F-049

特徴抽出器構造がクラス分離に及ぼす影響の調査

An Investigation into the Effects of Feature Extractor Architectures on Class Separability

髙橋 知里[1)]　神野 健哉[1)]
Chisato Takahashi　Kenya Jin'no

概要

クラス分類を行うモデルは入力データから特徴量を抽出する特徴抽出器と、得られた特徴量を用いてクラス分類を行う分類器から成る。本研究では ReLU 関数を用いた特徴抽出器の構造がクラス分離に与える影響を調査し、学習後のクラス分離度に最適値が存在するかどうかを確認する。具体的には、畳み込み層、Batch Normalization 層、ReLU 関数で構成された特徴抽出器の各パラメータに対し、最適なクラス分離度を持つ構造を探索する。

1　まえがき

クラス分類のモデルは入力から特徴量を抽出する特徴抽出器と得られた特徴量を用いてクラス分類を行う分類器で構成される。抽出された特徴量でクラス分類をする場合、各クラスに対応した特徴ベクトルが直交していることが望ましい。本研究で取り扱う特徴抽出器では ReLU 関数により 0 または正の値のみのベクトルとなる。分類するクラス数と特徴ベクトルの次元数が等しい場合、正規化された直交ベクトルは one-hot ベクトル以外存在しない。そこで本研究では特徴ベクトルを分類クラス数の次元とし、分類器の各クラスに対応した重みを one-hot 表現にした場合を教師信号として特徴抽出器のみを学習させる。この時の特徴抽出器の構造から特徴ベクトルの性質を予測し、分類性能が予測できるかついて検討する。構造を変化させた場合に分類性能の最適値が存在すれば、パラメータから学習後の分類精度が予測できる可能性がある。

2　クラス分離度

特徴抽出器は、畳み込み層、Batch Normalization 層、ReLU 関数からなるブロックを繰り返し用いる構造である。畳み込み層のカーネルサイズは 3x3、ストライドは 1、パディングは 1 とし、最終層には Global Average Pooling を使用する。この特徴抽出器の出力を用いて、分類器がクラス分離を行う。分類器は全結合と Softmax 関数で、特徴ベクトルと全結合層の各クラスに対応した重みとの内積値でクラス判別が行われる。このため各クラスに対応した重みが直交していれば理想的にクラス判別できる。ReLU によって特徴ベクトルの各次元要素は 0 または正の値のみとなる。特徴ベクトルの次元数とクラス数が等しい場合、直交する単位ベクトルは one-hot のみであり、各軸に対応する。すなわち、特徴ベクトル中の最大値に相当する次元が予測クラスとなるため、高次元空間を低次元で評価することができる。そこで本実験では分類器の各クラスに対応した重みを one-hot ベクトルに固定し、特徴抽出器のみを学習させる。

このモデルでは特徴ベクトルの各次元が各クラスに対応している。このため、あるクラスに属するデータの特徴ベクトルと当該次元の値の分布とその他のクラスの特徴ベクトルの当該次元のベクトルの分布がどのようになっているかを考える。すなわち、当該クラスとその他のクラスの特徴ベクトル値の分布が離れていれば精度良くクラス分類を行える可能性がある。これを測るために本稿では「クラス分離度」を以下のように定義する。クラス c_i の入力を $\boldsymbol{o}^{c_i} \in \mathbb{R}^{v\times h}$、その特徴ベクトルを $\boldsymbol{f}^{c_i} \in \{0\} \cup \mathbb{R}^{+10}$ とする。分類器の各クラスに対応した重みを以下のように定義する。

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_{10} \end{bmatrix} \tag{1}$$

特徴ベクトル f^{c_i} が入力のとき、クラス k に対応した重み w_k の Softmax 関数出力は以下のように表せる。

$$y_k = \text{softmax}\,(\boldsymbol{f}^{c_i}\boldsymbol{w}_k) \tag{2}$$

分類器の各重みを以下のように単位ベクトルとする。

$$\boldsymbol{w}_k = \text{one-hot} \quad \text{and} \quad \|\boldsymbol{w}_k\| = 1 \tag{3}$$

このとき、式 (2) の Softmax 関数は次式と等価である。

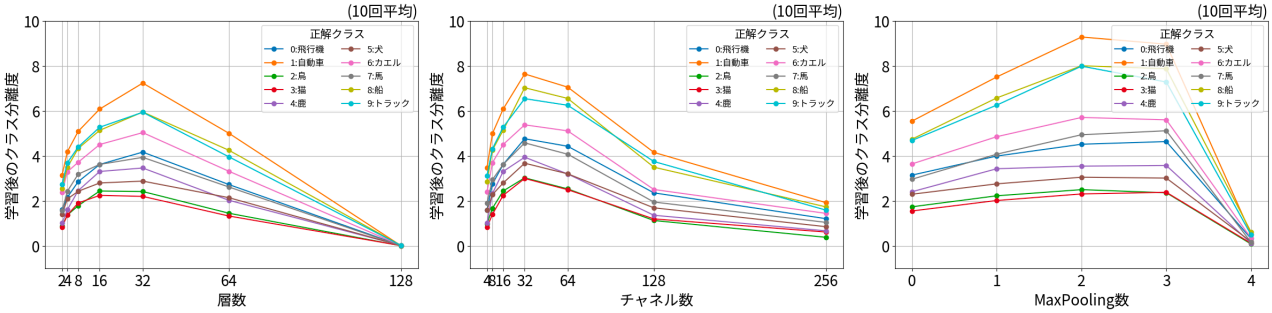
$$i = \arg\max_k\,(\boldsymbol{f}^{c_i}\boldsymbol{w}_k) \tag{4}$$

重みは one-hot であるため、各重みは下記のように直交関係となる。

$$\boldsymbol{w}_k \cdot \boldsymbol{w}_l = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

すなわち、10 次元の各軸は各クラスの特徴量となるため、その分布を考える。うまく分類されるためには各

1)　東京都市大学大学院総合理工学研究科情報専攻
Informatics, Graduate School of Integrative Science and Engineering, Tokyo City University



(a) 層数 (チャネル数 16、MaxPooling 数 0)　(b) チャネル数 (層数 16、MaxPooling 数 0)　(c) MaxPooling 数 (層数 60、チャネル数 16)

図 1: 各パラメータと各正解クラスの学習後分離度平均の関係。

軸上で該当クラスに対応した特徴量はその他のクラスとは重複しない分布になる必要がある。これを測るためフィッシャー判別基準 [1] ををを元に、クラス内平均とクラス内分散から異なるクラス間での特徴ベクトルの分離具合を定量的に示す指標であるクラス分離度を定義する。

$\boldsymbol{f}^{(c_i)}$ の m 次元目の要素を $f_m^{c_i}$、c_i に属するデータ数を N_{c_i} とし、μ_m^i、μ_m^j はそれぞれ各クラス c_i とその他クラス $\bar{c_i}$ のクラス内平均である。

$$\mu_m^i = \frac{1}{N_{c_i}} \sum_{k \in c_i} f_m^k \tag{6}$$

$$\mu_m^j = \frac{1}{N_{\bar{c_i}}} \sum_{l \in \bar{c_i}} f_m^l \tag{7}$$

σ_m^i、σ_m^j はそれぞれ各クラス c_i とその他クラス $\bar{c_i}$ のクラス内分散である。

$$\sigma_m^i = \frac{1}{N_{c_i}} \sum_{k \in c_i} \left(f_m^k - \mu_m^i \right)^2 \tag{8}$$

$$\sigma_m^j = \frac{1}{N_{\bar{c_i}}} \sum_{l \in \bar{c_i}} \left(f_m^l - \mu_m^j \right)^2 \tag{9}$$

これらを使用して、各軸上の特徴量のクラス分離度を式 (10) のように定義する。

$$S_i = \frac{\mu_m^i - \mu_m^j}{(\sigma_m^i)^2 + (\sigma_m^j)^2} \tag{10}$$

クラス分離度 S_i はクラス内平均値が離れていて、クラス内分散が小さければ大きな値となる。すなわちクラス分離度 S_i が大きければ i 番目のクラス分類においては分離しやすくなることを示唆する。

3　各パラメータに対する分離度の変化

各パラメータと学習後クラス分離度との関係を確認する。Floating Point Operations per second(FLOPs) は、1 秒間に実行される浮動小数点演算の数を示す指標であり、ニューラルネットワークの計算量やモデルの複雑さを測るために用いられる。本研究では、異なるモデル間での比較を公平に行うために、学習に必要な FLOPs を統一して実験を行った。本実験で用意したモデルの中で 1epoch の FLOPs が最も大きいのは層数 16、チャネル数 256 のモデルでであったので、この FLOPs を基準として学習回数を調節した。10 クラスの分類問題である CIFAR-10[2] を用い、モデルはスキップ接続などを含まない、層を順次接続したものを対象とする。

パラメータは、層数、チャネル数、MaxPooling 数である。層数をカウントする際は畳み込み、Batch Normalization、ReLU 関数のブロックを 1 層とみなす。同一モデル内のチャネル数は全て統一した。MaxPooling は各ウィンドウ内の最大値を選択することで特徴マップのサイズを縮小する操作であり、層に対し MaxPooling を均等に配置した。各パラメータに対する各正解クラスの学習後分離度平均の推移を図 1 に示す。

図 1 から、どのパラメータも最適値を有する可能性が示唆される。つまり、パラメータから学習後クラス分離度の予測ができる可能性がある。

4　まとめ

FLOPs を統一した状態で学習を行った場合、各パラメータに対して最適なクラス分離度が存在することが示唆された。これは、パラメータ設定によって学習後の分類精度を予測可能であることを示唆している。今後はこの知見をもとに、構造からの学習後クラス分離度を予測することを目指す。

謝辞

本研究の一部は JSPS 科研費 23K11266, 23H03387, 24K15115，東北大学電気通信研究所共同プロジェクト研究，東京都市大学重点推進研究未来知能ユニットの助成によるものです．

参考文献

[1] R.A. Fisher, “The use of multiple measurements in taxonomic problems,” Annals of Human Genetics, vol.7, no.2, pp.179–188, 1936.

[2] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Technical Report 0, University of Toronto, Toronto, Ontario, 2009.