特徴量圧縮モデルへの注意機構導入の検討

A Study on Introducing Attention Mechanism into Feature Compression Model

飯野景[*1]　高橋美穂[*1]　渡辺裕[*1]　江田毅晴[*2]　榎本昇平[*2]　坂本啓[*2]　史旭[*2]　森永一路[*2]
Kei Iino　Miho Takahashi　Hiroshi Watanabe　Takeharu Eda　Shohei Enomoto　Akira Sakamoto　Xu Shi　Ichiro Morinaga

[*1] 早稲田大学大学院 基幹理工学研究科　[*2]NTT ソフトウェアイノベーションセンター
Graduate School of Fundamental Science and Engineering, Waseda University　NTT Software Innovation Center

1. まえがき

協調型知能（Collaborative Intelligence[1]）では，DNN モデルを中間層で分割し，前段をエッジデバイスに後段をクラウドサーバに配置する．この方式では分割した前段の出力である特徴量がクラウド側に伝送されるため，従来の画像圧縮技術と同様に特徴量の圧縮技術が重要となる．特徴量圧縮のためにオートエンコーダ型の圧縮モデルを使用する研究[2,3]が近年発表されている．しかし，これらの研究の多くはモデルの学習法に注目しており，最適なモデルの構造についての研究はなされていない．そこで本研究では，圧縮モデルへの注意機構（Attention）の導入を検討する．

2. 提案手法

本研究では先行研究[2]同様，DNN モデルを中間層で分割し，オートエンコーダ型の圧縮モデルを挿入する．提案手法では先行研究[2]で使用されている圧縮モデルに Attention を取り入れ，圧縮性能の向上を図る．

Attention として Convolutional Block Attention [4]（CBAM）を採用した．図 1 に示すように，CBAM はチャネル方向の Channel Attention，空間方向の Spatial Attention を直列に接続する構成となっている．

提案手法では，エンコーダ内に二種類の方法（Attention1/2）で Attention を導入する．Attention1 ではエンコーダの第二畳み込み層の直後に CBAM を挿入する．Attention2 ではスキップ接続で分岐を行い，CBAM 通過前後の特徴量をチャネル方向で結合し Point-wise の畳み込みを施す．

3. 実験

画像分類のタスクに Resnet50 を用い，conv2_x の直後に圧縮モデルを挿入して Attention 導入の効果を検証する．データセットとして，ILSVRC2012 からランダムに選択した 100 クラスのサブセットを使用する．先行研究[2]にならい圧縮モデルの学習には式(1)の損失関数を用いる．

$$L = \lambda E(y, \hat{y}) + bpp \qquad (1)$$

$y, \hat{y}$はそれぞれ分割前の Resnet50 の出力，圧縮モデル挿入後の Resnet50 の出力である．Eはタスクの損失関数（ここではクロスエントロピー損失），bppはエントロピーモデルで計算されるデータサイズである．

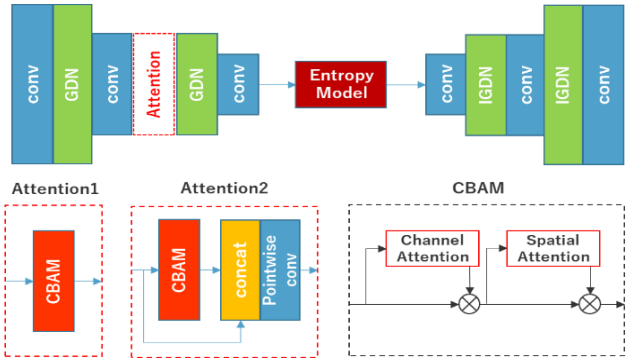

図 1. Attention を導入した圧縮モデル

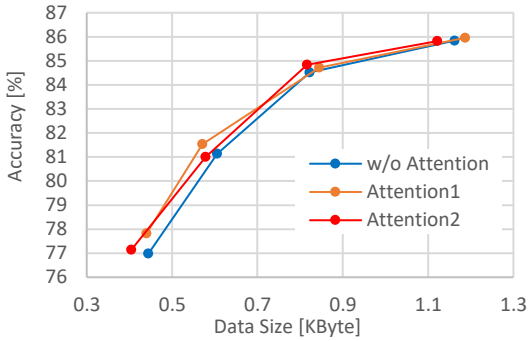

図 2. 各手法の RD 曲線

表 1. 計算量，パラメータサイズの比較

	GFLOPS	Param size [MB]
w/o Attention	3.7138	5.62
Attention1	3.7140	5.64
Attention2	3.7397	5.77

図 2 に Attention なし（w/o Attention），Attention1，Attention2 の圧縮モデルの RD 曲線を示す．図 2 より，Attention を導入することで圧縮性能が向上することがわかる．また表 1 より Attention 導入による追加のオーバーヘッドは非常に小さいことがわかる．

4. むすび

本研究では特徴量圧縮モデルにおける Attention 導入の検討を行い，その有効性を確認した．Attention の種類，位置，個数など，さらなる検討の余地があると考える．

参考文献

[1] Y. Kang et al.: “Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge,” ASPLOS '17, pp. 615- 629, Apr. 2017.

[2] M. Yamazaki et al.: "Deep Feature Compression using Rate-Distortion Optimization Guided Autoencoder," ICIP, pp. 1216-1220, Oct. 2022

[3] P. Datta et al.: "A Low-Complexity Approach to Rate-Distortion Optimized Variable Bit-Rate Compression for Split DNN Computing," ICPR, pp. 182-188, Aug. 2022

[4] S. Woo et al: “CBAM: Convolutional Block Attention Module,” ECCV, pp. 3-19, Sep. 2018