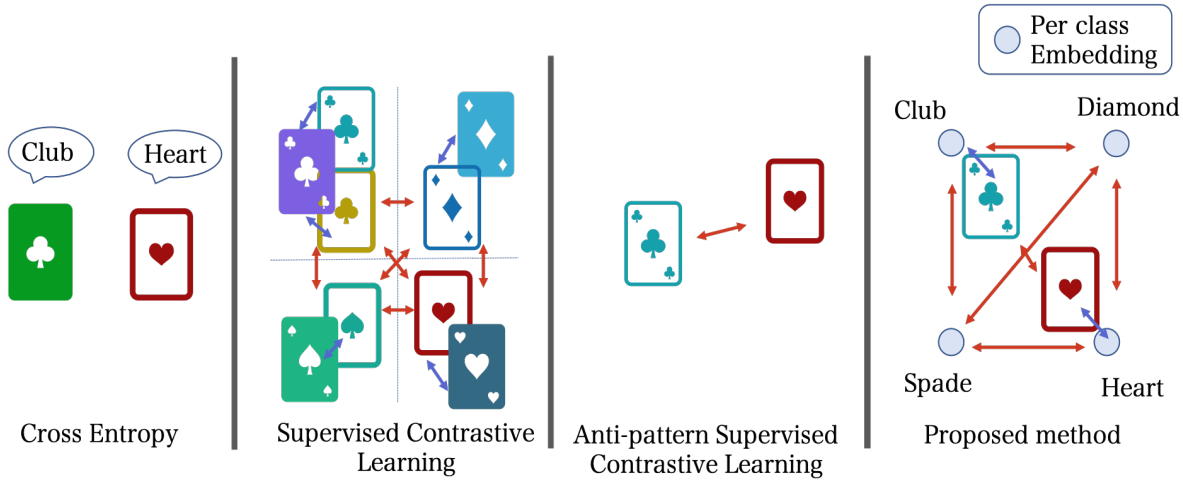

図 1　トランプを例にしたそれぞれのタスク．左が一般的な Cross-Entropy，中央左が教師あり対照学習でラベルを使って類似度を近づけている．赤の矢印は遠ざかるように，青の矢印は近づくように学習する．中央右は教師あり対照学習でのアンチパターンで，クラス数よりサンプルが少なくて学習ができない．提案手法では Proxy というクラスごとに Embedding を用意することで，学習可能にしている．※ Cross-Entropy 以外は表現学習．

ものは近づくように，違うラベルが付いているものは遠ざかるように学習する．しかし，中央右に示したように batch size が少なく，十分なネガティブサンプルやポジティブサンプルが得られない場合に損失が計算できないため学習ができない．そこで Embedding Layer を追加することで，擬似的にサンプリング対象を作成し学習することが可能となる．これは Deep Metric Learning の ProxyNCA[5] で提案されており，有効性が示されている．以降追加した Embedding Layer によるサンプルを Proxy と呼称する．損失関数を変更し，Proxy を追加した最終的な提案モデルを図.2 に示した．

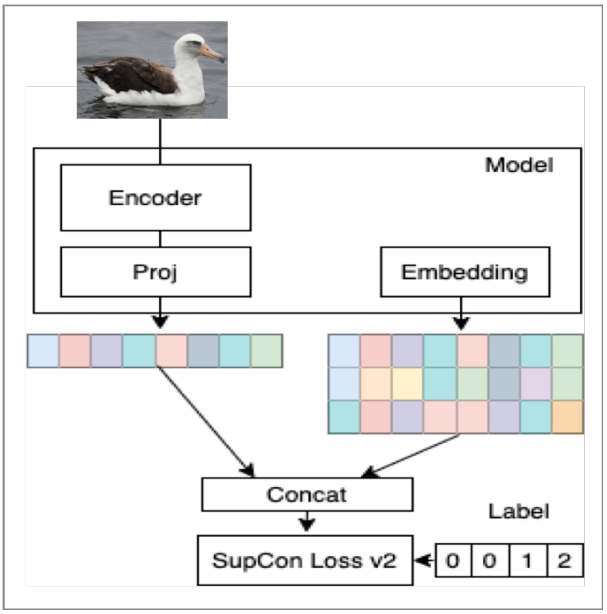

図 2　提案モデル

6　実験

本実験では提案モデルの学習及び精度評価を行い，提案モデルの有用性を検証することを目標とし，有用性を 3 段階に分けて検証した．

1. CIFAR-10, CIFAR-100 を用いて提案した損失関数と従来の損失関数を精度を用いて比較する

2. CIFAR-100 で batch-size=16 に設定した場合での，Proxy の追加したことによる精度を比較する．

3. CUB-200-2011 で提案モデルを精度による評価を行う

1,2 に関しては SupCon との比較を行うため，有志が再現したソースコードをもとに ResNet モデルを構築している．また，3 は GPU メモリの都合上 Pytorch の ResNet モデルを利用している．

6.1　データセット

使用した CIFAR-10, CIFAR-100, CUB-200-2011 のデータセットの内訳を 3 に示した．データかさ増しは，CUB-200-2011 に揃えるために Random Crop と Random Flip のみを適応した．

dataset	Training data	Test data
CIFAR-10	50000	10000
CIFAR-100	50000	10000
CUB-200-2011	5994	5794

表 3　データセット

6.2　実験環境

使用した SGD Optimizer は lr= 0.2, momentum= 0.9, weight_decay= $1e-4$, lr_decay_rate= 0.1 とした．先行研究と同様に temperature は 0.1，Embedding の出力は 128 次元と設定した．Proxy の Embedding はランダムに初期化されたものを利用した．また，CrossEntropy は 300epoch に設定し，SupCon や提案手法は表現学習に 200epoch, 追加した logit 層の学習は 100epoch と合計 300epoch で実験を行った．1. の実験では Batch size を 256 に設定し，2, 3. の実験では Batch size=16 と設定し，最終的な精度をもとに評価する．

6.3 結果と考察

6.3.1 提案した対照損失の有効性

Loss	CIFAR-10	CIFAR-100
Cross-Entropy	91.6	71.84
SupCon 2N	94.50	75.51
SupConv2(ours) N	94.40	75.57
SupConv2(ours) 2N	**94.77**	**75.67**

表 4 batch size=256 における CIFAR-10 と CIFAR-100 の実験結果

表 4 に Batch size=256 における従来の SupCon と提案損失の比較を示した．従来の SupCon では，Batch size=N とおいたときに損失が計算できなくなるため，別々の Augmenataiton がかかっているバッチを 2 つ入力する (図の 2N 表記). それぞれと比較するため，提案手法は N と 2N の両方の結果を示して比較している．Cross-Entropy と提案手法を比較すると，CIFAR-10 と CIFAR-100 のそれぞれにおいて，3.1%と 3.73%の精度向上を達成している．batch size=256 においては，従来の SupCon よりも提案手法のほうがわずかに優れた精度が得られた．

6.3.2 Proxy の有効性

Loss	CIFAR-100
Cross-Entropy	73.16
SupConv2 2N	70.74
SupConv2+Proxy(ours)	**76.22**

表 5 batch size=16 における CIFAR-100 の実験結果

表 5 に batch size=16 における CIFAR-100 の実験結果を示す．Cross-Entropy と SupConv2 を比較すると精度が 2.4%下がっている．batchsize が小さいことにより，比較するサンプル数が減り，十分な表現が得られなかったためと考えられる．提案手法の SupConv2+Proxy を見てみると，Cross-Entropy と比較して 3.06%の精度向上を達成している．これにより，Proxy が小さいバッチサイズ下での SupCon に有効であると考えられる．

6.3.3 CUB-200-2011 での実験

Loss	Acc@1	Acc@5
Cross-Entropy	69.73	87.69
Cross-Entropy + con loss[1]	67.0	85.71
SupCon 2N	5.42	17.28
SupConv2(ours) 2N	4.82	15.74
SupConv2+Proxy(ours)	**77.15**	**91.79**

表 6 CUB-200-2011 の実験結果

表 6 に CUB-200-2011 での実験結果を示した．比較対象として，TransFG[1] で利用されている対照損失を用意した．ResNet18 の pretrain モデルを利用して，batchsize=16 で学習した．その結果，提案手法は従来の Cross-Entropy と比較して，Top1 Accuracy が 7.42%精度向上することが確認できた．TransFG[1] の対照損失を追加したモデルは，今回は Cross-Entropy よりも精度向上が確認できなかった．ViT と ResNet モデルでは，全く違うためそれぞれの違いから対照損失が有効に働く要素が別にあると考えている．また，従来手法の SupCon と SupConv2 では，Proxy なしで実行した場合に分類に必要な表現を全く獲得できてないことがわかった．このことから小さいバッチ下で対照学習を利用する際には，Proxy が必要であると言える．

7 今後の展望

本研究では詳細画像分類における対照学習の有効性に着目し，SupCon のアーキテクチャに対し損失関数の変更と Proxy を追加することを提案した．実験の結果 CUB-200-2011 データセットにおいて ResNet18 モデルで精度向上することを確認した．今後の展望として Stanford Cars や NABirds データセットでの精度検証や ResNet 以外のモデルで提案した対照学習を行うことを検討する．また，提案した対照損失はポジティブサンプルが多い場合により適した損失を与えられることを考えているため，Sampler を用いてサンプルを調整した場合の検証を行う必要があると考えている．

参考文献

[1] He, Ju, et al. ”TransFG: A Transformer Architecture for Fine-grained Recognition.” arXiv preprint arXiv:2103.07976 (2021).

[2] Korsch, Dimitri, Paul Bodesheim, and Joachim Denzler. ”End-to-end Learning of a Fisher Vector Encoding for Part Features in Fine-grained Recognition.” arXiv preprint arXiv:2007.02080 (2020).

[3] Khosla, Prannay, et al. ”Supervised contrastive learning.” arXiv preprint arXiv:2004.11362 (2020).

[4] Wah, Catherine, et al. ”The caltech-ucsd birds-200-2011 dataset.” (2011).

[5] Movshovitz-Attias, Yair, et al. ”No fuss distance metric learning using proxies.” Proceedings of the IEEE International Conference on Computer Vision. 2017.

[6] Chang, Dongliang, et al. ”The devil is in the channels: Mutual-channel loss for fine-grained image classification.” IEEE Transactions on Image Processing 29 (2020): 4683-4695.

[7] Ge, Weifeng, Xiangru Lin, and Yizhou Yu. ”Weakly supervised complementary parts models for fine-grained image classification from the bottom up.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[8] Branson, Steve, et al. ”Bird species categorization using pose normalized deep convolutional nets.” arXiv preprint arXiv:1406.2952 (2014).

[9] Zhuang, Peiqin, Yali Wang, and Yu Qiao. ”Learning attentive pairwise interaction for fine-grained classification.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[10] Dosovitskiy, Alexey, et al. ”An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).