画像の圧縮・復元モデルと認識モデルの End-to-End 学習

○柴田蓮 †， 山内悠嗣 †

○ Ren SHIBATA † and Yuji YAMAUCHI †

†：中部大学， er20043-3876@sti.chubu.ac.jp, yuu@isc.chubu.ac.jp

＜要約＞エッジコンピューティングを活用した画像認識システムは，通信量が増加するとネットワークの負担が大きくなるため，通信速度が低下する等の問題を抱えている．通信量を削減するために圧縮・復元した画像を認識するアプローチが提案されているが，未圧縮の画像を用いて認識する場合と比べて認識精度が低下することが報告されている．そこで，本稿では認識精度の低下を抑制するために画像を圧縮・復元するモデルと認識するモデルを End-to-End で最適化するフレームワークを提案する．提案手法は，異なる 2 つのモデルを End-to-End で最適化するため，画像を圧縮・復元した際に認識に有効な情報の欠落を防ぐことが可能となり，高精度な画像認識の実現が期待できる．

＜キーワード＞画像の圧縮・復元，画像認識，エッジコンピューティング，End-to-End 学習

図 1　未圧縮画像と各手法における圧縮・復元画像の例．(a) は未圧縮の画像，(b) は (c) と同程度のデータ容量となるように圧縮・復元した JPEG[1]，(c) は深層学習により圧縮・復元した画像 [2]，(d) は提案手法により圧縮・復元した画像．

1　はじめに

Internet of Things(IoT) デバイスやモバイル端末の普及に伴い，エッジコンピューティングを活用したサービスが拡大している．エッジコンピューティングはクラウドコンピューティングを発展させたネットワーク技術であり，IoT デバイスの近くに配置されたエッジサーバでデータを処理することにより高速化と低遅延化を実現している．この技術を活用した画像認識システムは，端末の性能に関わらず高度な画像認識が利用可能であり，様々なサービスが実用化されている．しかし，大容量のデータを通信する場合にはネットワークに大きな負荷がかかるため，通信速度が低下する等の問題が発生する．

この問題を解決する方法として，JPEG[1] 等により圧縮した画像を通信する方法が採用されている．また，近年ではより効率良く圧縮するために深層学習を用いた手法 [3] も提案されている．他にも，画像から抽出した特徴量の圧縮 [4, 5, 6] や量子化 [7] などの手法が提案されている．これらの方法は，圧縮したデータをエッジサーバに送信し，エッジサーバ側で圧縮したデータに基づいて画像認識を行うが，未圧縮の画像を認識に用いた場合と比較して認識精度が低下することが報告されている [8]．図 1 に未圧縮画像と JPEG，深層学習により圧縮・復元した画像 [2]，提案手法により圧縮・復元した画像の例を示す．なお，図 1(b) と (c)，(d) の

データ容量は同程度となるように画像を圧縮・復元している．JPEG は低周波成分のテクスチャ情報が失われ，ブロックノイズが発生している．図 1(c) は，JPEG よりは高い品質で画像を圧縮・復元できているが，高周波成分のノイズが散見される．このようなノイズ成分は画像認識における性能の低下を引き起こす．また，圧縮・復元する際には認識することを考慮していないため，認識に寄与する重要な情報が欠落する恐れがある．

そこで本稿では，画像の圧縮・復元モデルと認識モデルを End-to-End で最適化するアプローチを提案する．2 つのモデルを End-to-End で最適化することで認識を考慮した画像の圧縮・復元が可能となる．図 1(d) に示すように画像を圧縮・復元する際，認識に有効な情報の欠落やノイズの発生を抑制することができるため，圧縮・復元した画像を用いた認識精度の低下の抑制が期待できる．

2 関連研究

2.1 画像の圧縮・復元

JPEG[1] をはじめとする非可逆圧縮手法は，画像データを効率的に圧縮できるため，画像ファイル形式のデファクトスタンダードとして普及している．JEPG は画像を効率良く圧縮できる一方で，ノイズの発生や元の画像に復元できない問題があるため，可逆圧縮手法である PNG[9] も広く普及している．

近年では，より効率的な圧縮を実現するために深層学習を用いた画像圧縮手法が提案されている．Toderici *et al.* は，Long Short Term Memory(LSTM)[10] に基づいた画像圧縮手法 [11] を提案した．この手法では，入力画像と圧縮・復元処理により得られた復元画像の差分を再びモデルに入力する．この処理を再帰的に繰り返し，復元画像を加算したものを最終的な出力画像とする．繰り返し回数を設定することで圧縮率と画像品質を制御することができる一方で，長期的なデータの保持はメモリ消費が大きいため，勾配消失や扱えるデータ容量が限られてしまう問題を抱えていた．そこで，LSTM の代わりに Gated Recurrent Unit(GRU)[12] を採用することで，この問題を回避する手法が提案された [2]．また，画像を復元する前の処理を繰り返すことで中間層の表現能力を上げるプライミングと復元画像の画像品質を目標値まで向上させるためビットレートを動的に調整する適応型ビットレートを導入することで圧縮率と画像品質を向上させる手法 [13] や，RNN の再帰的処理が増えることで発生する勾配消失問題を解決するための工夫として入力画像と復元画像の差分から入力画像を予測する Residual-to-Image と圧縮する前の画像から隣接するパッチ間の空間的な一貫性を利用する手法 [14] も提案されている．

2.2 画像圧縮技術を用いた画像認識

画像圧縮技術に関する研究が取り組まれる一方で，圧縮・復元した画像の認識に関する研究が進められている．Endo *et al.* は，深層学習により推定した JPEG 画像の品質係数を認識モデルに入力することで JPEG 画像から高精度に画像分類を行う手法 [15] を提案した．Park *et al.* は，符号化された JPEG から Vision Transformers(ViT)[16] に基づく認識モデルを最適化する手法 [17] を提案した．符号化された JPEG から画像にデコードする処理を省略することで高速化を実現している．

高瀬等は，画像の圧縮・復元モデルと認識モデルを結合したフレームワーク [18] を提案した．このフレームワークは，深層学習により画像を圧縮することでデータ通信量を削減し，復元した画像から画像分類や物体検出を行う．深層学習により画像を圧縮することでデータ通信量を大幅に削減できる一方で，未圧縮画像を用いた認識と比較して精度が低下する問題を抱えている．Janeiro *et al.* は，この問題を解決するために圧縮・復元した画像で認識モデルをファインチューニングすることにより，性能の低下を抑制した [8]．

2.3 提案手法の概要

本研究では，画像の圧縮・復元モデルと認識モデルを結合したフレームワーク [18] における認識性能の低下を抑制することを目的とする．これを実現するために，本研究では圧縮・復元モデルと認識モデルを End-to-End で最適化するフレームワークを提案する．本研究の貢献点は以下の通りである．

- 画像認識に適した画像の復元
  既存の手法は画像を圧縮・復元するモデルと認識するモデルを個々に最適化していた．そのため，認識に不要なノイズが画像に混入することや，認識に寄与する情報が欠落する恐れがあった．この問題

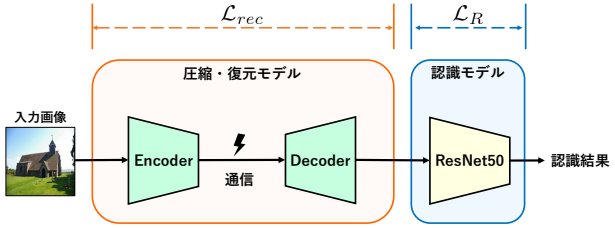

図 2　提案手法の概要．

を解決するために，本研究では 2 つのモデルを同時に最適化することで認識性能の低下を抑制する．

- 2 つのモデルの最適化
  提案手法は，画像を圧縮・復元するモデルと認識するモデルの損失を最小化するように End-to-End で最適化する．ただし，異なるタスクを解く 2 つのモデルを同時に最適化することは，必ずしも最適な解が得られるとは限らない [19]．そこで，本研究では 2 つのモデルを同時に最適化するアプローチと交互に最適化するアプローチの 2 つを提案する．

3　提案手法

提案手法の概要を図 2 に示す．提案手法は，画像を圧縮・復元するモデルと認識するモデルを直列に結合した構成となっている．エンコーダにより画像を圧縮したデータを送信することで通信量の削減が可能となる．また，圧縮・復元モデルと認識モデルを End-to-End に最適化することで認識に有効な情報の欠落とノイズを抑制が期待できる．

3.1　画像の圧縮・復元モデル

画像の圧縮・復元モデルには， Recurrent Neural Network(RNN)[20] を用いた再帰型オートエンコーダ [2] を採用する．図 3 に圧縮・復元モデルの流れを示す．本手法は，エンコーダ E，デコーダ D，バイナライザ B によって構成される．

次式に示すように入力画像 x はエンコーダ E_t によって圧縮後，バイナライザ B によってバイナリコード b_t に変換される．

$$b_t = B\left(E_t\left(r_{t-1}\right)\right) \tag{1}$$

ここで t は RNN の繰り返し回数，r_{t-1} は入力画像 x と復元した画像 $\hat{x}_t$ の差分画像を示す．なお，初回は $\hat{x}_0 = 0$，$r_0 = 0$ となる．圧縮されたバイナリコード b_t は，デコーダ D_t を用いて式 (2) により画像 $\hat{x}_t$ として復元される．

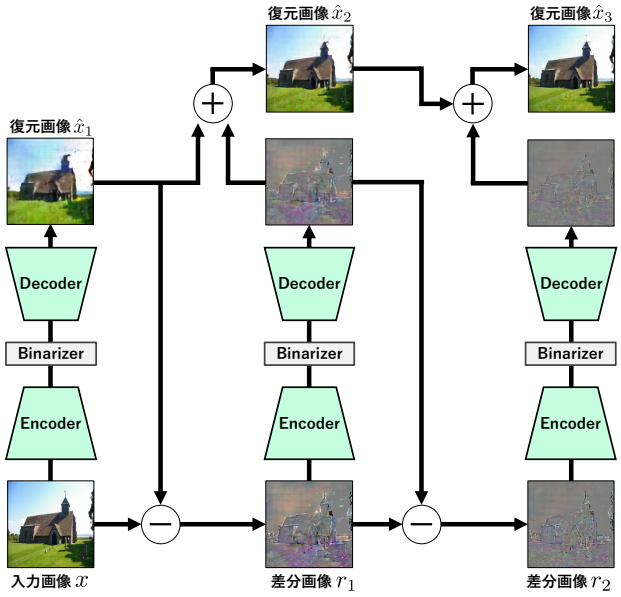

図 3　圧縮・復元モデルの流れ．

$$\hat{x}_t = D_t\left(b_t\right) + \hat{x}_{t-1} \tag{2}$$

上記の圧縮・復元処理を T 回繰り返し，処理ごとに得られる復元画像を加算した $\hat{x}_T$ を最終的な出力とする．

圧縮・復元モデルは，入力画像と出力画像の平均絶対誤差 $\mathcal{L}_{rec}$ を最小化するように最適化される．圧縮・復元モデルの損失関数を式 (3) に示す．

$$\mathcal{L}_{rec} = \sum_{t=1}^{T} |r_t| \tag{3}$$

RNN の繰り返し回数である T を大きくするほど元の画像に近しい画像を復元することができるが，計算量とデータ通信量が増加する．

3.2　画像の認識モデル

画像の認識モデルには，50 層からなる Residual Neural Network(ResNet)50[21] を採用する．ResNet50 は，畳み込み層の出力と入力を加算する残差ブロックを導入することで，ネットワークの深層化により発生する勾配消失問題を解決した畳み込みニューラルネットワークである．本研究では，圧縮・復元モデルにより復元した画像 $\hat{x}_t$ を認識モデル R に入力する．認識モデルの損失関数を式 (4) に示す．

$$\mathcal{L}_R = -p(R(\hat{x}_t)) \log q(R(\hat{x}_t)) \tag{4}$$

p は正解ラベルの確率分布，q は予測の確率分布を表し，交差エントロピー誤差を最小化するように最適化される．

3.3 提案手法による学習

圧縮・復元モデルと認識モデルを個々に最適化すると認識精度が低下する [18]．これは，画像の圧縮・復元モデルを最適化する際に認識モデルについて考慮していないため，画像を圧縮・復元する際にノイズの混入や認識に寄与する情報が欠落するためである．そこで提案手法は，直列に結合した圧縮・復元モデルと認識モデルを End-to-End に最適化する．提案手法における損失関数を式 (5) に示す．

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_R \tag{5}$$

λ は，2 つのモデルの重みを調整するハイパーパラメータであり，本研究では 0.6 に設定した．提案手法は，圧縮・復元モデルと認識モデルの損失の和を最小化するように最適化する．本研究では，式 (5) で示す損失関数を用いて 2 つの方法で最適化する．

- **2 つのモデルの同時最適化**

  圧縮・復元モデルと認識モデルを個々に事前学習し，式 (5) に示す損失関数により 2 つのモデルを同時にファインチューニングする．2 つのモデルを同時に最適化することで，認識モデルを考慮した画像の圧縮復元が可能となる．

- **2 つのモデルの交互最適化**

  先に述べた 2 つのモデルを同時に最適化するアプローチは，文献 [19] でも言及されているように最適化するパラメータが多いため，勾配の消失や局所解への収束等の問題が発生することが考えられる．そこで，本研究では，Generative Adversarial Network[22] のように 2 つのモデルを交互に最適化する．圧縮・復元モデルを最適化するために式 (5) の第 2 項を固定し，第 1 項のみを最適化する．次に，認識モデルを最適化するために式 (5) の第 1 項を固定し，第 2 項のみを最適化する．交互に最適化する処理を繰り返すことで 2 つのモデルを最適化する．

図 4 に繰り返し回数ごとの圧縮・復元画像を示す．どの学習方法においてもデータ通信量は同程度であるが，圧縮・復元される画像の見えに差が生じる．圧縮・復元モデルと認識モデルを個々に最適化した場合，繰り返し回数が 1 回の画像においてノイズが散見され，元の画像を正確に復元することができていないことがわかる．一方，提案手法では個別に最適化する方法よりは正確に復元できていることが確認できる．

4 評価実験

提案手法の有効性を確認するための評価実験を行う．本実験は，10 クラスの画像分類問題により評価する．下記の 4 つの手法における画像の分類精度とデータ通信量，未圧縮画像と圧縮・復元モデルの出力画像の PSNR(Peak Signal-to-Noise Ratio) を比較する．

- 未圧縮モデル
  未圧縮画像を用いて認識モデルを学習
- 個別最適化モデル
  圧縮・復元モデルを学習後，圧縮・復元モデルの出力画像を用いて認識モデルを学習
- 交互最適化モデル
  圧縮・復元モデルと認識モデルを交互に学習
- 同時最適化モデル
  圧縮・復元モデルと認識モデルを同時に学習

個別最適化モデルと提案手法は圧縮・復元モデルの繰り返し回数を制御することで，認識に用いる画像の画像品質を変更できる．本実験では，繰り返し回数を 1 から 10 回まで試行する．

4.1 データセット

本実験で使用する圧縮・復元モデルは CIFAR-10，認識モデルは ImageNet[23] を用いて事前学習を行う．圧縮・復元モデルと認識モデルのファインチューニング及び評価には，ImageNet から 10 クラスを抽出した ImageNette[24] を使用する．ImageNette は，学習用データとして 9,469 枚，評価用データとして 3,925 枚用意されている．

4.2 実験結果

4.2.1 認識精度とデータ通信量の評価

図 5 に繰り返し回数ごとの各モデルの平均認識精度と平均データ通信量を示す．個別最適化モデルと提案手法の平均認識精度を比較すると平均データ通信量が

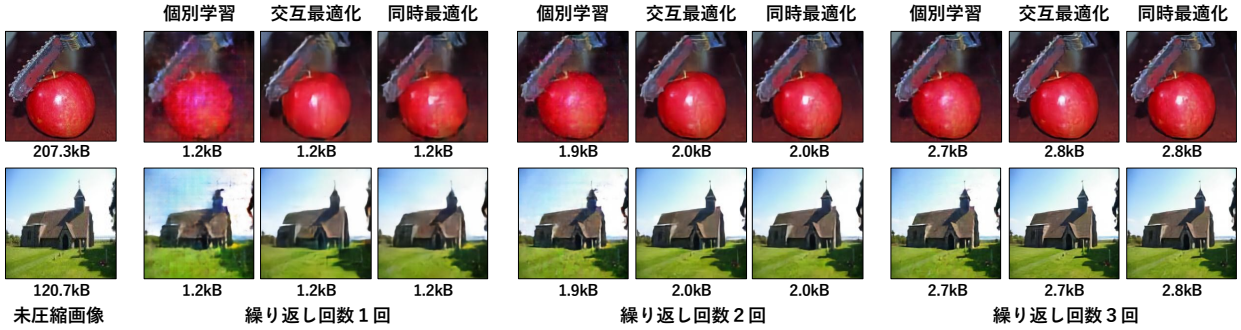

図 4　繰り返し回数ごとの個別最適化と同時最適化および交互最適化の圧縮・復元画像の例．

小さいほど，提案手法の方が認識精度が高いことがわかる．平均データ通信量が 1.2kB の画像を用いた平均認識精度は，個別最適化モデルは 91.6 %，同時最適化手法は 94.3 %，交互最適化手法は 94.9 %となった．2 つの提案手法は，どちらも個別最適化モデルに対して約 3 %以上高い認識精度を示した．これは，画像を圧縮・復元する際に認識に有効な情報の欠落を抑制することや，ノイズの発生を防ぐことができたためだと考えられる．また，繰り返し回数が多いほど認識精度が向上し，未圧縮モデルに近しい認識精度を得ることができた．また，平均データ通信量が小さい時には交互最適化の方が僅かに良い認識精度が得られた．個別最適化モデルと提案手法のデータ通信量を比較する．未圧縮モデルは，平均データ通信量が 118.2kB であったのに対し，データを圧縮する 3 つの手法は最大で約 99 %のデータ通信量を削減できた．繰り返し回数が多いほどデータ通信量が増加するが，繰り返し回数 10 回の場合においても，未圧縮モデルと比較して約 93 %のデータ通信量を削減することが可能であることがわかる．

表 1 に繰り返し回数ごとの各手法における圧縮・復元モデルの出力画像と未圧縮画像の PSNR の平均値を示す．PSNR は画像の品質を表し，PSNR が大きいほど画像の劣化が小さいことを表す．全ての繰り返し回数において同時最適化モデルと交互最適化モデルは，個別最適化モデルより高い PSNR が得られた．このことから，画像の圧縮・復元によるノイズの発生を抑制し，元の画像に近い画像を出力できたといえる．

4.2.2　復元画像と認識モデルの判断根拠の比較

各手法における認識時の判断根拠の変化を確認する．なお，判断根拠の可視化には，Grad-CAM [25] を使用する．図 6 に繰り返し回数 1 回の圧縮・復元モデルの出

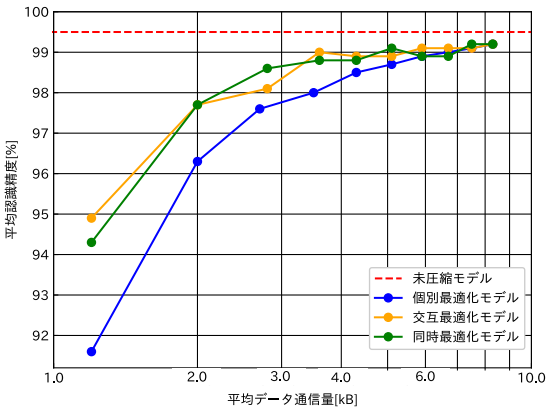

図 5　繰り返し回数ごとの各手法の平均認識精度 [%] と平均データ通信量 [kB]．

力画像と認識の判断根拠を可視化した画像を示す．個別最適化モデルは，多くの画像でノイズが散見していることがわかる．交互最適化モデル及び同時最適化モデルでは，ノイズの発生を抑制できていることがわかる．また，認識の判断根拠を可視化した画像を比較すると，個別最適化モデルよりも提案手法の方が，未圧縮モデルと近い領域を注視していることがわかる．そのため，認識に有効な情報の欠落を抑制できた可能性が高いといえる．

5　おわりに

本稿では，圧縮・復元モデルと認識モデルを End-to-End で最適化するフレームワークを提案した．提案手法により，ノイズの発生と認識に有効な情報の欠落を抑制した画像の圧縮・復元が可能となり，認識精度が向上することを確認した．今後は，本アプローチに適した画像の圧縮・復元モデルを提案する予定である．

表 1　繰り返し回数ごとの各手法における平均 PSNR[dB]

繰り返し回数	1	2	3	4	5	6	7	8	9	10
個別最適化モデル	20.2	23.1	24.5	25.5	26.1	26.7	27.2	27.5	27.8	28.1
交互最適化モデル	21.9	24.1	25.3	26.2	26.9	27.4	27.8	27.5	28.5	28.8
同時最適化モデル	21.6	23.9	25.1	26.0	26.7	27.2	27.7	28.0	28.3	28.6

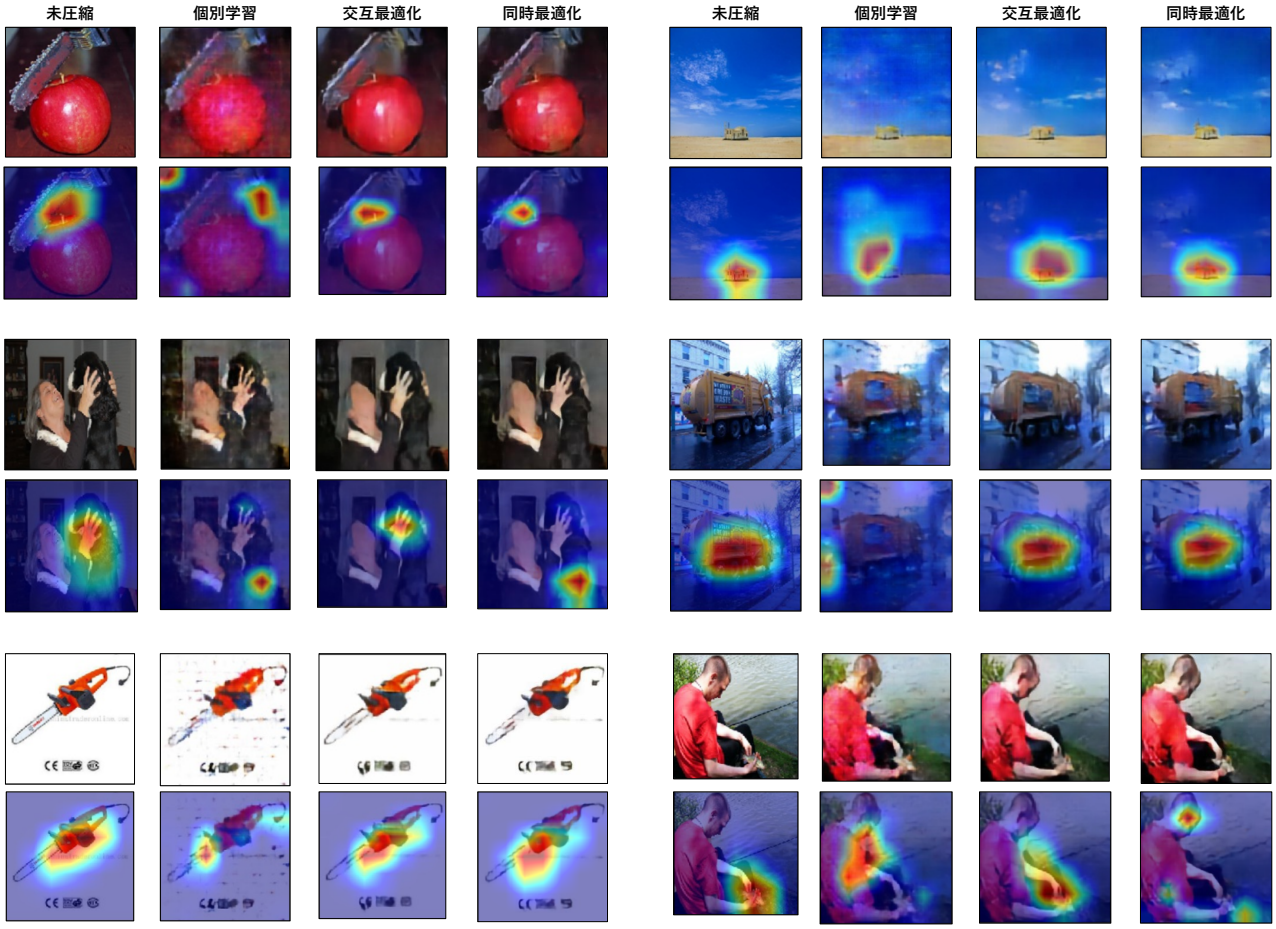

図 6　Grad-CAM による判断根拠の可視化結果の例．上段は入力画像，下段は Grad-CAM により判断根拠をヒートマップとして可視化した画像．

参考文献

[1] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, “An overview of jpeg-2000”, *Data Compression Conference*, pp. 523–541, 2000.

[2] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell, “Full resolution image compression with recurrent neural networks”, *Conference on Computer Vision and Pattern Recognition*, pp. 5306–5314, 2017.

[3] M. Nakahara, M. Nishimura, Y. Ushiku, T. Nishio, K. Maruta, Y. Nakayama, and D. Hisano, “Edge computing-assisted dnn image recognition system with progressive image retransmission”, *IEEE Access*, 2022.

[4] O. Wiles, J. Carreira, I. Barr, A. Zisserman, and M. Malinowski, “Compressed vision for efficient video understanding”, *Conference on Computer Vision and Pattern Recognition*, 2022.

[5] Y. Kang, J. Hauswald, C. Gao, A. Rovinski, T. Mudge, J. Mars, and L. Tang, “Neurosurgeon: Collaborative intelligence between the cloud and mobile edge”, *ACM SIGARCH Computer Architecture News*, 2017.

[6] H. Choi, and I. V. Baji, “Deep feature compression for collaborative object detection”, *International Conference on Image Processing*, 2018.

[7] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning”, *Advances in Neural Information Processing Systems*, 2018.

[8] J. M. Janeiro, S. Frolov, A. El-Nouby, and J. Verbeek, “Are visual recognition models robust to image

compression?" *Conference on Computer Vision and Pattern Recognition*, 2023.

[9] G. Roelofs, "PNG: the definitive guide", *Linux Journal*, 1999.

[10] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[11] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks", *Conference on Computer Vision and Pattern Recognition*, 2016.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation", *Conference on Empirical Methods in Natural Language Processing*, 2014.

[13] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor, and G. Toderici, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks", *Conference on Computer Vision and Pattern Recognition*, pp. 4385–4393, 2018.

[14] M. H. Baig, V. Koltun, and L. Torresani, "Learning to inpaint for image compression", *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] K. Endo, M. Tanaka, and M. Okutomi, "Classifying degraded images over various levels of degradation", *International Conference on Image Processing*, 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", *International Conference on Learning Representations*, 2021.

[17] J. Park, and J. Johnson, "RGB no more: Minimally-decoded jpeg vision transformers", *Conference on Computer Vision and Pattern Recognition*, pp. 22 334–22 346, 2023.

[18] 高瀬 俊希, 戸谷 響, 西片智広, 山内悠嗣, "エッジコンピューティングのための圧縮画像認識", ビジョン技術の実利用ワークショップ, 2022.

[19] T. Glasmachers, "Limits of end-to-end learning", *Proceedings of Machine Learning Researchs*, pp. 17–32, 2017.

[20] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 5, no. 2, pp. 157–166, 1994.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", *Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[24] J. Howard, and S. Gugger, "Fastai: A layered api for deep learning", *Information*, vol. 11, no. 2, p. 108, 2020.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", *International Conference on Computer Vision*, pp. 618–626, 2017.