様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ６ 日現在

機関番号：１１３０１
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７７５６

研究課題名（和文）ドメインシフトを解決する深層ネットワークアーキテクチャの設計

研究課題名（英文）Designing deep network architectures to solve domain shift

研究代表者
菅沼　雅徳（Suganuma, Masanori）

東北大学・情報科学研究科・助教

研究者番号：００８１５８１３

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：深層ニューラルネットワークは，様々な問題を高精度に解ける一方で，学習データとわずかであれ異なる性質を持つデータに対し，期待するような性能を示せないという課題がある（ドメインシフト問題）．この課題に対し，様々な深層ニューラルネットワークを評価し解析することで，ドメインシフトに対してより頑健性を示すネットワーク構造の発見を目指した．ドメインシフトの例として，劣化画像に対する画像分類問題に取り組み，ドメインシフトに頑健性を示すネットワーク構造の知見を得ることができた．

研究成果の学術的意義や社会的意義
深層ニューラルネットワークは，近年発展の著しい人工知能（AI）の中枢技術であり，実応用も数多く行われている．しかしながら，実応用の上で頻出するドメインシフト問題に対して決定打となる方法がないという課題があった．本研究は，ネットワーク構造の観点からこの問題に取り組み，ネットワーク構造のドメインシフトに与える影響について，多くの知見を得た．これらの成果は，今後の研究や実応用において参考になりうる．

研究成果の概要（英文）：Deep neural networks can solve various problems with high accuracy, but they face the challenge of not performing as expected when dealing with data that slightly differs from the training data (the domain shift problem). To address this issue, various deep neural networks were evaluated and analyzed with the goal of discovering network structures that are robust to domain shifts. As an example of domain shift, this study tackled the problem of image classification for degraded images and was able to gain insights into network structures that demonstrate robustness to domain shifts.

研究分野：コンピュータビジョン

キーワード：深層学習　画像認識

１．研究開始当初の背景

研究構想時，深層学習はコンピュータビジョンのほとんどのタスクにおいて従来の機械学習を含む問題解決の方法論を圧倒し，成功を収めていた．深層学習の成功が，大量のデータを用いた学習によってもたらされていることに疑いはないが，逆に，学習データへの依存性が「高すぎる」と言える部分があり，それによって新たな課題が生み出されていた．中でも，学習に用いたデータとわずかであれ異なる性質を持つデータに対し，期待するような性能を挙げないこと（ドメインシフトと称される）は，実用上，大きな問題となっていた．

一例として図１に，畳み込みニューラルネット（以下，CNN）を用いた物体検出手法の，入力画像の画質変動に対する振る舞いを示す．この CNN は，比較的良好な画質の画像を用いて学習を行ったものであり，それらと同様の画質の入力に対しては，図の左列のように十分以上の性能を見せる．しかしながら，図 1 右列のように，ノイズやモーションブラーなどの外乱を含む画像が入力されると，正しくない結果が出力されることがよく知られていた．

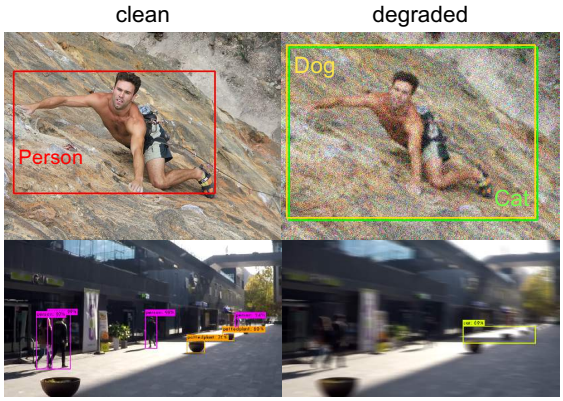

図 1．画像劣化（ドメインシフト）による物体検出器の出力変化の例．左列：良好な画質の画像に対する検出結果．右列：外乱を含む画像に対する検出結果．誤検出をしていることがわかる．

このようなドメインシフト（＝学習データ分布外のデータに対する性能低下）を解決すべく行われている，関連分野での取組は２つに大別されていた．一つは学習データの規模を拡大することであり，上の例で言えば劣化画像をも一緒に学習することに当たる．これは，実行できれば大抵の場合にうまくいく一方で，コストがかかりすぎる欠点があった．例えば，ドメインシフトを誘発する要因が複数ある場合（例えば，ノイズ，モーションブラー，降雨，霧霞などの要因），その組み合わせの数が爆発し，集めるべきデータが膨大になり，現実的には実行できなくなる場合も多い．もう一つのアプローチが，ドメイン適合と呼ばれる方法である．これは，モデルを特殊な方法で学習することで，想定できる範囲内のドメインの違いに寛容な特徴を使って，推論を行えるようにする．しかし，膨大な量の研究が行われているものの，実用的な性能は低く，決定打にはなっていなかった．

以上を踏まえ，本計画では「どのような構造をもつネットワークが，ドメインシフトに頑健なのか？」といった問いに答えることを目的とした．現在までの深層学習の発展は，良いネットワーク構造の発見が牽引してきたと言っても良く，それほどネットワーク構造が性能に与える影響は大きく，タスクやデータごとに最適な構造は変化する．そのため，様々な試行錯誤での構造の探求や，構造の自動設計（NAS）の研究が盛んに行われてきた．しかしながら，ドメインシフトに強いネットワークのデザインを考えることは，これまでほとんど行われてこなかった．

２．研究の目的

本研究では，ドメインシフトの典型例として，図 1 に例示した画像劣化への対処を主なターゲットに選び，画像劣化に頑健なネットワークの構造の探求および理解を目指した．画像の劣化は実環境において頻出することから，解決できればその意義は大きいことが採用理由であった．本研究では，次の二つの観点から上述の目標にアプローチした．

(1) 画像劣化に強いネットワーク構造を発見的に見出す
(2) 様々なネットワークの自動設計技術を応用し，画像劣化に強いネットワーク構造を設計する

３．研究の方法

本研究では，ImageNet-C[1]と呼ばれる人工的に生成した劣化画像を用いて，様々なネットワーク構造のドメインシフト（画像劣化）に対する頑健性評価を行った．ImageNet-C に含まれる劣化画像は，ガウシアンノイズ，JPEG 圧縮ノイズ，モーションブラーなど 19 種類の外乱が付与された画像である．そこから 11 種類を選択し評価に用いた．ImageNet-C は 1000 クラスの画像分類タスクである．

まずは，幅広く使用されている既存モデルを ImageNet-C 上で評価することで，どのようなネットワーク構造が画像劣化（ドメインシフト）に対して頑健性を示すのかを理解することを目指した．そして，得られた知見を軸に，ネットワーク構造の自動設計手法（NAS）を用いて，より網羅的に，頑健なネットワーク構造を探索する．

[1] D. Hendrycs+, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, ICLR, 2019

Model	RepVGG	EfficientNet-v2	ConvNeXt-v2	PVT-v2	EfficientViT
Top1-Acc.	59.71	72.24	72.50	72.98	73.20
Params (M)	57.4	53.2	50.2	45.2	52.7

表 1. 主要な CNN および ViT の ImageNet-C 上での性能比較結果.

Model	ViT-small	ViT-small
Top1-Acc.	58.04	47.48
patch size	16	32

(a)

Model	ViT-small	ViT-base	ViT-large
Top1-Acc.	58.04	68.68	76.85
Params (M)	22.1	86.6	304.3

(b)

表 2. ViT におけるパッチサイズおよびモデル規模が与える影響.

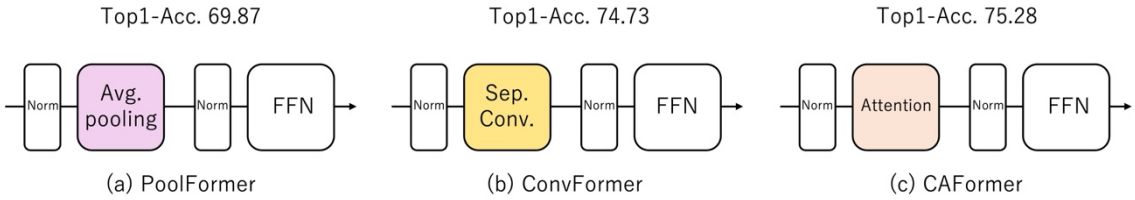

図 2. ViT 内の演算ブロックが頑健性に与える影響.

４．研究成果

(1) ネットワーク構造の分析および理解

畳み込みニューラルネットワーク（CNN），Vision Transformer（ViT）を含む，様々な深層ニューラルネットワークを ImageNet-C 上で評価することで，頑健性の優れたネットワーク構造を見出すことを試みた．表 1 に主要な CNN と ViT による比較結果を載せる．表内の各数値は 11 種類の劣化画像に対する平均画像分類精度を示している．

表 1 より，同程度のモデル規模をもつ CNN と ViT では，CNN（RepVGG，EfficientNet-v2，ConvNeXt-v2）よりも ViT（PVT-v2，EfficientViT）の方が基本的に優れた頑健性を示すことがわかった．CNN のモデル間で比較すると，ConvNeXt-v2，EfficientNet-v2，RepVGG の順で性能が優れており，これは同時に ViT 構造との類似度合いにも一致する（ConvNeXt-v2 が最も ViT の構造に近い）．以上の実験結果から，ViT のモデル構造の方が画像劣化に頑健性を示す傾向があることがわかる．

続いて，ViT におけるどのような構成要素が頑健性に影響しているのかを検証した．まずは，ViT に入力するパッチサイズの影響を同じく ImageNet-C 上で調査した．表 2(a)からわかるように，パッチサイズが小さい方が優れた性能を示すことがわかる．また，ViT のモデルサイズが大きくなるほど，より優れた頑健性を示すことが判明した（表 2(b)）．

さらに，ViT の主要演算である自己注意機構が画像劣化への頑健性に与える影響を調査した．具体的には，図 2 に示すような 3 種類のモデルを用意し，性能比較を行った．いずれのモデルも演算群であるブロックを積み重ねる点は同じだが，ブロックの構成要素が異なる．図 2(a)のモデルは通常の ViT における自己注意機構を平均プーリングに，図 2(b)では可分畳み込みに置換したモデルである．図 2(c)のモデルでは，モデル前半は図 2(b)の構造を採用し，後半は自己注意機構を採用する．これらの比較によって，自己注意機構の与える影響が明らかになる．図 2 上部に示しているように，自己注意機構を有するモデルが最も優れた頑健性を示すことが明らかになった．

最後に，ネットワーク構造以外の要素として，学習方法が画像劣化に対する頑健性への影響を調査した．具体的には，より洗練化された学習方法[2]，ViT を自己教師あり学習の一種である Masked autoencoder（MAE），画像・テキストペアデータによる対照学習（Contrastive Language-Image Pre-training; CLIP）のそれぞれの効果について検証した．表 3 からわかるように，いずれの学習方法を用いても，ベースとなる ViT よりも優れた頑健性を示すことがわかる．その中でも，MAE による事前学習が特に効果を発揮することが判明した．

以上の研究を通じて，画像劣化（ドメインシフト）に対して頑健性を示すネットワーク構造に関する多くの知見を得た．

Model	ViT	ViT-DeiT-III	ViT-MAE	ViT-CLIP
Top1-Acc.	62.16	75.88	76.58	75.57
Params (M)	86.6	86.6	69.9	86.9

表 3. 学習方法が与える頑健性への影響.

[2] H. Touvron+, DeiT III: Revenge of the ViT, arXiv:2204.07118, 2022.

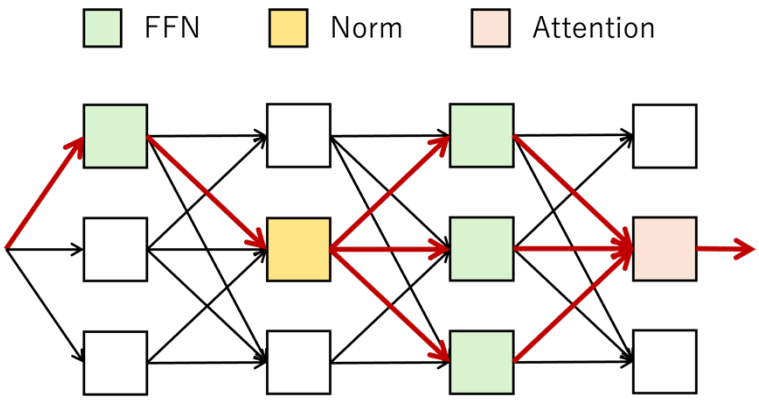

図 3. 提案した ViT の探索空間.

(2) ネットワーク構造の自動設計

より網羅的かつ多種多様な ViT モデルを評価し，ネットワーク構造がドメインシフトに与える影響の理解を深めるために，NAS 手法の適用を検討した．NAS を用いることで，人手では思いもつかないようなネットワーク構造を設計可能であり，さらに人手で設計したものよりも高性能なモデルを獲得できる可能性がある．

そこで，ViT のネットワーク構造を NAS で自動設計可能なように，図 3 に示すような探索空間を開発した．本手法では，ViT を構成する各ブロックは 2 次元の非巡回有向グラフ（DAG）で表現され，各ノードが自己注意機構などの演算子，各エッジがデータの流れを表す．対象とするタスク上での性能を最大化するように，各ノードの種類ならびにノード間の接続関係を最適化することで，所望のモデルを獲得する．しかしながら，NAS には膨大な計算コストがかかるという問題がある．具体的には，NAS 手法では何かしらの方法でネットワーク構造を探索空間からサンプリングし，サンプリングされたモデルを実際に対象データを用いて学習し，そのネットワーク構造の良し悪しを評価，といった処理を繰り返す．この処理のうち，サンプリングされたモデルの学習部分に膨大な計算時間がかかる．

これに対し，本研究ではサンプリングされたモデルの学習なしに，そのモデルの良し悪しを定量化可能な方法を研究した．具体的には，サンプリングされたモデルの重みの初期値分布や出力される特徴ベクトル分布をもとに，モデルの良し悪しを判定する方法を提案した．ImageNet-C 上での評価を行う前に，まずは通常の ImageNet 上で既存モデルと同程度以上の性能を示すモデルを自動設計可能かどうか検証を行った．検証の結果，既存モデルと同程度の性能を示す ViT モデルを設計可能であることを確認したが，設計されたモデルは標準的な ViT に近い構造をもつものが多く，NAS を利用する利点が見出せなかった．本研究期間中に，効率的かつ多種多様な ViT モデルを設計可能な NAS 手法の開発を試みたが，決定打となる方法を見出せなかった．そのため，NAS による ViT モデルの解析は今後の研究課題である．

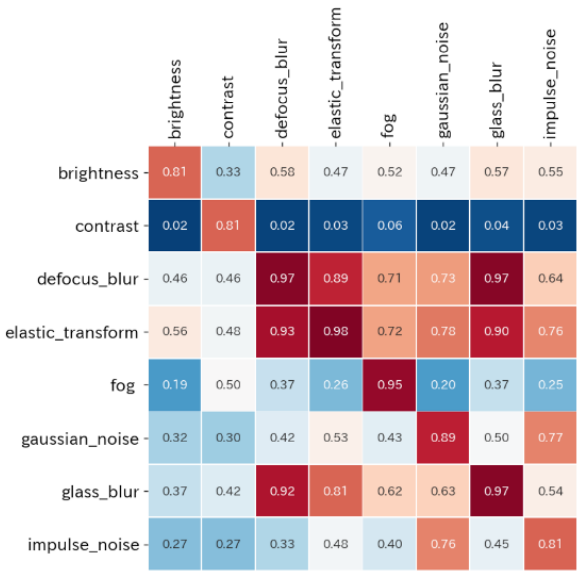

図 4. 学習時の外乱が与える影響.

(3) 特定の外乱が与える影響の解析

上述の研究を遂行中に，特定の外乱を含む画像を用いてモデルを学習すると，学習に用いなかった外乱に対しても頑健性が向上する事象を偶然発見した．学習に用いた外乱に対して頑健性を示すのは当然であるが，それ以外の外乱に対しても頑健性を示す事実はこれまで知られていなかった．

図 4 に学習に用いた外乱と推論時の外乱の分類精度の関係を示す．図 4 の各行が学習に用いた外乱の種類を示しており，各列が推論時の外乱を表している．また，各セルの数値は分類精度を示す．対角線部分が優れた精度を示すのは予想通りだが，非対角部分においても精度が向上しているケースが散見される（例えば，defocus blur を学習した場合）．このことから，ネットワーク構造だけでなく，学習に用いる外乱の種類も大きな影響を与えることが判明した．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関