# SwinTransformerとSwinUnetのマルチタスク学習を用いた画像認識

平田 航大[1,a)]　大北 剛[1,b)]

**概要**：本論文では, マルチタスク学習のトランスフォーマーによる手法である MTL-Swin-Unet を提案した. 疑似相関問題に陥った場合に, このマルチタスク学習のトランスフォーマーによる手法を用いることにより, セグメンテーションの表現と, 画像再構成タスクの表現の手助けを得ることにより, 分類器 SwinTransformer のみであるより, マルチタスク学習により洗練された表現を得ることができるのではというアイデアである. 実験では, この提案手法 MTL-Swin-Unet は, テストデータが同じ患者のスライスを含む設定 (共分散シフトではない設定) では F 値が, 他の分類器の性能より高い結果を得た. また、テストデータが同じ患者のスライスを含まない設定 (共分散シフトの設定) では,AUC が最も高い結果を得た.

# Image Recognition Using SwinTransformer and Multitask Learning of SwinUnet

***Abstract:*** This paper proposes a method MTL-Swin-Unet which is multi-task learning using transformers for classification and semantic segmentation. For sprious-correlation problems, this method allows us to enhance the image representation with two other image representations: representation obtained by semantic segmentation and representation obtained by image reconstruction. In our experiments, the proposed method outperformed in F-value measure than other classifiers when the test data included slices from the same patient (no covariance shift). Similarly, when the test data did not include slices from the same patient (covariance shift setting), the proposed method outperformed in AUC measure.

## 1. はじめに

画像認識は, 畳み込みニューラルネットワークをベースとする認識からビジョントランスフォーマーによる認識へと移行し, 事前学習を用いた形で大幅に性能を改善している. 画像の場合, ImageNet[1], CoCo などのデータセットにある物体を対象として事前学習をして、この事前学習したモデルを用いて転移学習やファインチューニングを行う. この形を取ることにより、事前学習をした画像データのドメインと対象とするドメインが異なる場合には, ドメイン間の差を縮めるドメイン適応と呼ばれる手法を用いたことになる. 事前学習した物体と類似するドメインであれば, このアプローチを行うことで目的とする結果を得ることができるが, 非常に例は少ないが, 対象とするドメインのデータが疑似相関 (spurious correlation)[2] に陥る場合には, この通常の対処方法では対処できないこともある.

疑似相関問題 [2] とは, 対象とする物体を認識するのが目的であっても, 対象ではない物体が不幸にも対象とする物体として機械学習モデルが構築されてしまう現象を言う. たとえば, 水鳥の画像が対象であった場合, 背景画像は往々にして, 湖/池/川/海である. 水鳥だからである. つまり, 湖/池/川/海などという水に関係する物体が、画像内に映っていることが多い. このよう場合、不幸にも、水鳥ではなく, 背景画像にある水に関係する物体が誤った対象として, 学習器が構築されてしまう. この現象を疑似相関と言う. ところで, 画像認識では, 予測の根拠を確認する方法があり, Grad-CAM [3] で色付けした画像, もしくは, アテンション重みを色付けした画像を見れば、このような疑似相関に陥っているかどうか判断できる. 構築した学習器が, 水鳥を学習していれば、水鳥が色付けされ, 水に関係する物体を学習していれば、別の場所が色付けされる.

[1] 九州工業大学大学院情報工学研究院知能情報工学研究系大北研究室
[a)] hirata.kodai612@mail.kyutech.jp
[b)] tsuyoshi@ai.kyutech.ac.jp

© 1959 Information Processing Society of Japan

本論文の対象は, 血腫の一種であるハイポデンシティーというマーカーを分類することである．この対象となる血腫マーカー以外にも, 他の血腫マーカーが 3 種類ある設定である．つまり, 正のクラスのハイポデンシティーというマーカーを分類する際に, 負のクラスでは 3 種類の血腫の画像が存在し, これらだけではなく, 血腫でない画像が多数含まれる．このような設定のため, 正のクラスが血腫を含む画像、負のクラスが血腫を含まない画像が, 疑似相関しやすい設定となる．試しに構築した学習器では, 血腫以外の頭骨や脳の中心部などに色付けされ, 疑似相関に陥っていることが容易に確認できた．我々の論文の対象の分類問題は, このような非常に特殊な場合に相当する．医用画像だからという理由ではなく, つまり, 対象とする部分が血腫であるから陥る場合ではないので注意されたい．

疑似相関問題に対するアプローチには, 重み付けによる方法 [4], サンプリングによる方法 [5], データ拡張による方法 [6], ポストキャリブレーションによる方法 [4], 最悪のグループエラーによる方法 [2] などが試されてきた．本論文では, これらとは異なり, 画像の表現を, 別の目的を用いて精練することにより, 全体としての精度を上げる方法を取る．自己教師あり学習は, ラベルづけされていない (大規模な) データを用いて, 画像の表現を学習する．この大規模なデータから学習した画像の表現を用いて, 後続タスクである本来の目的にラベルつきデータとして訓練する際に、補助的な情報として与えてやることで後続タスクの精度を上げる方法である．

## 2. 関連研究

直接的に同じ脳画像 CT を用いて、同じ血腫を目的とした関連研究は [7] である．ここでは, U-Net[8] を用いたセマンティックセグメンテーションと, Efficient-b3 畳み込みニューラルネットワーク [9] を用いたジョイント学習を行う．Efficient-b3 畳み込みニューラルネットワークを用いる際に, 本論文ではないが, ウェーブレット変換を CNN アーキテクチャに組み込むことで, 通常の CNN では失われやすいテクスチャ情報を保持する機構として Wavelet CNN[10] も用いる．Efficientnet は, ImageNet[1] で事前学習したモデルを用いている．セマンティックセグメンテーションと畳み込みニューラルネットワークのジョイント学習が効果的であったことを報告している．なお, この論文では CT 画像を患者ごとに選択した設定 (つまり, 共分散シフトの設定) も試しており、こちらの設定では性能が芳しくないことも報告されている．また、Wavelet CNN の効果は思ったほどの効果はなかったことも報告されている．

## 3. 提案手法

本章では，複数タスクを組み合わせるためにどのようにモデルを構築したかを述べる．マルチタスク学習とジョイント学習の 2 種類の方法で学習を行った．マルチタスク学習では，Swin-Unet[11] が持つセマンティック・セグメンテーションに，分類タスク・再構成タスクを追加して結果を比較した．ジョイント学習では，Swin-Unet[11] でセマンティック・セグメンテーションを行った後，SwinTransformer のエンコーダの表現を分類タスクに利用した．実験では 1 つ目の手法を MTL-Swin-Unet, 2 つ目の手法を Joint-SwinTransformer と呼んでいる．

### 3.1 Swin-Unet を用いたマルチタスク学習

#### 3.1.1 アーキテクチャ

1 つ目の提案手法のアーキテクチャを図 1 に示す．セマンティックセグメンテーションを目的とした Swin-Unet に，再構成タスク（reconstruction）と分類タスク（classification）を追加した．各タスクは，Swin Transformer エンコーダを共有している．エンコーダでは，画像の段階的なダウンサンプリングを行い，階層ごとの表現を生成する．セグメンテーションタスク，再構成タスクでは，エンコーダを対称にした形の Swin Transformer デコーダ [11] を用いる．デコーダでは，エンコーダによって抽出された表現と，スキップ接続を介して送られる各階層ごとの表現を用いて，入力画像の解像度に復元する．分類タスクでは，エンコーダから得られた表現を入力として線形層による分類を行う．

**セグメンテーション・画像再構成タスクのデコーダ**

セグメンテーションタスクと画像再構成タスクのデコーダには，Swin-Unet[11] のセグメンテーションタスクに利用されているデコーダを利用する．デコーダは，Swin Transformer エンコーダと対称的な構造を持っている．

各ステージの最初にエンコーダで得られた表現のアップサンプリングをパッチ拡張層によって行う．この層では，まず入力特徴量のチャネル方向を線形層を用いて 2 倍に拡張する $(h, w, C) \rightarrow (h, w, 2C)$．次に，形状変換を行う rearrange 処理によって，チャネル方向から解像度の方向にアップサンプリングを行う $(h, w, 2C) \rightarrow (2h, 2w, \frac{C}{2})$．この処理によって，パッチマージ層とは対称的に次元数を下げつつ解像度を上げている．

SwinTransformer ブロックの前には，U-Net[8] と同様にスキップ接続を行う．同じ階層のエンコーダの表現を保存しておき，チャネル方向に結合を行う．最後に，線形層によって 2 倍になった次元は元の次元に戻される．これによって，ダウンサンプリングによって失われる空間情報を補完する．

**分類ヘッド**

分類はエンコーダの最終出力にグローバル平均プーリングを適用し，線形層を通すことで行われる．ViT[12] や DeiT[13] などは最終出力を CLS-token としているためここではパッチを出力としているため，グローバル平均プー

© 1959 Information Processing Society of Japan

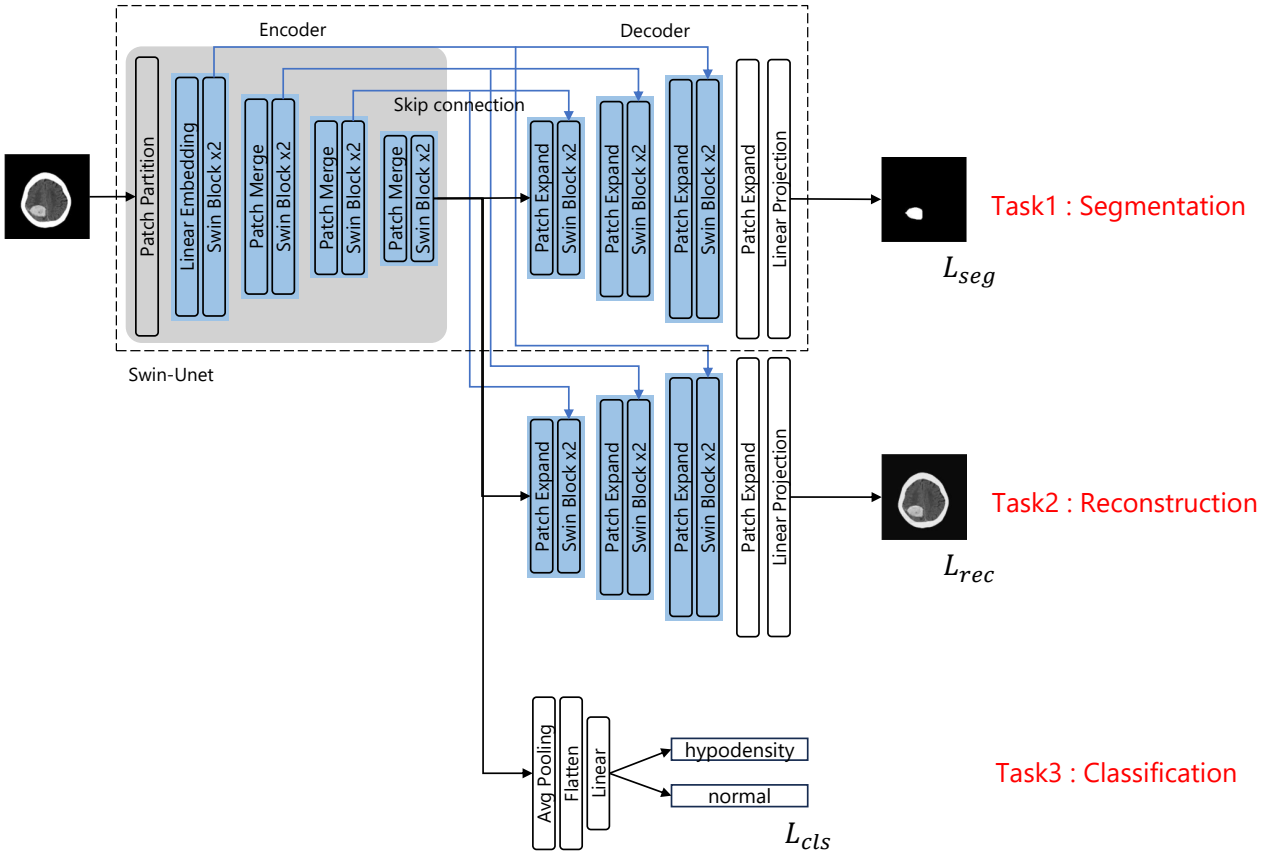


図 1: MTL-Swin-Unet のアーキテクチャ

リングで表現の集約を行う．

### 3.1.2 損失関数

セグメンテーションタスクは，画像内の対象物をピクセル単位で予測する．セグメンテーションタスクは難しいタスクのために，2 つの損失関数の重み付き和を用いる．1 つ目の損失関数は，セグメンテーションタスクの損失を直接最適化するためのクロスエントロピー関数である．以下の式で示される．

$$L_{ce}(p,q) = -\sum_i p_i \log q_i \tag{1}$$

ここで，$p$ は正解ラベル，$q$ は推論結果である．しかし，画像に占める対象物体が小さい場合は不均衡データとなってしまい，最適化がうまくいかない可能性がある．そのため，2 つ目の損失関数として Dice 損失を用いる．Dice 損失は正解の領域と推論結果の領域の重なり具合を計算し，以下の式で示される．

$$L_{dice}(p,q) = 1 - \frac{2\sum_i p_i q_i}{\sum_i (p_i + q_i)} \tag{2}$$

最終的なセグメンテーションタスクの損失は，2 つの損失関数の重み付き和となるので，以下の式で示される．

$$L_{seg} = \lambda_{ce} \cdot L_{ce} + \lambda_{dice} \cdot L_{dice} \tag{3}$$

ここで，$\lambda$ は各タスクの重みであり，実験では $\lambda_{ce} = 0.4$, $\lambda_{dice} = 0.6$ を用いている．画像再構成タスクは，デコーダの出力が入力画像を予測するように学習させる．再構成タスクでは平均二乗誤差を用いた．平均二乗誤差は以下の式で示される．

$$L_{rec}(p,q) = \frac{1}{n}\sum_i (p_i - q_i)^2 \tag{4}$$

クラス分類タスクでは，セグメンテーションタスクの 1 つ目の損失関数と同じように，クロスエントロピー関数を用いて損失 $L_{cls}$ を求めた．マルチタスク学習は，一般的に各タスクの損失の重み付き和で最適化を行い最小値を求める．最終的な損失は，次の式のように各タスクの損失の重み付き和によって算出される．

$$L = \lambda_{cls} \cdot L_{cls} + \lambda_{seg} \cdot L_{seg} + \lambda_{rec} \cdot L_{rec} \tag{5}$$

実験では，3 タスクのマルチタスク学習の場合 $\lambda_{cls} = 0.3$, $\lambda_{seg} = 0.4$, $\lambda_{rec} = 0.4$ を用いている．また，2 タスクのマルチタスク学習の場合 $\lambda_{cls} = 0.4$, $\lambda_{seg} = 0.6$，もしくは $\lambda_{cls} = 0.4$, $\lambda_{rec} = 0.6$ を用いている．最後に，セグメンテーションタスクと分類タスクを同時に行うとき，分類にはラベルが存在するがセグメンテーションにはラベルが存在しない場合がある．例えば，血腫のない画像にはセグメンテーションのマスク画像が存在しない．そのため，セグメンテーションのマスクが存在する画像のみで損失の計算

© 1959 Information Processing Society of Japan

を行い，バッチごとに平均することでセグメンテーションの損失としている．

### 3.2 Swin-Unet を用いたジョイント学習

#### 3.2.1 アーキテクチャ

2 つ目の提案手法のアーキテクチャを図 2 に示す．この学習方法は, 先行研究 [7] に基づいている．最初に, Swin-Unet を用いて血腫の画素を予測するセグメンテーションタスクを行う．次に，Swin-Unet で学習させたエンコーダのパラメータを凍結させる．最後に，通常の SwinTransformer エンコーダと凍結させたエンコーダの表現を結合させて，その表現を分類タスクに利用する．

**セグメンテーションタスクの表現を持つエンコーダ**

ジョイント学習の事前のセグメンテーションタスクとして，Swin-Unet を利用して学習した．Swin-Unet のエンコーダは，SwinTransformer のエンコーダと同じ構造である．SwinTransformer ブロックでは表現の学習を行う．セグメンテーションタスクによって学習されたエンコーダは，セグメンテーションタスクの固有の表現を持つ．これを分類タスクに利用する．事前のセグメンテーションタスクの学習結果は節**??**に示している．

**クラス分類時のエンコーダ**

クラス分類タスクを行うときのエンコーダとして，Swin-Transformer エンコーダと，セグメンテーションタスクによって学習された SwinTransformer エンコーダを用いる．具体的には，入力された画像は学習可能なエンコーダと，節 3.2.1 で学習を行ったエンコーダを学習不可能にしたエンコーダにそれぞれ入力する．それぞれのエンコーダから出力された特徴ベクトルは，チャネル方向に結合を行い 1 つのベクトルとする．SwinTransformer エンコーダのパッチマージ層はダウンサンプリングの役割を持つ．各ステージの前にパッチマージングがされるため，ステージの深さが同じであれば解像度も同じである．セグメンテーションタスクによって学習されたエンコーダと，学習可能なエンコーダは同じ深さのステージなので，2 つのエンコーダから出力されるベクトルはそのままチャネル方向に結合している．

**分類ヘッド**

分類ヘッドは，節 3.1.1 と同じ構造を用いている．ただし，結合された表現を入力できるように線形層の入力次元数を変更している．

#### 3.2.2 損失関数

1 つ目の手法と同様に，事前のセグメンテーションタスクではクロスエントロピー関数と Dice 損失の混合の損失関数を用いて，クラス分類タスクではクロスエントロピー関数を用いた．

## 4. 設定と実験結果

この章では，血腫がハイポデンシティであるかどうかに対して，2 つの手法を用いて実験を行った．

### 4.1 データセット

#### 4.1.1 使用するデータセット

本研究では，CT 画像を用いるが, これは先行研究 [7] で使用したものと同じである．データは，DICOM 形式で保存されており，ピクセル値の単位はハウンズフィールド（HU）である．収集されたデータは，HU 範囲に従ってコントラストが調整されている．また，1 から 4 の施設に対してはセグメンテーションタスクのアノテーションデータが存在する．このデータは，血腫が存在する画素に対してマスクされている．

血腫にはハイポデンシティ，マージンイレギュラーサイン，ブレンドサイン，フルイドレベルなどと言った血腫の見え方の異なる名前のついた急性脳内血腫（ICH）のマーカーが存在する (図 3 参照). これらは別々の医師が ICH のマーカーとして有用そうなものを示し, これらが ICH にどう関わるかは, Boulouis ら [14] が議論する．このため，4 つのクラスは重複可能であり, マルチラベル問題となる．実験では，ハイポデンシティのみに注目して, ハイポデンシティであるかどうかを分類するタスクとして考える．つまり，データセットには，血腫が無い画像と 4 つの血腫クラスを重複して持つ画像があるが，ハイポデンシティである画像をすべて分類することをタスクとして設定した．

#### 4.1.2 データ設定

実験では，1 から 11 の施設の 11780 枚を使用した．施設 1 の 1 割である 179 枚をテストデータ，施設 1 の残りの 9 割と施設 2 から 4 の 9912 枚のうち 8 割を訓練データ，残りの 2 割を検証データとして扱った．このデータセットは，同じ患者の CT 画像を複数のスライスに分割しているため，訓練データとテストデータに同じ患者のスライスが入っている可能性がある．そこで，同じ患者のスライスがテストデータに入らない場合を評価するために，5 から 11 の施設の 1868 枚から 179 枚をランダムに選択しテストデータとした．ここで，陽性の比率が施設 1 からとったテストデータと同じになるように選択している．実験では施設 1 からとったテストデータをテストデータ（病院 1-4），5 から 11 の施設からとったテストデータをテストデータ（病院 5-11）として，2 種類のテストデータに評価している．

訓練データは病院 1-4 から構成される患者の CT 画像であり, テストデータ (病院 1-4) は同じ病院 1-4 から構成される患者の CT 画像である。このことから, 同じ患者の同じ撮影日の CT 画像のスライスが使われている可能性がある．一方, テストデータ (病院 5-11) は, 訓練データが病院

© 1959 Information Processing Society of Japan

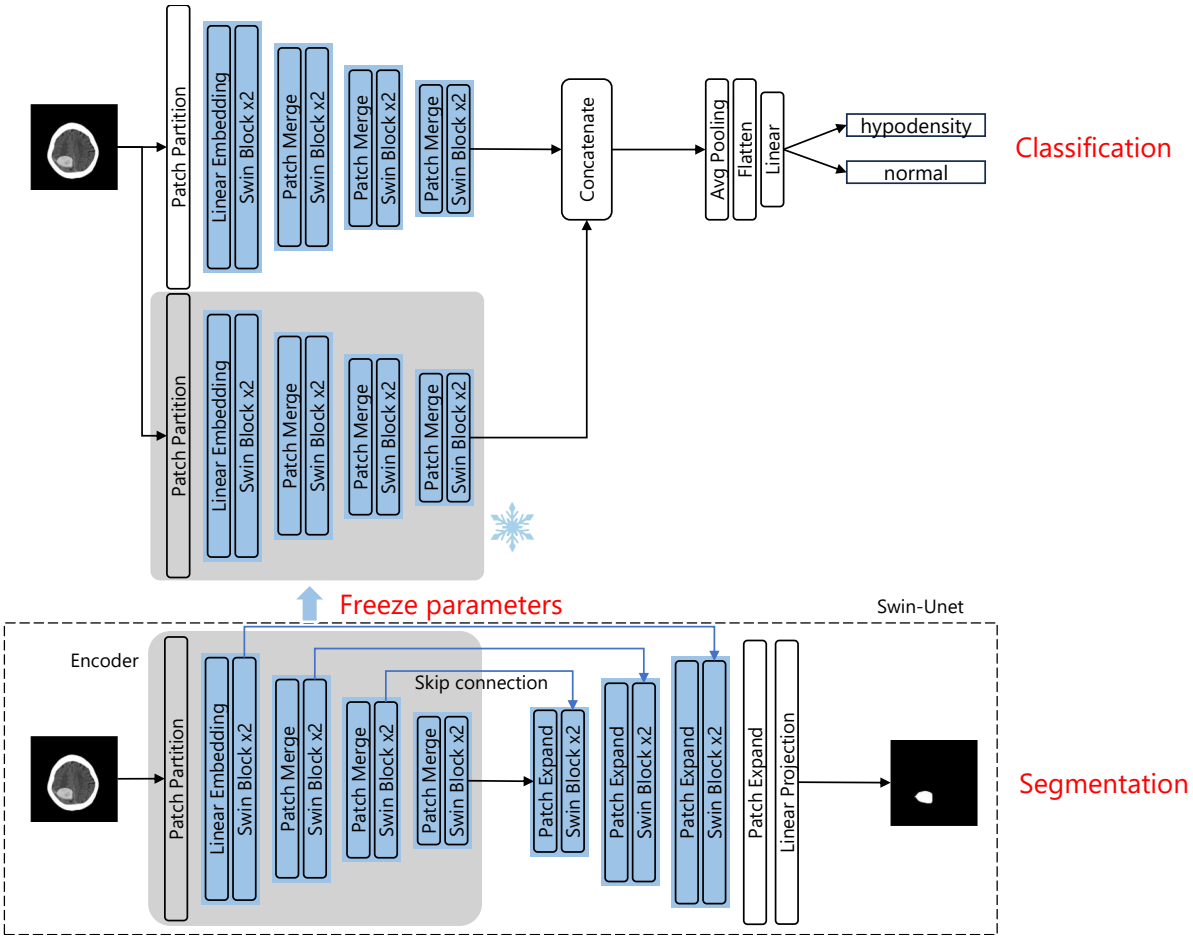


図 2: Joint-SwinTransformer のアーキテクチャ


図 3: （a）ハイポデンシティ （b）ハイポデンシティ以外の，血腫クラスが重複していない画像（左から，血腫のない画像，マージンイレギュラーサイン，ブレンドサイン，フルイドレベル）

1-4 からの患者の CT 画像である. つまり, 同じ患者の同じ撮影日の CT 画像のスライスが使われていない. 前者は, 分類として見た場合に共分散シフトの設定ではなく, 一方, 後者は共分散シフトの設定だと言える.

### 4.2 学習器の設定

1 つ目の提案手法である MTL-Swin-unet では，4 種類の設定で実験を行った．分類タスクがメインタスクなので，分類タスクと画像再構成タスク（cls + rec），分類タスクとセグメンテーションタスク（cls + seg），分類タスクとセグメンテーションタスクと画像再構成タスク（cls + seg + rec）の 3 つの組み合わせで実験を行った．これに加えて，3 つのタスクを行うモデル（cls + seg + rec）ではモデルサイズを大きくした tiny で実験を行った．2 つ目の提案手法である Joint-SwinTransofrmer では，通常サイズと tiny サイズで実験を行っている．分類タスクの比較手法として，ResNet152[15]，Swin Transformer[16]，Swin Transformer tiny[16]，joint-Learning[7] で実験を行った．

© 1959 Information Processing Society of Japan

セグメンテーションタスクの比較手法として，U-Net[8] と Swin-Unet[11] で実験を行った．

すべての学習について，Python3.7, pytorch1.13.1 で行った．学習の設定はほとんど [11] に従っている．入力の画像サイズは 224 × 224 にリサイズしている．各モデルのエンコーダは，ImageNet[1] によって事前学習されたパラメータによって初期化している．最適化手法はモーメンタム 0.9，L2 正則化 0.0001 の確率的勾配降下法を用いて，分類タスクを含む学習は 600 エポック，セグメンテーションタスクのみの学習は 200 エポックで行った．学習率の初期値とバッチサイズは学習によって異なる．MTL-Swin-Unet-tiny（cls + seg + rec）ではバッチサイズ 32，学習率 0.01 を用いた．他の Swin-Unet を用いたマルチタスク学習と Swin Transformer[16] ではすべてバッチサイズ 64，学習率 0.01 を用いている．Joint-SwinTransformer と Joint-Learning[7] ではすべてバッチサイズ 128，学習率 0.004 で学習させた．ResNet[15] はバッチサイズ 128，学習率 0.01 で学習させた．最後に，U-Net[8] と Swin-Unet[11] はバッチサイズ 32，学習率 0.01 で学習させた．すべての学習で学習率の減衰を行っている．学習率 $lr$ はイテレーション $iter$ ごとに以下の式に従って減少する．

$$lr = lr_{base} \times (1 - \frac{iter}{iter_{max}})^{0.9} \qquad (6)$$

ここで $lr_{base}$ は学習率の初期値，$iter_{max}$ はイテレーションの最大値である．

入力画像とセグメンテーションのマスク画像に同様のデータ拡張を行っている．この実験では，次の処理を組合せることでデータ拡張を行っている．

- 0，90，180，270 度の回転がランダムに適用された後，上下か左右のフリップをランダムに行う．
- -20 から 20 度の間で， ランダムに回転を行う．

それぞれ，5 割の確率で独立に適用される．つまり，データに変形が加えられない場合，どちらか一方のデータ拡張が適用される場合，どちらも適用される場合が存在する．

最後に，明示しない場合には Swin Transformer[16] をアーキテクチャに持つモデルのエンコーダの 3 ステージ目のブロック数を 2 個に，入力時のチャネルを 96 に設定している．Swin Transformer[16] のモデルサイズは，この 3 ステージ目のブロック数と入力チャネル数によって決まる．実際に tiny はブロック数 6，チャネル数 96，base はブロック数 18，チャネル数 128 と異なる．Swin-Unet[11] の研究ではモデルサイズによる性能の影響を調べているが，モデルを大きくしても大きな性能の向上はみられなかった．そのため [11] のコードではエンコーダの 3 ステージ目のブロック数を 2 個にしたものをデフォルトにしている．明示しない場合はこの設定に従っている．

### 4.3 実験結果

表 1 は，テストデータ（病院 1-4）とテストデータ（病院 5-11）に対する，提案する MTL-Swin-Unet とその比較手法の結果である．結果から，MTL-Swin-Unet（cls + seg + rec）はテストデータ（病院 1-4）において accuracy が 0.961，F 値が 0.903 であり最も高い値を得た．テストデータ（病院 5-11）では，MTL-Swin-Unet（cls + seg + rec）は F 値では Joint-SwinTransformer と比べて劣っていたが，AUC は 0.799 で最大の値になった．

**異なるタスクによって学習される表現が分類タスクに与える影響**

表 1 から，MTL-Swin-Unet（cls + seg）は Swin Transformer[16] に比べて AUC がテストデータ（病院 1-4）で 0.004，テストデータ（病院 5-11）で 0.016 上回った．Joint-SwinTransformer も Swin Transformer[16] に比べて AUC がテストデータ（病院 1-4）で 0.011，テストデータ（病院 5-11）で 0.015 上回った．一方で，MTL-Swin-Unet（cls + rec）は Swin Transformer[16] に比べて，AUC がテストデータ（病院 1-4）で 0.008，テストデータ（病院 5-11）で 0.017 下回った．このことから，画像再構成タスク単体では分類タスクと共通する有利な要因はなく，むしろ相反する情報を持つと考えられる．一方で，セグメンテーションタスクは分類タスクにとって有利な表現を学習できることが分かる．また，MTL-Swin-Unet（cls + seg + rec）の 3 つのタスクを行う場合は，MTL-Swin-Unet（cls + rec）に比べて，AUC がテストデータ（病院 1-4）では変化がみられなかったが，テストデータ（病院 5-11）で 0.011 上回った．このことから，画像再構成タスクは単体では分類タスクに有利な影響を与えないが，セグメンテーションタスクと同時に学習させることで，MTL-Swin-Unet（cls + rec）よりも分類タスクに有利な表現を学習できることが分かった．

**モデルサイズの影響**

MTL-Swin-Unet，Joint-SwinTransformer，Swin Transformer でモデルサイズの大きい tiny で実験を行った．表 1 から，テストデータ（病院 1-4）に対する MTL-Swin-Unet-tiny（cls + seg + rec）で MTL-Swin-Unet（cls + seg + rec）と比較して AUC の 0.006 の上昇，テストデータ（病院 5-11）に対する Joint-SwinTransformer-tiny で Joint-SwinTransformer と比較して AUC の 0.001 の上昇がみられたが，それ以外の場合では 0.004 から 0.031 の低下がみられた．

#### 4.3.1 予測根拠の確認

脳の CT 画像では，モデルは頭骨に注目して分類を行ってしまう場合がある．しかし，実際は血腫や脳のテクスチャに重要な情報がある．そこで，予測根拠の可視化を Grad-CAM[3] で行った．Grad-CAM[3] は入力画像のどの部分が予測に最も影響を与えたかをヒートマップとして出力することで，予測に対する視覚的説明を得ることがで

© 1959 Information Processing Society of Japan

| | TestData(Hospital1-4) | | | | | | TestData(Hospital5-11) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | iou-seg | acc | prec | rec | f1 | auc | acc | prec | rec | f1 | auc |
| U-Net[8] | .692 | - | - | - | - | - | - | - | - | - | - |
| Swin-Unet[11] | .808 | - | - | - | - | - | - | - | - | - | - |
| ResNet152[15] | - | .901 | .655 | .333 | .692 | .804 | **.822** | .126 | .095 | .504 | .511 |
| Swin Transformer[16] | - | .956 | .813 | **.819** | .895 | .963 | .746 | .254 | .600 | .599 | .772 |
| SwinTransformer-tiny[16] | - | .952 | .825 | .752 | .880 | .959 | .755 | .240 | .505 | .587 | .741 |
| Joint-Learning[7] | - | .945 | .792 | .724 | .863 | .913 | .787 | .208 | .305 | .560 | .631 |
| Joint-SwinTransformer | - | .956 | **.859** | .752 | .888 | **.974** | .797 | **.317** | .638 | **.650** | .787 |
| Joint-SwinTransformer-tiny | - | .953 | .857 | .724 | .879 | .950 | .782 | .287 | .581 | .626 | .788 |
| MTL-Swin-Unet(cls+rec) | - | .949 | .805 | .743 | .871 | .955 | .736 | .238 | .571 | .586 | .755 |
| MTL-Swin-Unet(cls+seg) | **.815** | .956 | .823 | .800 | .893 | .967 | .757 | .256 | .648 | .617 | .788 |
| MTL-Swin-Unet(cls+seg+rec) | .811 | **.961** | .852 | .809 | **.903** | .967 | .756 | .276 | .667 | .618 | **.799** |
| MTL-Swin-Unet-tiny(cls+seg+rec) | .773 | .960 | .841 | .810 | .901 | .973 | .752 | .282 | **.705** | .623 | .790 |

表 1: 半分より上にベースラインシステムとして 6 モデル, 半分より下に提案手法の 6 モデルを示す. 提案手法はマルチタスク学習である MTL-SwinUnet で始まる 4 モデル, Joint-SwinTransformer で始まる 2 モデルを指す. 後者は, SwinTransformer を使った形で, Joint-Learning の畳み込みニューラルネットワーク (EfficientNet) を SwinTransformer を使った形に変形する形のモデルを指す. 左列のテストデータ (病院 1-4) は, 訓練とテストデータで, 同じ患者からのオーバーラップしないスライスを許した場合を示す. テストデータ (病院 5-11) は訓練とテストデータで, スライスは別の患者のスライスを使った場合を指す. マルチタスク学習では 3 クラスのマルチタスクの場合と 2 クラスのマルチタスクの場合があり, タスク名はそれぞれ，cls：クラス分類タスク，seg：セグメンテーションタスク，rec：画像再構成タスクを示す.

きる.

図 4 は，MTL-Swin-Unet（cls + seg + rec）の Grad-CAM[3] による可視化結果である.（a）はハイポデンシティの画像に対する可視化結果であり，（b）は血腫はあるがハイポデンシティでない画像の可視化結果である. どちらの場合も，血腫やその周りに注目して予測を行っていることが分かる.

## 5. 結論

本論文では, マルチタスク学習のトランスフォーマーによる手法である MTL-Swin-Unet（cls+seg+rec）を提案した. 疑似相関問題に陥った場合に, このマルチタスク学習のトランスフォーマーによる手法を用いることにより, セグメンテーションの表現と, 画像再構成タスクの表現の手助けを得ることにより, 分類器 SwinTransformer のみであるより, マルチタスク学習により洗練された表現を得ることができるのではというアイデアである. 実験では, この提案手法 MTL-Swin-Unet（cls+seg+rec）は, テストデータが同じ患者のスライスを含む設定 (共分散シフトではない設定) では F 値が, 他の分類器の性能より高い結果を得た. また、テストデータが同じ患者のスライスを含まない設定 (共分散シフトの設定) では, AUC が最も高い結果を得た.

## 参考文献

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[2] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations, 2020.

[3] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, Vol. abs/1610.02391, , 2016.

[4] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2023.

[5] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy, 2022.

[6] Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. Selecmix: Debiased learning by contradicting-pair sampling, 2022.

[7] Hokuto Hirano and Tsuyoshi Okita. Classification of hematoma: Joint learning of semantic segmentation and classification, 2021.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

[9] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, Vol. abs/1905.11946, , 2019.

[10] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. Wavelet convolutional neural networks. *CoRR*, Vol. abs/1805.08620, , 2018.

[11] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmen-

© 1959 Information Processing Society of Japan

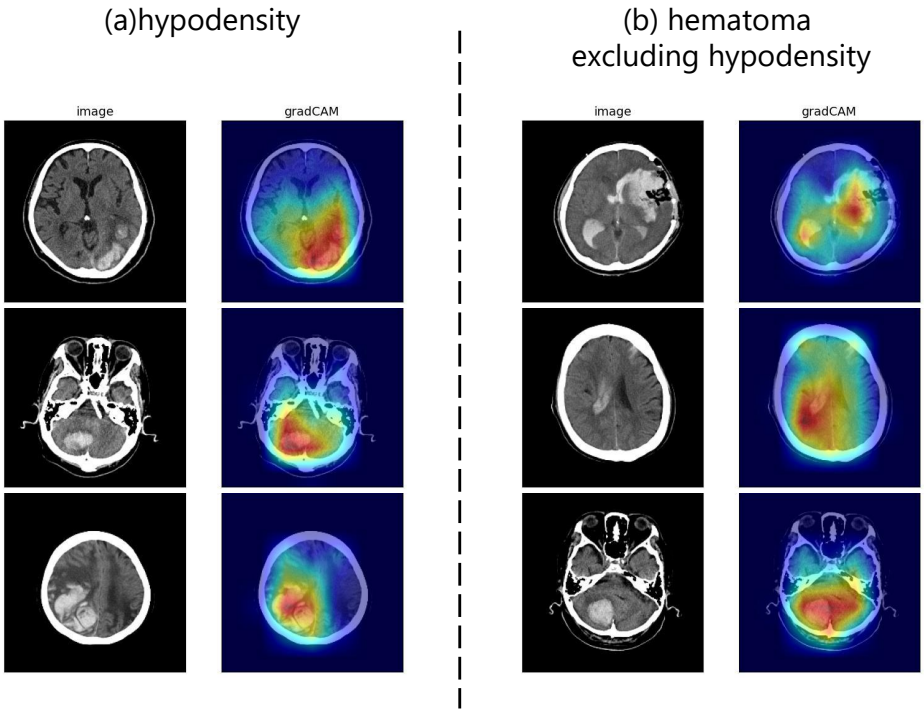


図 4: MTL-Swin-Unet（cls + seg + rec）による Grad-CAM[3] の可視化結果 （a）ハイポデンシティの画像 （b）血腫はあるがハイポデンシティでない画像

tation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

[14] Gregoire Boulouis, Andrea Morotti, H Bart Brouwers, Andreas Charidimou, Michael J Jessel, Eitan Auriel, Octávio Pontes-Neto, Alison Ayres, Anastasia Vashkevich, Kristin M Schwab, et al. Association between hypodensities detected by computed tomography and hematoma expansion in patients with intracerebral hemorrhage. *JAMA neurology*, Vol. 73, No. 8, pp. 961–968, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, Vol. abs/1512.03385, , 2015.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

© 1959 Information Processing Society of Japan