機械学習モデルによる画像認識に関する簡易コンサルティング

大見　一樹† 川合　諒†† 吉田　登†† 劉　健全††

† 名古屋工業大学工学専攻情報工学プログラム　〒 466–0061 愛知県名古屋市昭和区御器所町
†† NEC ビジュアルインテリジェンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部
E-mail: †k.omi.646@nitech.jp, ††{r-kawai.az,n-yoshida14,jqliu}@nec.com

あらまし　現在，コンピュータビジョンによる様々なタスクが多くの分野で用いられているが，全ての画像がどんなタスクでも期待通りの認識結果が得られるわけではない．画像認識の知識がない素人には期待通りの認識結果が得られる見込みのあるタスクの判断が難しく，コンピュータビジョンの利用を制限してしまう．そこで，本論文では画像の適切なタスクを分類するという新しいタスクとその手法を提案する．また，認識率に大きな影響を与える被写体サイズに着目し，適していないと判断されたタスクでも画像の拡大，縮小を行うことで認識可能となる場合はそれらをレコメンドする手法を提案する．本研究は画像認識に関する簡易コンサルティングとなり，コンピュータビジョンの普及やビジネスチャンスの拡大に繋がると考えられる．

キーワード　画像認識，ユーザー支援，評価・データセット

1 はじめに

現在，顔認識 [30]，歩行者検出 [18]，文字認識 [29]，医用画像診断 [34]，画像キャプション生成 [33], 深度推定 [6] など様々なコンピュータビジョンによるタスクが多くの分野で用いられており，今後も応用先は広がり続け，コンピュータビジョンはさらに身近に利用できるものになると考えられる．近年では，コロナの影響もあり，人と人との距離を分析するタスクなども研究されており [2]，コンピュータビジョンにおけるタスクは今後もさらに増え続けていくと予想される．

ただし，画像それぞれに適応可能なタスクがあり，どんなタスクでも期待通りの認識結果が得られるわけではないと考えられる（図 1）．姿勢推定や顔認識などのタスクが適する条件の 1 つは画像内に人が存在することであるが．それだけでは十分ではない．例えば被写体の大きさによっても適切なタスクは異なる．被写体が小さい場合は姿勢推定などの全身を認識するタスクに向いているが，顔認識のような被写体の詳細を認識するようなタスクには向いていない．逆に被写体が大きい場合はこれらのタスクに関する向き不向きが逆になることがある．また，ある 1 つのタスクにおいても，画像の解像度，画質，オクルージョンの程度によって検出可能であるかは異なる [3, 14]．他にも多くの要因が絡み，画像によって適切なタスクは様々である．

これらはコンピュータビジョンの専門家であれば知識や経験から画像の適切なタスクを判断することが可能かもしれないが，画像認識の知識のない素人には判断が難しい．また，専門家であっても判断の難しい画像がある．しかし，どういった画像がどのタスクに適しているかは研究されておらず，モデルに画像を入力して出力が得られるまで期待通りの認識結果が得られるかどうかはわからない．また，全てのタスクで試行するのはコストや時間が必要となってしまう．このような現状は，特に素人によるコンピュータビジョンの利用を制限してしまう．

そこで本論文では与えられた画像の適切なタスクを分類する手法を提案する．コンピュータビジョンにおける様々なタスクが活発に研究されているが，画像がどのタスクに向いているかは注目されていない．しかし，機械学習モデルにより画像がどのタスクに適しているか分類することができれば，画像認識の知識がなくても画像さえあれば適切な用途が分かり，より一層コンピュータビジョンの普及を広げ，さらにはビジネスチャンスの拡大も考えられる．言い換えれば，これは画像認識に関する簡易的なコンサルティングを可能にするといえる．さらに、より優れたコンサルティングの第一歩として，特に認識率に大きな影響を与える被写体サイズに関するレコメンドについても本研究で扱う．具体的には，適していないと判断された場合でも，拡大もしくは縮小することで認識可能な場合はその旨出力することで，ユーザーに画像の調整や撮影条件の変更などを促す．なお，ここで拡大・縮小というのはリサイズではなく，拡大縮小行列をかけること，すなわち，全体の画像のサイズはそのままで被写体サイズを変更することを指す．

本論文で提示するタスクは与えられた画像の適切なタスクを判定するものであり，タスクごとに適しているか適していないかの 2 値分類を行うため，マルチラベル学習で実現できる．ただし，そのようなラベル付けされたデータセットは存在しないため，画像を収集しラベル付けを行う必要がある．その方法については 3.1 節で述べる．また，1 つのタスクを対象とした一般的なマルチラベル分類とは異なり，顔認識，文字認識，天気認識のように多種多様なタスクで共通して適用できるモデルを構築する必要がある．我々の提案するモデルが多種多様なタスクで共通して有効であることを実験的に示す．本研究の貢献は以下の通りである．

- 画像の適切なタスクを分類する手法を提案する．著者の知る限り，本研究は画像の適切なタスクの分類に初めて取

図 1: それぞれの画像ごとに適応可能なタスクがあり，どんなタスクでも適しているわけではない．コンピュータビジョンの知識がない素人には特にこれらの判断は難しい．

り組んだ研究である．

- 画像認識の知識がなくてもどういったタスクで用いることができるか判断することを可能にし，よりコンピュータビジョンを身近な存在にする．
- 認識率に大きな影響を与える被写体サイズに着目し，画像の拡大または縮小を行うことで適切なタスクがある場合はそれらのレコメンドを行う．

2 関 連 研 究

2.1 マルチラベル分類

近年，シングルラベル分類の発展に伴いマルチラベル分類も盛んに研究されている．マルチラベル分類とは，複数のラベルの有無を予測するタスクのことを指す．マルチラベル分類では，ラベル間の相関関係をモデル化するための手法 [15, 16] やそのためのグラフニューラルネットワークやリカレントニューラルネットワークなど [5, 35] が提案されている．他にも，ラベル分布が不均衡であるという問題に対して，画像のサンプリングや損失計算などの手法が提案されている [4, 20, 27]．これらの一般的なマルチラベル分類はある 1 つのタスクにおけるクラス識別において，識別するクラスを 1 つから任意の個数に拡張する目的のために用いられる．目的は異なるが，本研究で我々が提示するタスクも，タスクごとに適しているか適していないかの 2 値分類を行うマルチラベル学習が利用できる．

2.2 被写体サイズと認識率

物体検出やクラス識別などのタスクにおいて，被写体サイズは認識率に大きな影響を与えることが知られている [13, 37]．そのため，マルチスケールな特徴量の使用，アンカー設定の工夫，畳み込みにおけるフィルタサイズの動的変更などにより，被写体サイズにロバストな手法は数多く研究されている [11, 12, 19, 38] ．特にマルチスケールな特徴量を扱う手法は近年活発に研究されており，処理速度向上や，異なるサイズの特徴量の融合やプーリングまで様々な手法が提案されている [8, 19, 22, 32]．ただし，これらの手法はいずれも被写体サイズのバリエーションが豊富な訓練データで学習する必要

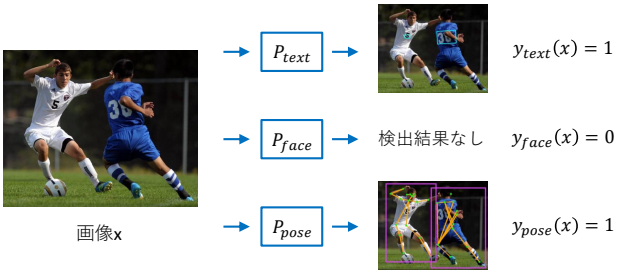

図 2: 適切なタスクのラベリング方法．

がある．また，姿勢推定では被写体が小さく全身が写っていることが好ましい，顔認識では被写体が大きく顔などの詳細が写っていることが好ましいといった，タスクごとに本来認識しやすい被写体サイズがあり，上で述べた手法であっても認識しやすい被写体サイズが存在すると考えられる．そこで本研究では，認識結果が得られなかった場合でも画像の拡大または縮小を行い，認識しやすい被写体サイズに変えることで検出可能となる場合にそれらのレコメンドを行う．

3 提 案 手 法

3.1 アノテーション

3.1.1 適切なタスクのアノテーション

初めに，画像の適切なタスクのラベルを作成する処理について説明する．そのようなラベル付けがなされた画像データセットは存在しないので，タスク t をあらかじめ学習したモデル P_t において高い出力スコアが得られた画像はそのタスクに適していると考え positive ラベルとし，そうではなかったものを negative ラベルとする（図 2）．すなわち，画像 x のタスク t におけるラベル $y_t(x)$ を以下のように定義する．

$$y_t(x) = \begin{cases} 1 & (P_t(x) \geqq \tau_t) \\ 0 & (P_t(x) < \tau_t) \end{cases} \tag{1}$$

ここで τ_t はタスク t における閾値である．なお，検出された物体ごとにスコアが出力される物体検出タスクのように，出力スコアが複数ある場合は最も高いスコアを用いる．また，学習時においてモデルの出力スコア $P_t(x)$ の値をそのままソフトラベルとして用いることができ $(y_t(x) = P_t(x))$，その有効性は 4.2.2 節で示す．

3.1.2 拡大縮小アノテーション

次に拡大または縮小レコメンドを学習するためのラベルを作成する方法について説明する．元画像 x に回転角度を 0，回転中心の座標 c，スケール s の回転行列をかけた画像を $x^{c,s}$ とし，タスク t において座標 c を中心にスケール s で拡大または縮小をレコメンドすべきかのラベル $y_t^{c,s}(x)$ を以下のように定義する．

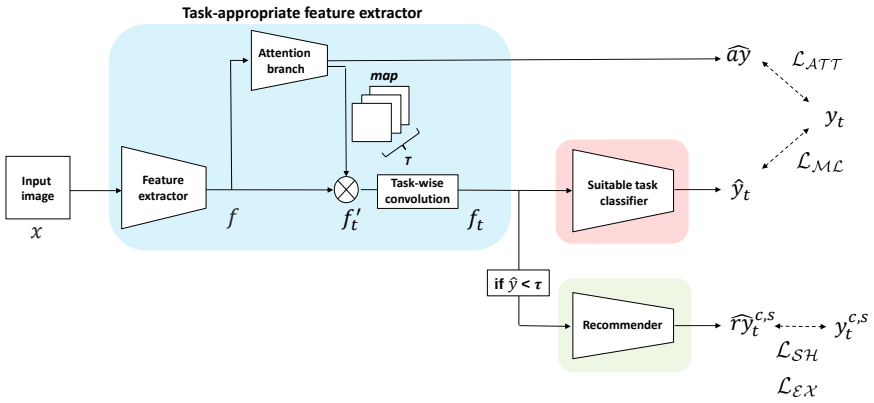

図 3: 提案手法の全体像．タスクごとに注目領域に重みをかけ適切な特徴量を抽出する Task-appropriate feature extractor，それぞれのタスクで適切であるかを分類する Suitable task classifier，適切でないと判断された場合でも入力画像を拡大縮小することで認識可能となる場合にそれらのレコメンドを行う Recommender からなる．

$$y_t^{c,s}(x) = \begin{cases} 0 & ((y_t(x)=0) \wedge (y_t(x^{c,s})=0)) \\ 1 & ((y_t(x)=0) \wedge (p_t(x^{c,s})=1)) \\ 2 & ((y_t(x)=1) \wedge (y_t(x^{c,s})=0)) \\ 3 & ((y_t(x)=1) \wedge (y_t(x^{c,s})=1)) \end{cases} \quad (2)$$

縮小をレコメンドする際に縮小を行う中心座標は重要ではないと考えられるため，c は元画像の中心とする．また，元画像で適切なタスクではないと判断された場合のみレコメンドを行いたいので，学習時には $y_t^{c,s}(x) \in \{0,1\}$ の画像のみを用いる．

3.2 アーキテクチャ

提案手法の全体像を図 3 に示す．我々の提案するモデルはそれぞれのタスクごとに効率よく適切な特徴抽出を行う Task-appropriate feature extractor，それぞれのタスクで適切であるかを分類する Suitable task classifier，適していないと判断された場合でも拡大または縮小によって適切なタスクがある場合はそれらのレコメンドを行う Recommender の 3 つからなる．Recommender の学習は Task-appropriate feature extractor と Suitable task classifier の学習後にそれらのパラメータを固定して行う．

3.2.1 Task-appropriate feature extractor

ある画像 $x \in \mathbb{R}^{3\times H\times W}$ から特徴抽出器 E を用いて得られた特徴量を $f = E(x) \in \mathbb{R}^{C\times H'\times W'}$ とする．ニューラルネットワークの表現力に期待し，特徴抽出器は全てのタスクで共通のものを用いる．しかし，同じ画像であっても対象のタスクごとに注目すべき領域は異なると考えられる [24]．そこで Attention Branch Network [7] を用いてタスクごとに異なる attention map を作成し，それを反映させることでタスクごとに必要な領域に注目させる．

$$f'_t = f \cdot map_t + f \quad (3)$$

ここで $t = 1, 2, ..., T$ であり，T はタスクの数を表す．また map_t は attention branch により得られたタスク t における attention map である．[7] に従い attention branch の出力 $\hat{ay}$ からも損失を計算し，この損失を $\mathcal{L}_{ATT}$ とする．また，それぞれのタスクごとに抽出すべき特徴も異なると考えられるので，アテンションマップが適用された特徴量はその後，タスクごとに異なる畳み込み TC_t を用いてさらに洗練される．

$$f_t = TC_t(f'_t) \quad (4)$$

3.2.2 Suitable task classifier

Task-appropriate feature extractor によって得られた特徴量はタスクごとに異なる出力層を持つ 2 値分類器 C に入力され予測 $\hat{y}$ を得る．

$$\hat{y}_t = C(f_t) \quad (5)$$

損失は以下のように定義する．

$$\mathcal{L}_{ML} = -\sum_{t=1}^{T} (y_t \log \hat{y}_t + (1-y_t) \log (1-\hat{y}_t)) \quad (6)$$

最終的な損失は次式で表される．

$$\mathcal{L} = \mathcal{L}_{ML} + \lambda \mathcal{L}_{ATT} \quad (7)$$

ここで λ は重みである．以上の損失により Task-appropriate feature extractor と Suitable task classifier を学習させる．

3.2.3 Recommender

推論時に，Suitable task classifier によって適していないと判断された画像は Recommender に入力され，拡大または縮小によって適切なタスクとなる場合はそれらのレコメンドを行う．Recommender の学習は Task-appropriate feature extractor と Suitable task classifier の学習後にそれらのパラメータを固定してから行う．これらのレコメンドはタスクごとに学習が行われる．学習時はバッチサイズをそろえるため，あらかじめ元画像で negative ラベルであった画像のみ

を用いる．

まずは縮小のレコメンドを行う Shrink Recommender について説明する．タスク t においてスケール $s(<1)$ で縮小すべきかどうかを予測する 2 値分類器を R_t^s とし，その出力 $\hat{ry}_t^{c,s} = R_t^s(f_t)$ から損失を以下のように定義する．

$$\mathcal{L}_{\mathcal{SH},t} = -y_t^{c,s} \log \hat{ry}_t^{c,s} - (1 - y_t^{c,s}) \log (1 - \hat{ry}_t^{c,s}) \quad (8)$$

ここで c は元画像の中心座標とする．

次に拡大のレコメンドを行う Expand Recommender について説明する．ここでは拡大のスケールを 2 の場合について説明し，スケールを拡張する方法は後に述べる．タスク t においてスケール 2 で拡大をすべきかどうかを予測する 2 値分類器を R_t^2 とする．ある領域を拡大すべきかどうかはその領域にのみ焦点を当てればよいので，元画像から得られた特徴量 $f \in \mathbb{R}^{C \times H' \times W'}$ から対象の領域を取り出す．スケール 2 の場合は拡大領域 4 箇所を以下のように取り出す．

$$\begin{aligned} f_t^{ul,2} &= f_t[:, 0 : \frac{H'}{2}, 0 : \frac{W'}{2}] \\ f_t^{ur,2} &= f_t[:, 0 : \frac{H'}{2}, \frac{W'}{2} : W'] \\ f_t^{ll,2} &= f_t[:, \frac{H'}{2} : H', 0 : \frac{W'}{2}] \\ f_t^{lr,2} &= f_t[:, \frac{H'}{2} : H', \frac{W'}{2} : W'] \end{aligned} \quad (9)$$

ここで ul, ur, ll, lr は元画像を 4 分割した左上，右上，左下，右下それぞれの中心座標を表す．これらの特徴量はすべて R_t^2 に入力され，出力 $\hat{ry}_t^{p,2} = R_t^2(f_t^{p,2})$ とそれぞれのラベルとロスを取る．

$$\begin{aligned} \mathcal{L}_{\mathcal{EX},t} = -\sum_{p \in \mathcal{P}} (& y_t^{p,2} \log \hat{ry}_t^{p,2} \\ &+ (1 - y_t^{p,2}) \log (1 - \hat{ry}_t^{p,2})) \end{aligned} \quad (10)$$

ここで $\mathcal{P} = \{ul, ur, ll, lr\}$ である．以上が拡大のスケールが 2 の場合についてであり，スケールを $s(>1)$ に拡張する場合はスケールに合わせて (9) 式における取り出す特徴量の数と領域を変更し，それに合わせた $\mathcal{P}$ を用いて 2 値分類器 R^s を学習させる．

4 実　　験

4.1 実験設定

4.1.1 データセット

実験では顔認識，人型認識，姿勢推定，物体検出，車両検出，文字認識，動物認識，料理認識，天気認識の 9 つのタスクを対象とする．手書き数字のみが収集された MNIST [17] や顔のスケールが大きい画像のみが収集された CelebA [23] などの適切なタスクが明確であるデータセットは本実験では望ましくない．そのため様々な物体や景観が収集された COCO [21], 人物のスケール，ポーズ，オクルージョンのバリエーションが豊富な WIDER [36], ソースが映画であり豊富なシーンが含まれる AVA [10] の 3 つのデータセットを 3.1 節に従ってラベル付けを行い実験で使用する．どのデータセットも公式のスプリットに従って，訓練セット，検証セットを構築する．画像データセットである COCO, WIDER はランダムに選択された訓練セット 10,000 画像，検証セット 2,000 画像にラベル付けを行った．30 分以上の長い動画で構成される AVA はすべての動画で 10 秒ごとに 1 枚画像をサンプルし，そのうえでランダムに選択された訓練セット 10,000 画像，検証セット 2,000 画像にラベル付けを行った．

レコメンドに関しては特に被写体サイズが認識結果に影響を与える顔認識，人型認識，姿勢推定の 3 つのタスクを対象とし，拡大のスケールを 2，縮小のスケールを 0.7 に対して実験を行う．また，ランダムに選択された WIDER の訓練セット 10,000 画像，テストセット 8,000 画像の計 18,000 画像を訓練セットとし，検証セット 2,000 画像，テストセット 2,000 画像の計 4,000 画像を検証セットとして用いる．ただし，前述したとおり，あるタスクにおける Recommender の学習は訓練セットのうちそのタスクにおいて元画像で negative ラベルが付与された画像のみを学習に用いる．

4.1.2 モ デ ル

アノテーション付与のためのモデルとして顔認識，人型認識は NeoFace [1]，姿勢推定は [25]，物体検出は YOLOv5 [9]，車両認識，文字認識は OpenVINO で公開されているモデル，動物認識，料理認識，天気認識は OpenAI が公開している CLIP [26] でそれぞれのタスクのクラスを定義したものを用いた．

提案手法のモデルはバックボーンとして，ImageNet [28] で事前学習済みの EfficientNet [31] を用いる．EfficientNet の stage8 までを特徴抽出器として用い，得られた特徴量から Attention branch により生成したアテンションマップを適用させ，stage9 における畳み込みを task-wise とした．Suitable task classifier はプーリング層と 2 層の全結合層からなる．Recommender は 1 層の畳み込み層と 1 層の全結合層からなる．

4.1.3 学　　習

モデルの入力サイズである 224x224 の画像を作成する処理を以下に述べる．訓練時は画像の長辺が 224 pixel となるようにアスペクト比を保ったままリサイズを行い，解像度 224x224 の画像の中心に配置し，短辺方向の余白を黒で埋める．これは拡大レコメンドを学習するうえで，(9) 式により中間特徴量を切り出した領域と元画像における領域を対応させるためである．また，Suitable task classifier の学習時のみ 50%の確率で画像を水平方向に反転させる．学習率 1e-4, $\beta 1 = 0.9$, $\beta 2 = 0.999$, 重み減衰 0.01 の AdamW により最適化される．エポック数は 8 とし，特に言及しない限り，$\lambda = 1$ とする．

4.1.4 推　　論

推論時は水平方向の反転を除いて訓練時と同様の処理を行う．また，Recommender は Suitable task classifier で適切なタスクではないと判断された画像のみを推論し評価を行う．

表 1: AUC における性能の比較.

model \ task	face	humanoid	pose	object	vehicle	character	animal	food	weather	mean
EfficientNet	0.976	0.899	0.940	0.845	0.758	0.767	0.822	0.681	0.608	0.811
EfficientNet+ours	0.978	0.905	0.944	0.852	0.771	0.778	0.851	0.776	0.680	0.837

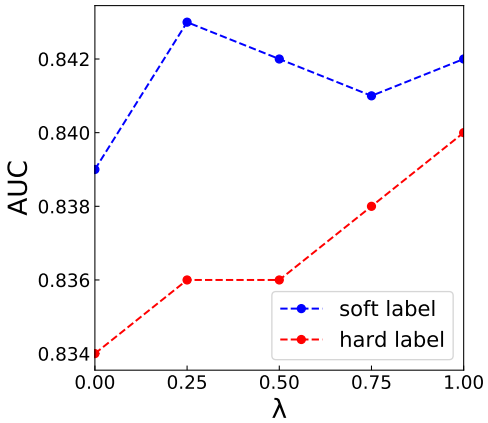

図 4: 学習時にソフトラベルを用いた場合とハードラベルを用いた場合の全タスクの平均 AUC における比較.

4.2 実験結果

4.2.1 Task-appropriate feature extractor の有効性

Task-appropriate feature extractor において，タスクごとのアテンションマップと task-wise convolution を用いなかった場合と提案手法の AUC における比較を表 1 に示す．全てのタスクにおいて提案手法の性能が向上していることが確認できる．特に，料理認識では 0.081，天気認識では 0.087 の向上が見られる．これは，料理認識や天気認識など注目すべき領域が他のタスクとは大きく異なると考えられるタスクに Task-appropriate feature extractor は特に有効であるということが考えられる．

4.2.2 ソフトラベルの有効性

本手法ではあるタスクを学習済みのモデルにおいて出力スコアが閾値を超えた場合に対象のタスクが適切な画像であると判断し positive ラベルを付与している．しかし，閾値よりスコアが高くても間違った認識をしている場合や，閾値よりスコアが低くても正しい認識をしている場合がある．このような画像があるとモデルが何を根拠に予測すべきか学習することが難しいと考えられる．そこで，学習時は positive ラベルと negative ラベルを明確に分ける必要はないと考え，モデルの出力スコアをそのままソフトラベルとして用いることで性能の改善を図る．ソフトラベルを用いた場合と閾値によるハードラベルを用いた場合との比較をいくつかの λ の値において行った実験結果を図 4 に示す．どの λ の値においてもソフトラベルを用いることで全タスクの平均 AUC が向上していることが確認できる．このことから，学習済みのモデルにおける出力スコアの閾値でラベル付けを行った本手法において，学習時は出力スコアをそのままソフトラベルとして用いる手法が有効であると考えられる．

4.2.3 attention branch と task-wise convolution の組み合わせの有効性

提案手法において attention branch と task-wise convolution のどちらも有効であることを検証するために，attention branch と task-wise convolution のいずれか一方のみを用いた実験を行う．結果を表 2 に示す．いずれか一方を用いた場合は，両方用いない場合より性能が向上し，両方組み合わせることでさらに性能が向上していることが確認できる．これは，あくまで attention branch は注視領域に関する情報を付加するだけで，大元の特徴マップが全タスクで共通なので attention branch だけでは効果は小さく，タスクごとに異なる畳み込みも行うことで効果があると考えられる．また，同様にタスクごとに異なる畳み込みを用いるだけでは効果は小さく，注視領域に関する情報と組み合わせることが有効であったと考えられる．

4.2.4 タスクごとに学習させる場合との比較

今回提示するタスクは画像がどんなタスクで適切かを分類するものであるので，タスク固有のモジュールは必要最低限に抑えることが望ましい．そのため，本手法はニューラルネットワークの表現力に期待し，特徴抽出器はすべてのタスクで共有し，タスクごとに必要な領域にアテンションをかけた後にタスク固有の畳み込みを用いる手法である．しかし，動物認識，料理認識，天気認識のように抽出すべき特徴が大きく異なると考えられるものは別々の特徴抽出器を用いたほうが性能が高い可能性がある．そこで，動物認識，料理認識，天気認識の 3 つのタスクを対象とし，タスクごとにそれぞれ別々の特徴抽出器を用い学習させた場合と，単に出力層のみを分けて同時に学習させた場合と，提案手法により同時に学習させた場合との比較を行う．タスクごとに学習させる場合は学習の収束が速いため，タスクごとに学習させる場合はエポックを 5 とし，同時に学習させる場合はエポックを 7 とする．結果を表 3 に示す．単に出力層のみを分けて同時に学習させた場合は，タスクごとに学習させた場合に比べ性能低下が見られるが，提案手法により同時に学習させた場合は大きな性能低下は見られないことが確認できる．この結果から提案手法を用いることでタスクごとに学習させた場合と同等の性能が得られることが分かった．また，タスクごとに特徴抽出器を必要としないためパラメータ数の削減ができ，さらに全てのタスクを同時に学習することが可能となる．

4.2.5 レコメンドの結果

縮小と拡大のそれぞれの Recommender の AUC における性能を表 4，定性的な結果を図 5 に示す．実験では，ソフトラベルを用いて学習させたモデルに Recommender を追加し，Recommender のみを学習させる．人型認識では，

表 2: attention branch (AB) と task-wise convolution (TC) の AUC に与える影響.

AB	TC	face	humanoid	pose	object	vehicle	character	animal	food	weather	mean
		0.976	0.899	0.940	0.845	0.758	0.767	0.822	0.681	0.608	0.811
✓		0.976	0.901	0.940	0.847	0.762	0.772	0.831	0.722	0.634	0.821
	✓	0.977	0.896	0.942	0.847	0.761	0.775	0.833	0.751	0.636	0.824
✓	✓	0.978	0.905	0.944	0.852	0.771	0.778	0.851	0.776	0.680	0.837

表 3: タスクごとに学習させた場合との比較. (a) タスクごとに個々に学習. (b) 出力層のみをタスクごとに分け同時に学習. (c) 提案手法により同時に学習.

	animal	food	weather	mean
(a)	0.864	0.796	0.687	0.783
(b)	0.843	0.742	0.623	0.736
(c)	0.861	0.779	0.664	0.768

表 4: Recommender の AUC における性能.

	face	humanoid	pose
Shrink	0.787	0.806	0.801
Expand	0.908	0.878	0.700

図 5(a)，姿勢推定では図 5(b) のような画像に対して Shrink Recommender が縮小をレコメンドしている．これらの画像は縮小画像では positive ラベルが付与されているが，元画像では被写体が大きく映りすぎたために negative ラベルが付与されたと考えられ，そのような画像に対して，縮小をレコメンドしていることが確認できる．また，顔認識においては，図 5(c),(d) のような画像に対して Expand Recommender が拡大をレコメンドしている．これらの画像は拡大画像では positive ラベルが付与されているが，元画像では被写体サイズが小さいために negative ラベルが付与されたと考えられ，そのような画像に対して，拡大をレコメンドしていることが確認できる．このように元画像のままでは適切ではないタスクにおいても画像の縮小や拡大を行うことで適切なタスクであることをレコメンドすることが可能である．

5 おわりに

本論文では画像がそれぞれのタスクで適切であるかを分類するタスクの提示とそれを解く手法を提案した．提示したタスクを解くことはコンピュータビジョンの普及をさらに広げ，ビジネスチャンスの拡大にもつながると考えられる．実験では，顔認識，天気認識，文字認識など多種多様なタスクを対象とし，我々の提案する手法の有効性を実験的に示した．また，特に認識率に大きな影響を与える被写体サイズに着目し，拡大や縮小をすることで認識可能となるタスクがある場合はそれらのレコメンドも行った．

今後は，今回のラベリング方法ではラベル付けが行えないセグメンテーションタスクや画像ではなく動画として考えるべきである速度推定などのタスクに対するラベリング方法も考え，より優れた画像認識に関するコンサルティングを行う

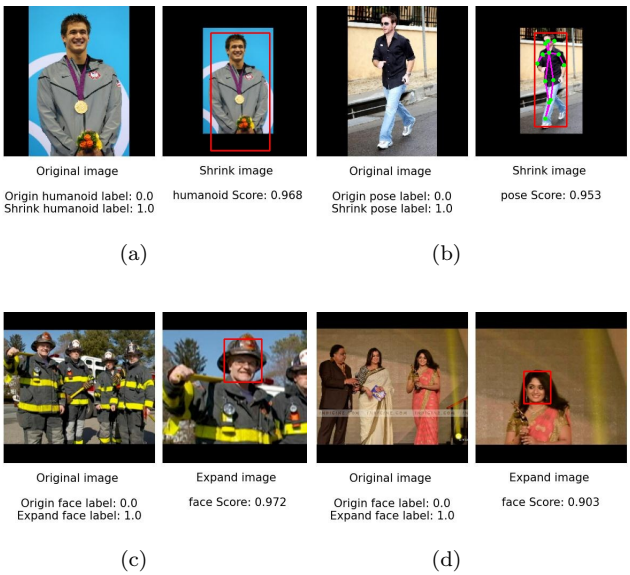

図 5: (a) 人型認識において縮小レコメンドが行われた画像. (b) 姿勢推定において縮小レコメンドが行われた画像. (c),(d) 顔認識において拡大レコメンドが行われた画像.

機械学習モデルを構築したい．また，レコメンドに関して今回行ったのは，ある固定のスケールで拡大や縮小をすべきかどうかの 2 値分類にすぎないが，今後は拡大や縮小のスケールを直接レコメンドする様なモデルと学習方法を考えたい．さらに，拡大や縮小だけでなく，障害物の除去や明るさの調整といったリコメンドの導入も検討していく．

文　献

[1] あらゆるシーンで活躍する顔認証. https://jpn.nec.com/biometrics/face. [Accessed 2023/1/8].

[2] 人と人の距離を ai で可視化 密集度把握の技術開発相次ぐ. https://www3.nhk.or.jp/news/html/20200611/k10012466701000.html, Jun 2020. [Accessed 2022/12/28].

[3] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

[4] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.

[5] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019.

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[7] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] glenn jocher et al. yolov5. `https://github.com/ultralytics/yolov5`, 2021.

[10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[11] Chaoxu Guo, Bin Fan, Qian Zhang, Shiming Xiang, and Chunhong Pan. Augfpn: Improving multi-scale feature learning for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Shizhong Han, Zibo Meng, Zhiyuan Li, James O'Reilly, Jie Cai, Xiaofeng Wang, and Yan Tong. Optimizing filter size in convolutional neural networks for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[15] Jun Huang, Qian Xu, Xiwen Qu, Yaojin Lin, and Xiao Zheng. Improving multi-label learning by correlation embedding. *Applied Sciences*, 11(24), 2021.

[16] Sheng-Jun Huang and Zhi-Hua Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 949–955. AAAI Press, 2012.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Lihuan Li, Maurice Pagnucco, and Yang Song. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2231–2241, June 2022.

[19] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[24] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[25] Yadong Pan, Ryo Kawai, Noboru Yoshida, Hiroo Ikeda, and Shoji Nishimura. Training physical and geometrical mid-points for multi-person pose estimation and human detection under congestion and low resolution. *SN Comput. Sci.*, 1(4):208, 2020.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[27] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, October 2021.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[29] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017.

[30] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[32] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[34] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[36] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[37] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[38] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.