中部大学 工学研究科 ロボット理工学専攻



MACHINE PERCEPTION AND ROBOTICS GROUP

2022年度 藤吉研究室 修士論文発表 アブストラクト

Deep Learning, Attention mechanism, Visual explanation

Complement Objective Training によるアテンションマップの最適化に関する研究 岩吉孝明

Reservoir computing, Echo State Network, DARTS

Echo State Network における構造分析と最適構造探索に関する研究

濵地 優輝

Deep Learning, Semi-supervised Learning

知識転移グラフによる半教師あり深層共同学習に関する研究

村本 佳隆





Complement Objective Training によるアテンションマップの最適化に関する研究

TR21002 岩吉 孝明

指導教授:藤吉弘亘

1.はじめに

Attention Branch Network (ABN)[1] は、ネットワー クの注視領域をアテンションマップとして可視化するだけ でなく、認識処理に活用することで高精度な画像認識を実 現した.一方、鳥の種類のようなサブカテゴリを分類する 詳細画像分類では、クラス間の相違が小さいため、より識 別に有効な領域を注視して分類する必要がある.

本研究では,識別に有効な領域をアテンションマップと して獲得するために,1.認識に有効な領域の探索,2.特徴 空間の改善,3.出力空間の改善の3つのアプローチを提 案する.具体的には,ABN に有効な注視領域を探索する Attention mining branch (AMB),特徴空間を改善する Prototype conformity loss (PC Loss)と出力空間を改善 する Complement objective training (COT)を導入する.

2. Attention Branch Network

ABNは、入力画像から特徴マップを獲得する Feature extractor, アテンションマップを獲得する Attention branch, 最終的な認識結果を出力する Perception branch の3つか ら構成されている. Attention branch で獲得したアテン ションマップは、Attention 機構により Feature extractor で得られた特徴マップに重みづけする.これにより、重要 な領域を強調した特徴が Perception branch に入力する. しかしながら、クラスの異なる物体が複数存在する場合、 ABN のアテンションマップは認識対象以外の物体を注目 することがあり、認識精度を劣化させる可能性がある.こ の問題に対し、人の知見により修正したアテンションマッ プを用いて ABN を再学習する手法 (人の知見) [2] が提案 された. ABN で誤認識した際のアテンションマップを人の 知見を介して修正し,理想的なアテンションマップを作成 して、ネットワークを再学習する.しかし、アテンション マップの修正には人的コストがかかるという問題がある.

3.提案手法

本研究では学習によって詳細画像分類に有効な注視領域 を自動で獲得することを目的とする.提案手法は、ABN に3つのアプローチを導入し、詳細画像分類に有効な領 域の獲得を目指す.1つ目に、ABN に Attention mining branch (AMB)を導入し、認識に有効な領域のみに注視す るよう学習する.2つ目に、ABN に Prototype conformity loss (PC Loss)を導入して特徴空間を改善する.PC Loss は、特徴空間において同じクラスの特徴量を近づけ、異な るクラスの特徴量を離すように学習することから、クラス 特有の注視領域の獲得を可能とする.更に、注視領域の改 善により、認識精度の向上も見込めると考える.3つ目に、 AMB に不正解クラスを平坦化する Complement objective training (COT)を導入し、不正解クラス確率を抑えるこ とで、認識対象のみの注視領域の獲得を目指す.

3.1. Attention mining branch

提案手法は、入力画像から特徴マップを獲得する Feature extractor, アテンションマップを獲得する Attention module, 最終的な認識結果を出力する Perception branch, に加え、識別に有効な領域を探索する AMB で構成されて いる. AMB は、認識に有効な領域のみに注視するように 学習を行う. AMB は, Perception branch と同様の構造 であり、アテンションが閾値より高い領域をマスク処理し た特徴マップを入力し、クラス確率を出力する. このとき、 対象クラスのクラス確率が低いほど、アテンション領域が 認識対象に影響しているといえる. そこで、対象クラスの クラス確率を最小化するよう学習する. これにより、認識 対象のみを注視するようにアテンションマップを最適化す る. また, AMB は, Perception branch と重みを共有する. これにより、認識対象の物体のみを注視するように学習し た AMB の重みを最終的な認識結果を出力する Perception branch に利用することができる.



図 1: 提案手法における学習の流れ

3.2. Prototype conformity loss

特徴空間の改善として, PC Loss を Perception branch と Attention module に導入する. PC Loss は同じクラス 内の特徴量を近づけ,異なるクラス間の特徴量を離す損失 関数である. PC Loss L_{PC} は、サンプル数を N、クラス 数を k、サンプル i の正解クラスを g、不正解クラスを j、 学習可能なクラス重心 w^c と特徴量 f_i , $\alpha_i = ||f_i - w_g^c||_2$, $\beta_i = ||f_i - w_j^c||_2$, $\gamma_i = ||w_g^c - w_j^c||_2$, とすると、式 (1) で 表現できる.

$$L_{PC} = \sum_{i}^{N} \{ \alpha_i - \frac{1}{k-1} \sum_{j \neq g} (\beta_i + \gamma_i) \}$$
(1)

これを最小化することで、 α_i は特徴量を正解クラスの重 心に集め、 $(\beta_i + \gamma_i)$ はクラス間を離すように働く.これに より、クラスごとに特有の特徴が獲得され、その結果、よ り良いアテンションマップを獲得できると考える.

3.3. Complement objective training

AMBの出力空間の改善として、COTを導入する. COT は Complement entropy が最小になるよう重みを更新す ることで、不正解クラス確率を平坦化する. Complement entropy $C(\hat{y}_{\bar{C}})$ は、 \hat{y} を入力 x_i に対するクラス確率、gを 正解クラスのインデックスとすると、式(2)のように表さ れる.

$$C(\hat{y}_{\bar{C}}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j \neq g}^{K} \frac{\hat{y}_{ij}}{1 - \hat{y}_{ig}} \log\left(\frac{\hat{y}_{ij}}{1 - \hat{y}_{ig}}\right) \quad (2)$$

AMB に COT を導入する際, $\hat{y}_{ig} = 0$ とすることで, 正解 クラス確率が 0,残りの不正解クラス確率が平坦になるよ う算出される.これにより,注視領域に悪影響を及ぼす突 出した不正解クラス確率を削減する.

3.4. 学習手順

提案手法における学習の流れを図1に示す.提案手法の 学習手順を以下に述べる.



図 2: アテンションマップの例

Step1 ABN と同様のモデル構造で事前学習を行う.

Step2 ABNの損失に Attention module, Perception branch から算出した PC Loss を加え,最小になるよう学 習する. COT を用いない場合は,AMBの損失も加 える.

Step3 AMB の出力から Complement entropy を算出し, 最小になるよう学習する.

Step2,3をイタレーションごとに繰り返す.

4.評価実験

提案手法の有効性を評価するために詳細画像分類タスク にて評価実験を行う.

4.1.実験条件

本実験では、200 クラスの鳥の画像で構成される Caltech-UCSD Birds 200-2010 (CUB-200-2010) データセットを 用いる. ベースネットワークとして ResNet-50 を使用し, バッチサイズは 16 とする. マスク処理の閾値は 0.78 とし, 学習の更新回数は ABN の事前学習,提案手法それぞれで 300 エポックとする.提案手法との比較として,ABN,人 の知見を用いる.

4.2.実験結果

各手法における分類精度の比較を表1に示す.表1より, AMBは、ABNより分類精度が向上するが、人の知見より 低い. PC Loss もしくは COT を導入すると、人の知見よ りも分類精度が向上することを確認した.また、PC Loss と COT の両者を AMB に組み合わせることで分類精度が 更に向上しており、ABN より Top-1 の分類精度が 14.41 ポイント高い結果となることをことを確認した.

	AMB	PC Loss	COT	Top-1 acc.	Top-5 acc.
ABN	-	-	-	31.68	57.01
人の知見	-	-	-	37.42	62.08
提案手法	\checkmark	-	-	33.33	58.56
	\checkmark	\checkmark	-	45.10	71.68
	\checkmark	-	\checkmark	39.76	66.57
	\checkmark	\checkmark	\checkmark	46.09	69.24

表 1: 認識精度の比較 [%]

4.3.アテンションマップの可視化

各手法におけるアテンションマップの比較を図2に示す. アテンションマップの下に、モデルが認識したクラスとそ のクラス確率を示す.図2に示すように、提案手法により ABN で獲得していた認識対象以外の領域、すなわち不要な 注視領域が軽減されていることが分かる.また、PC Loss を導入することで、ABN より局所的な領域に着目してい ることが分かる.一方で、AMB、COT は認識対象を広域 に捉えつつ、不要な注視領域を軽減していることが分かる. また, COT は PC Loss より高いクラス確率を獲得してい ることから, PC Loss では捉えられていない重要な領域を 獲得できていると考えられる.

4.4.アテンションマップの定量的評価

各手法における Insertion, Deletion 結果を表 2 に示す. 提案手法により, ABN や人の知見と比ベスコアが向上し ており, 認識に有効な領域を捉えられるといえる.また, PC Loss と COT を組み合わせることでスコアが更に高く なった. これらのことから, AMB に PC Loss と COT を 組み合わせることで認識に有効な領域を獲得し, それに伴 い認識精度が向上したことが分かった.

表	2:	Insertion.	Deletion	結果
~	<u> </u>	THOUT OTOTIN	DOIOUIOII	111/18

	AMB	PC Loss	COT	$\operatorname{Insertion}\uparrow$	$Deletion \downarrow$	$\mathrm{Score}\uparrow$
ABN	-	-	-	0.2095	0.0881	0.1214
人の知見	-	-	-	0.1685	0.1576	0.0109
提案手法	~	-	-	0.2184	0.0807	0.1377
	~	~	-	0.2908	0.1251	0.1657
	\checkmark	-	~	0.2838	0.1241	0.1597
	~	~	~	0.3329	0.1263	0.2066

5.おわりに

本研究では、ABN に注視領域が認識に有効か考慮しな がら学習する AMB,特徴空間を改善する PC Loss,出力 空間を改善する COT の 3 つのアプローチにより ABN の 注視領域を改善する手法を提案した.評価実験では、提案 手法を用いることにより、不要な注視領域を軽減し、認識 精度が向上したことを確認した.今後は、学習方法やモデ ル構造の検討による更なる注視領域の改善を行う.

参考文献

- H. Fukui, et al., "Attention branch network: Learning of attention mecha-nism for visual explanation", CVPR, 2019.
- [2] M. Mitsuhara, et al., "Embedding human knowledge in deep neural network via attention map". VISAPP, 2021.

研究業績

- 岩吉孝明,等,"Attention mining branch を導入した ABN によるアテンションマップの最適化",第 24 回画 像の認識・理解シンポジウム, 2021.
- [2] T. Iwayoshi, et al., "Attention mining branch for Optimizing Attention Map", MVA, 2021.
- [3] T. Iwayoshi, et al., "Complement Objective Mining Branch for Optimizing Attention Map" VISAPP, 2023.

→:接続評価重み
:非接続評価重み

1.はじめに

人間の脳には、役割により機能する箇所が分かれていく 機能分化と呼ばれる現象がある.一方,人工ニューラルネッ トワークの1つである, Echo State Network (ESN) [1] に おいても、特定の時間または空間パターンに反応するニュー ロンを獲得する機能分化が確認されている [2]. 本手法は, Genetic Algorithm (GA) を用いて ESN の Reservoir 層 のニューロンの接続構造と重みを探索する.しかし、GA による最適な重みの探索には膨大な時間が必要になる. そ こで本研究では、DARTS を用いて ESN の接続構造探索 を行い、探索した接続の重みを誤差逆伝播により最適化す る方法を提案する.

2. 先行研究

ESN は入力層, Reservoir 層, 出力層の3層で構成され, 出力層のみを学習する.入力層と Reservoir 層の重みは 0 から1の間でランダムに決定する. Yamagutiら [2]は、図 1のように Reservoir 層のニューロンを Input Neurons と Output Neurons に分けた拘束条件付き ESN を提案した. 更に,この拘束条件付き ESN を対象に Genetic Algorithm (GA)を用いて構造探索を行う手法を進化 ESN として提案 した. 拘束条件付き ESN の計算を式 (1) と式 (2) に示す.



図 1. 拘束条件付き ESN のネットワーク構造

$$h_T = f(W_{in}X_T + W_{res}h_{T-1}) \tag{1}$$

$$y_T = f(W_{out}h_T) \tag{2}$$

 W_{in} は入力重み, X_T は入力であり, h_T は時刻 Tにおける 出力である. Wres は Reservoir 重みであり, 前時刻の出力 である h_{T-1} との積により過去の情報を考慮した Reservoir 層の出力を求める. この Reservoir 層の出力と Wout の積 に活性化関数 f を適用し, ESN の最終出力を求める. 拘 束条件付き ESN を含めて一般的な ESN は出力層のみを 学習対象としているため Wout のみが最適化される.進化 ESN はこの拘束条件付き ESN の出力層の学習後,遺伝的 アルゴリズム (Genetic Algorithm) を用いて Reservoir 層 の重みと接続構造を探索する.このとき、全ニューロンの 構造の組み合わせは膨大なため,進化 ESN によって最適 な重みと接続構造の両者を十分に探索することは困難であ るという問題がある.

3.提案手法

本研究では、DARTS を用いて ESN の接続構造の探索を 行い,探索した接続に対する重みの最適化を行う DARTS-ESN を提案する.提案手法では, DARTS [3] を用いて ニューロン間に評価重みを定義し,接続構造を探索する. その後,探索した接続に対する重みの値を誤差逆伝播によ り最適化する.この接続構造探索と重み最適化の構造探索 手法を DARTS-ESN と定義する. DARTS-ESN の構造探 索の流れを図2に示す.

3.1. DARTS による接続構造探索

ESN の接続構造の探索に DARTS を適用する. この接 続構造探索時の ESN モデルを図 3 に示す.DARTS を適 用した ESN は Reservoir 層の各ニューロンが接続してい る場合の接続評価重みと、接続していない場合の非接続評 価重みの2つを持っており、それらの出力を Softmax 関数 によってスケーリングする. $\alpha_0^{(i,j)}$ はニューロン $i \ge j$ の



図 2. DARTS-ESN の構造探索の流れ

0-0

接続重み

非接続評価重み、 $\alpha_1^{(i,j)}$ を接続評価重みを表す.式 (3) に DARTS を適用した Reservoir 層の出力計算を示す.

$$\bar{o}^{(i,j)}(X_T^i) = \sum_{o \in O} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in O} \exp(\alpha_o^{(i,j)})} o(X_T^i)$$
(3)

ここで、Oは接続を表す2つの候補(非接続を0,接続を 1) の集合である.ニューロン間の接続 (i, j) において考え



図 3. DARTS による接続構造最適化モデル

られる全ての演算候補の集合である評価重み $lpha_o^{(i,j)}$ を誤差 逆伝播により評価重みを更新する. これにより, 精度に良 い影響を及ぼす,接続状態 $\alpha_o^{(i,j)}$ の評価重みが増大する. 誤差が収束するまで重みを更新した後,接続構造毎により 大きい評価重みを持つ構造を選択することで、最適な接続 を決定する.

3.2. 誤差逆伝播による重みの最適化

DARTS により接続構造を探索した後、全ての層を誤差 逆伝播を行い重みを最適化する. 拘束条件付き ESN は通 常、入力層と Reservoir 層の学習を想定していないため、 Reservoir 層において拘束条件と接続構造を維持するように 計算式を変更する.重み最適化する際の拘束条件付き ESN の計算式を式 (4), 式 (5), 式 (6), 式 (7) に示す.

$$\widehat{h_T^1} = W_{in} x_T + W_{res}^1 b_1 h_{T-1}^1 + W_{res}^{21} b_{21} h_{T-1}^2 \qquad (4)$$

$$h_T^2 = f(W_{res}^{12}b_{12}\hat{h}_T^1 + W_{res}^2b_2h_{T-1}^2)$$
(5)

$$h_T^1 = f(\widehat{h_T^1}) \tag{6}$$

$$y_T = f(W_{out}h_T^2) \tag{7}$$

式 (4) で Input Neurons の内部状態を式 (5) で Output Neurons の内部状態をそれぞれ計算することで、拘束条 件と接続構造を維持したまま重みを最適化できる.式(4), 式 (5) において, Input Neurons の内部状態は $\widehat{h_{T}^{1}}$, Output Neurons の内部状態は h_T^2 とする. Reservoir 層重み は, Input Neurons と Output Neurons の関係を考慮して $W_{res}^{12}, W_{res}^{2}, W_{res}^{12}$ に分割する.また、学習対象と $W_{r\epsilon}^1$ せず, 誤差逆伝播後も拘束条件と接続構造を維持する b1, b2, b12, b21 を定義する. また,活性化関数は式 (2) の hT と式 (7) の h_T² の計算結果が同一になるよう式 (6) で計算 する. 最後に,式(7)において,Reservoir 層出力 h_T^2 と出 力重み Wout の積によって計算される出力 yr と正解ラベ ルとの誤差を算出し誤差逆伝播によって重みを最適化する. 4.評価実験

本研究では,進化 ESN と DARTS-ESN の構造探索手 法の評価を行う.

4.1.実験概要

学習データは空間情報と時間情報を組み合わせたパター ンを用いる.このとき,縦方向の分割数を空間情報,横方 向の周波数を時間情報と見立てることでマルチタスクを実 現する.また,クラス分類の数を変化させ,6クラス分類, 8クラス分類,10クラス分類と分ける.これにより,タス ク難易度毎の初期化条件,ニューロン数と接続率の関係性 を分析する.実験に使用したコサイン波の波形パターンの 一例を図4に示す.



図 4. 実験用データの波形パターン

4.2.ESN の構造探索

進化 ESN と DARTS-ESN の構造探索を行う.初期化手 法には,He の初期化を用いる.拘束条件付き ESN を対象 として進化 ESN は Reservoir 層と出力層の構造探索を行 い,DARTS-ESN は全ての層の重みと Reservoir 層の接続 構造の構造探索を行う.Reservoir 層のニューロン数は 32, 64,128 とする.進化 ESN では世代数 10,個体数 100,生 存個体数 20, Epoch 数 1 で GA を行い,DARTS-ESN は 接続構造探索 100 Epoch,重み最適化 ESN モデルでの重 み最適化 100 Epoch とする.各手法の各クラス分類数に おいての構造探索の結果から,最も良い性能を獲得したと 考えられるモデルを選択した.それらモデルの空間情報精 度,時間情報精度とニューロン数を比較する.各クラス分 類数毎の進化 ESN の分類精度とDARTS-ESN の分類精 度を表 1 に示す.

		進化 ESN	DARTS-ESN
		(ニューロン数)	(ニューロン数)
で カニョハ海	空間情報	50.21(32)	81.95 (128)
0 ジ ノ ヘ 力 頬 	時間情報	45.25 (32)	65.84(128)
0 カニッム桁	空間情報	47.46 (64)	84.62 (64)
0 ジ ノ ヘ <i>力</i> 頬	時間情報	34.89 (64)	61.25(64)
10 クラフ分類	空間情報	51.85(64)	85.52 (128)
10 9 7 入力規	時間情報	20.89(64)	54.81(128)

表 1. 分類精度 [%]

表1より,進化 ESN と比較して DARTS-ESN の方が 全てのクラス分類数において空間情報,時間情報の両方の 分類精度が高い結果となった. DARTS-ESN は進化 ESN より高精度なモデルを獲得する構造探索ができた.また, DARTS-ESN は8クラス分類において,64ニューロン,6, 10クラス分類においては128ニューロンが高い精度を示 している.一方,進化 ESN はニューロン数が少ない,32 もしくは64の時に高い精度を示した.これは進化 ESN の構造探索はGA で行うため,ニューロン数128 におい て十分な構造探索できなかったためであると考えられる. DARTS-ESN は接続構造の探索と重み最適化を分けて行 うため,ニューロン数が多くても最適な構造探索が可能で ある.

4.3.ESN の構造分析

入力情報の空間情報と時間情報の正解ラベルと各ニュー ロンの出力の相互情報量の算出を行った.この算出により ニューロンが情報を特化して処理するようになる機能分化 を分析する.横軸に空間情報量,縦軸に時間情報量を示す. 図 5(a) に 8 クラス分類 64 ニューロンの時の進化 ESN の 機能分化の分析を,図 5(b) に 8 クラス分類 64 ニューロン の時の DARTS-ESN の機能分化の分析を示す.



図 5.8 クラス分類における機能分化の結果

図 5(a) から,進化 ESN は空間情報と時間情報の両者に おいて各情報に特化したニューロンが発生する機能分化が 生じている.図 5(b) から,DARTS-ESN は空間情報のみ に特化したニューロンが多く発生する機能分化であること が分かる.その他の実験条件における機能分化の分析から も進化 ESN と DARTS-ESN では,機能分化は異なる傾向 である事が分かった.具体的には,進化 ESN は両方の情 報に対してそれぞれ特化したニューロンが発生する機能分 化を誘発し,DARTS-ESN では空間情報のみに特化した ニューロンが発生する機能分化を誘発することが分かた.

また,精度と機能分化の傾向についても分析する.表1 の8クラス分類の進化 ESN と DARTS-ESN の空間情報 の結果と図5(a),図5(b)から,DARTS-ESN は進化 ESN と比べて,空間情報量へのニューロンの機能分化が多く, 空間情報精度も高いことが分かる.一方で,表1の8クラ ス分類の進化 ESN と DARTS-ESN の時間情報の結果と 図5(a),図5(b)から,DARTS-ESN は進化 ESN と比べ て,時間情報量へのニューロンの機能分化が少ないにもか かわらず時間情報精度が高いことが分かる.この結果から, DARTS-ESN の機能分化は進化 ESN とは違う傾向の機能 分化を示す構造を獲得しているものの,問題設定に対して 有効な構造を獲得していると言える.

5.おわりに

本研究では、DARTS を用いて ESN の接続構造探索を 行い、誤差逆伝播を用いて重み最適化を行う手法を提案し た.結果、従来の手法である進化 ESN と比較して、高精 度な ESN の構造を獲得し、進化 ESN とは異なる機能分化 の特徴を持つことを確認した.今後の予定としては、獲得 できた ESN の構造のより詳しい分析やニューロンカット による分類精度の確認などが挙げられる.

参考文献

- H. Jaeger, et al., "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", Science, 2004.
- [2] Y. Yamaguti, et al., "Functional differentiations in evolutionary reservoir computing networks", Chaos 2021.
- [3] Liu, *et al.*, "Darts: Differentiable architecture search", arXiv preprint arXiv:1806.09055, 2018.

研究業績

濱地優輝, 平川翼, 山下隆義, 藤吉弘亘, 山口裕, 津田一 郎, "Hessian-Free 法を用いた Echo State Network の構 造探索", FIT2021(第 20 回情報科学技術フォーラム), 2021.

指導教授:藤吉弘亘

1.はじめに

半教師あり学習は、ラベルありデータとラベルなしデー タを用いた学習法である.半教師あり学習の分野で手法を 大別すると、一致性正則化法、擬似ラベリング法、それら を組み合わせた手法がある.これらの学習手法は、人が設 計した構成であり、最適な学習法であるとは限らない.本 研究では、従来法を内包しつつ、新しく高精度な半教師あ り学習法の自動設計を目的とする.従来の相互学習を内包 する知識転移グラフに半教師あり学習法の主要な構成要素 を導入し、グラフ構造を自動設計することで、従来法より 高精度な半教師あり学習法を獲得する.

2.半教師あり学習

半教師あり学習において一致性正則化法は, 摂動とラベル なしデータを活用した正則化効果によりモデル性能を改善 する. 一致性正則化法の代表例としてΠ-model [1] がある. Π-model の派生手法として, 代表的な Mean Teacher [2] の他に, 擬似ラベリング法を導入した FixMatch [3] があ る. FixMatch は, ラベルなしデータに対して弱変換時に 推測値をもとに生成した擬似ラベルと強変換時の推測値が 一致するように学習する. そのため FixMatch は, 従来法 で個別に行うような, 一致性正則化と擬似ラベリングを上 手く合わせた正則化により, 高い精度を発揮する.

3.提案手法

FixMatch は一致性正則化と擬似ラベリングの組み合わ せが効果的であることを示した.本研究では,半教師あり 学習の従来法を内包しつつグラフで統一の形式に表現し, グラフの構成要素を組み合わせて新たな半教師あり学習法 を自動設計する手法を提案する.半教師あり深層学習の従 来法として,一致性正則化法,擬似ラベリング法,複合的な 学習法をノードとエッジを用いてグラフで表現する.ノー ドはネットワークモデルを表し,エッジは誤差計算と誤差 逆伝播による知識転移の方向を表す.

3.1. 一致性正則化法のグラフ表現

一致性正則化の代表的な学習法として Π -model と Mean Teacher の学習の流れをグラフに表現する. Π -model の グラフ表現を図 1(a) に示す.





(b) Mean Teacher

図1:一致性正則化法のグラフ表現

 $f(\theta, x)$ はパラメータ θ を持つネットワークに画像 x を 入力した時のクラス確率を表す. 1 枚の画像に対して異な る摂動を付与した 2 枚の画像 $x + \zeta_1, x + \zeta_1'$ をネットワー クに入力し、2つのクラス確率の差を誤差とする一致性損 失を最小化するように学習する. 一致性損失 \mathcal{L}_{con} を式 (1) に示す.

 $\mathcal{L}_{con}(x) = \mathbb{E}_x \mathcal{R}(f(\theta_1, x + \zeta_1), f(\theta_1, x + \zeta_1')). \quad (1)$

ここで、 $\mathcal{R}(\cdot, \cdot)$ は2つの確率分布間の KL-divergence による誤差を表す. Π -model は、誤差の算出に用いるネット

ワークと更新するネットワークが同一である. そのため, パラメータ θ_1 をもつネットワークモデルをノード m_1 と して, m_1 から m_1 への知識転移としてグラフに表現でき る. 一致性損失は始点と終点が同じノードにあるエッジで 表現する.

Mean Teacher のグラフ表現を図 1(b) に示す. Mean Teacher は、学習を補助する役割を持つ指数移動平均 (EMA) モデルを使用し、EMA モデルのクラス確率 $f(\text{EMA}(\theta_1), x + \zeta_1)$ に $f(\theta_1, x + \zeta_1')$ が類似するように学習する. EMA モデルのパラメータ EMA(θ_1) の更新方法を式 (2) に示す.

$$EMA(\theta_{1,t}) = \alpha EMA(\theta_{1,t-1}) + (1-\alpha)\theta_{1,t} \qquad (2)$$

ここで、 α はハイパーパラメータ、tは学習ステップである. Mean Teacher は、指数移動平均モデルのパラメータ EMA(θ_1)は θ_1 の EMA で更新する. そのため、EMA モ デルを表すノードを用いて、EMA モデルと m_1 による双 方向の知識転移としてグラフに表現する.

3.2.擬似ラベリング法のグラフ表現

擬似ラベリング法の代表的な手法として Pseudo-Label [4] のグラフ表現を図2 に示す. Pseudo-Label は, $f(\theta_1, x+\zeta_1)$



図 2: Pseudo-Label のグラフ表現

の推測クラスを1とする one-hot ベクトルを生成し,生成 した one-hot ベクトルを入力画像の擬似的なラベル y' と して, y' と $f(\theta_{1}, x + \zeta'_{1})$ が類似するように学習する.損 失関数 \mathcal{L}_{pse} を式 (3) に示す.

$$\mathcal{L}_{pse}(x) = -\mathbb{E}_x y' \log f(\theta_1, x + \zeta_1') \tag{3}$$

Pseudo-Label の損失関数を PseudoLoss としてエッジ に表現し, m₁ から m₁ への知識転移としてグラフに表現 する.また,疑似ラベルを生成するネットワークと推測を 行うネットワークが異なる場合,始点と終点のノードが異 なるエッジを用いて容易に拡張できる.

3.3. 複合的な学習法のグラフ表現

複合的な学習法として FixMatch のグラフ表現を図3に 示す.擬似ラベルは、ネットワーク性能が不十分な学習序



図 3: FixMatch のグラフ表現

盤において誤識別により学習を阻害する可能性がある. そのため FixMatch は,擬似ラベルとクラス確率が一致するように学習する際,クラス確率が閾値 τ を越える場合のみ擬似ラベルを生成することで損失値を制御する. グラフでは入力する損失値を a として,擬似ラベリングに対する損失値の制御をゲート関数 $G^{Threshold}(a)$ で表し,計算方法を式 (4) に示す.

$$G^{Threshold}(a) = \mathbb{1}(\max f(x) > \tau) \cdot a \tag{4}$$

FixMatchは、PseudoLoss で得た損失を Threshold Gate で制御するエッジを用いてグラフに表現する.また、摂動 ζ_1' を RandAugment などの強い摂動とすることで、さら に正則化効果を高めることができる.

3.4. グラフ最適化による半教師あり学習法の自動設計

従来の学習法を内包しつつ,図4のようにノードとエッ ジを組み合わせて学習法を自動設計する.知識転移を制御 するために複数種のゲート関数を導入することで,多様な 共同学習を実現することができる.ゲート関数は,誤差値 を制御して対象ノードを効率的に学習させる.



図4:各エッジの処理の流れと設計空間の例

グラフ構造の自動設計には、知識転移グラフ [5] の最適 化法を用いる.最適化したい評価対象ノードを除くノード を補助ノードとして、評価対象ノードの精度が最大化する ように補助ノードとエッジをハイパーパラメータサーチに より設計する.グラフ要素の組み合わせ数は膨大になるこ とが多く、並列かつ非同期的に実行可能な最適化法 ASHA を用いる.ASHA は、ランダムサーチによって提案された パラメータを用いて試行し、学習途中の評価値が過去に試 行したパラメータと比較した時に既定の割合を下回る順位 であれば枝刈りを行う.

4.評価実験

自動設計した新しい半教師あり学習法を評価する.デー タセットには CIFAR-100 を用い,ラベルありデータ数を 2,000~10,000,試行回数を 15,000,ノード数は 2 と 3,評 価対象ノードは ResNet32 とする.設計空間は,モデル が ResNet18, WideResNet28-2 と 28-6,EMA モデルの 4 種,損失計算が KL-divergence,PseudoLoss の 2 種,ゲー ト関数が Through, Cutoff,Threshold, Positive/Negative linear の 5 種である.Through はノード間の接続,Cutoff は切断,Positive linear と Negative linear は動的な接続を 表す.Positive linear は学習が進むにつれて重みが大きく なり学習序盤は Cutoff と同様に切断する役割をもつ.そ れに対して,Negative linear は学習終盤に切断する役割を もつ.

4.1.評価結果

ノード数とラベルありデータ数のそれぞれ学習条件に対して個別に学習法を自動設計し,自動設計した学習法と従 来法の認識精度の比較を表1に示す.

衣 : 目動設計した子習法と促米法の有度比較

	Number of labeled data				
	2,000	4,000	6,000	8,000	10,000
Supervised	20.85	32.35	41.56	50.38	53.61
Pseudo-Label	29.01	40.59	47.79	54.51	56.67
Π -model	29.54	41.98	50.69	53.44	55.98
Mean Teacher	31.28	43.10	48.85	49.98	54.52
FixMatch	29.31	41.87	47.33	51.28	54.03
Ours(2 nodes)	42.04	51.92	55.94	60.65	63.08
Ours(3 nodes)	42.55	54.60	57.12	62.49	63.81

Supervised の行はラベルありデータのみで ResNet32 で教師あり学習した場合の認識精度である.表1より,ラ ベルありデータ数がいずれの場合においても自動設計した 学習法が従来法に比べて高精度であり,ノード数が2か ら3になると更に精度が向上する.これは,自動設計した 学習法は新しく高精度な学習構造を持ち,ノード数が増え て更に多様なグラフ構造から自動設計することの有効性を 表す.また,従来法に注目すると,ラベルありデータ数が 2,000~6,000 と少ない場合は一致性正則化法の II-model や Mean Teacher の精度が高く,ラベルありデータ数が8,000, 10,000 と多い場合は擬似ラベリング法の Pseudo-Label の 精度が高い.そのため,ラベルありデータ数によって適切 な学習法が異なることがわかる.



(c) ゲート関数 (ラベル数 10k) (d) 学習方法 (ラベル数 10k)
 図 5: ノード数 2 における自動設計した学習法のグラフ構造



(c) ゲート関数(ラベル数 10k) (d) 学習方法(ラベル数 10k)
 図 6: ノード数3における自動設計した学習法のグラフ構造

4.2.自動設計した半教師あり学習のグラフ構造

自動設計したグラフの代表例として、ラベルありデータ 数が少ない6,000と多い10,000の場合に、CIFAR-100を 用いて自動設計した学習法のグラフ構造を図5,6に示す. エッジがゲート関数と損失関数という2つの要素を表すこ とから、ゲート関数と学習方法の2つの観点で表記する. 赤のノードは評価対象、青と灰色のノードは補助ノードで あり、青は教師あり学習による学習済みモデルを表す.

学習経過に注目すると、図5よりノード数2の場合、ラ ベルありデータ数が少ない場合は学習序盤に教師あり学習、 多い場合は学習終盤にラベル情報を持つ学習済みモデルか ら擬似ラベリングをすることから、半教師あり学習におい ては学習序盤と終盤で異なる学習法が効果的である.また、 図6よりノード数3の場合においても、ラベルありデータ 数が少ない場合は Negative linear Gate により学習序盤の ラベル情報活用が強調され、ラベルありデータ数が多い場 合は Positive linear Gate と擬似ラベリングにより学習終 盤のラベル情報活用が強調される.したがって、ラベルあ りデータ数が少ない場合は学習序盤、多い場合は終盤にラ ベル情報を活用する学習法が効果的であることがわかる.

5.おわりに

本論文では、従来の半教師あり学習法をグラフで統一的 に表現し、自動設計により新しく高精度な半教師あり学習 法を獲得した.今後は、ベイズ最適化の導入、自己教師あ り学習への拡張について検討する.

参考文献

- S. Laine, et al., "Temporal Ensembling for Semi-Supervised Learning", ICLR, 2017.
- [2] A. Tarvainen, et al., "Mean teachers are better role models: Weight-averaged consistency targets improve semisupervised deep learning results", NeurIPS, 2017.
- [3] K. Sohn, et al., "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence", NeurIPS, 2020.
- [4] D. Lee, et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks", ICML, 2013.
- [5] S. Minami, et al., "Knowledge Transfer Graph For Deep Collaborative Learning", ACCV, 2020.

研究業績

 村本 等, "Refined Consistency による知識蒸留を用いた半 教師あり学習",人工知能学会全国大会,2021. (他1件)