Refined Consistencyによる知識蒸留を用いた半教師あり学習

Refined Consistency for Semi-Supervised Learning with Knowledge Distillation

村本 佳隆 Yoshitaka Muramoto
岡本 直樹 Naoki Okamoto
平川 翼 Tubasa Hirakawa
山下 隆義 Takayoshi Yamashita
藤吉 弘亘 Hironobu Fujiyoshi

中部大学
Chubu University

Semi-supervised learning is a method that uses both labeled and unlabeled data for training the model. Dual Student (DS), which transfers knowledge between two networks, and Multiple Student (MS), which expands the number of DS networks to four or more, have been proposed as semi-supervised learning. MS achieves higher accuracy than DS, but learning MS is inefficient because knowledge transfer between all networks is not performed at once in the MS learning. In this paper, we propose refined-consistency, which transfers knowledge between all networks at once, to improve accuracy through an efficient knowledge transfer method. In the experiment with the CIFAR-100 dataset, we show that the proposed method improves the accuracy more than MS.

1. はじめに

半教師あり学習は，ラベルありデータに加えてラベルなしデータを活用する学習法である．そのため，アノテーションなどにおけるコスト削減や大量の学習データを容易に確保できるなどの利点がある．半教師あり学習手法として，2 つのネットワーク間で知識を転移する Dual Student [Zhanghan 19]，Dual Student のネットワーク数を 4 つ以上に拡張した Multiple Student [Zhanghan 19] が提案されている．Dual Student は，各ネットワークが一貫性損失に基づいた半教師あり学習を行い，ネットワーク間で知識転移を行う．Multiple Student は，1 度にランダムな 2 つのネットワークを用いて，Dual Student と同様の方法で学習する．Multiple Student は Dual Student に比べて高い精度であるが，全てのネットワーク間での知識転移を一度に行うことができないため，学習効率が悪い．

そこで本研究では，半教師あり学習における効率的な知識転移方法による精度向上を目的として，1 度にすべてのネットワーク間で知識を転移する Refined Consistency を提案する．Refined Consistency は，一貫性損失を知識の指標として，最も一貫性損失が小さいネットワークから他のすべてのネットワークに知識を転移する．また，一貫性に基づいた半教師あり学習手法では，RandAugment [Cubuk 19] などの大きな摂動を付与するデータ拡張手法を用いて高い精度を発揮する Unsupervised Data Augmentation [Xie 20] と FixMatch [Sohn 20] が提案されている．本稿では，Refined Consistency にデータ拡張として RandAugment を適用し，RandAugment の有無による精度変化を評価する．

2. 関連研究

2.1 一貫性に基づいた半教師あり学習法

一貫性に基づいた半教師あり学習法として，様々な手法が提案されている．Laine らは，それぞれ異なる摂動を付与した 2 つの入力データに対して，Dropout の影響下でネットワークが一貫した推測をするように学習する Π-model [Laine 17]，単一のネットワークによる時間方向のアンサンブルを用いて Π-model の計算コストと精度の双方を改善した Temporal Ensembling [Laine 17] を提案した．Tarvainen らは，教師ネットワークの重みを生徒ネットワークの重みの指数移動平均値とし，中間表現をもつ教師ネットワークにより目標値を生成する Mean Teacher [Tarvainen 17] を提案した．Radosavovic らは，ラベルありデータで学習済みのネットワークを用いて，複数の画像変換それぞれに対する事後確率の平均値をラベルなしデータのラベルとして付与する Data Distillation [Radosavovic 18] を提案した．Verma らは，ラベルなしデータに Mixup [Zhang 18] を用いることでデータ間の補間を促すように学習する Interpolation Consistency Training [Verma 19] を提案した．Suzuki らは，2 つの入力データのうち一方に敵対的な変換を適用する Adversarial Transformations [Suzuki 20] を提案した．Xie らは，2 つの入力データのうち一方に RandAugment などの強変換を適用する Unsupervised Data Augmentation [Xie 20] を提案した．Sohn らは，2 つの入力データのうち一方に RandAugment などの強変換，もう一方には左右反転や平行移動などの弱変換を適用し，弱変換を適用した際の事後確率を用いてラベルなしデータに擬似ラベルを付与する FixMatch [Sohn 20] を提案した．Zhanghan らは，それぞれ 2 つのネットワークが一貫性に基づいた学習を行い，ネットワーク間で安定性という指標に基づき知識転移を行う Dual Student [Zhanghan 19]，Dual Student のネットワーク数を 4 つ以上に拡張する Multiple Student [Zhanghan 19] を提案した．

2.2 Dual Student

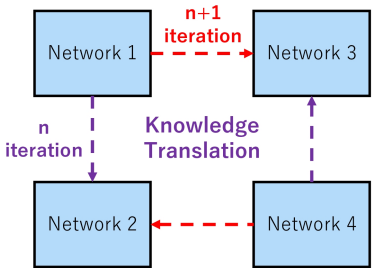
Dual Student は，2 つのネットワークを用いた半教師あり学習手法である．各ネットワークには，同じ入力に対してランダム性によって異なる摂動を加えた 2 つの入力を与える．そして，摂動に対して頑健になるよう学習する．さらに，安定性という指標を用いてネットワーク出力の一貫性を評価し，安定性が高いネットワークから優れた知識を転移させている．知識は多くのネットワークから転移させる方が高い精度を達成できる．Multiple Student は Dual Student を拡張し，4 つ以上のネットワークを用いている．Multiple Student は，1 度にランダムな 2 つのネットワークを用いて，Dual Student と同様の

連絡先:
村本 佳隆：yoshitaka@mprg.cs.chubu.ac.jp
岡本 直樹：naok@mprg.cs.chubu.ac.jp
平川 翼：hirakawa@mprg.cs.chubu.ac.jp
山下 隆義：takayoshi@isc.chubu.ac.jp
藤吉 弘亘：fujiyoshi@isc.chubu.ac.jp

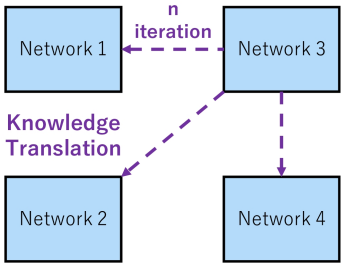
方法で学習する．

3. 提案手法

Multiple Student は，Dual Student のネットワーク数を 4 つ以上に拡張することで高い精度を発揮する．しかし，1 度にランダムな 2 つのネットワークを用いて学習するため，1 度に全てのネットワーク間での知識転移を行うことができず学習効率が悪い．そこで本研究では，効率的な知識転移方法による精度向上を目的として，1 度に全てのネットワーク間で知識を転移する Refined Consistency を提案する．Multiple Student と提案手法について，1 イタレーションごとの知識転移の方向を図 1 に示す．



(a) Multiple Student



(b) Refined Consistency

図 1: 知識転移方向の例 (ネットワーク数 4)

3.1 損失関数

提案手法の学習方法を図 2 に示す．入力 x に対するネットワーク i の損失関数 $\mathcal{L}^i(x)$ を式 (1) に示す．

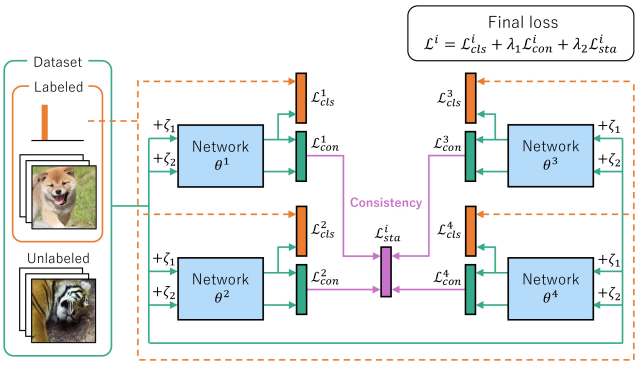


図 2: Refined Consistency の学習方法 (ネットワーク数 4)

$$\mathcal{L}^i(x) = \mathcal{L}^i_{cls} + \lambda_1\mathcal{L}^i_{con} + \lambda_2\mathcal{L}^i_{sta} \tag{1}$$

ここで，λ_1, λ_2 は重み係数である．損失関数は，3 つの損失式 $\mathcal{L}_{cls}, \mathcal{L}_{con}, \mathcal{L}_{sta}$ の重み付き和で表される．$\mathcal{L}_{cls}$ は，推測値とラベル間の交差エントロピーであり，x がラベルありデータである場合のみ計算する．$\mathcal{L}_{con}$ は一貫性損失，$\mathcal{L}_{sta}$ は安定化損失である．安定化損失は，各ネットワークの一貫性損失を用いて計算する．

3.2 一貫性損失

各ネットワーク i に対する一貫性損失 $\mathcal{L}_{con}$ を式 (2) に示す．

$$\mathcal{L}^i_{con}(x) = \mathbb{E}_{x\in\mathcal{D}}||f(\theta^i, x+\zeta_1) - f(\theta^i, x+\zeta_2)||^2 \tag{2}$$

ここで，$\mathcal{D}$ はデータセット，ζ_1, ζ_2 はランダム性のある摂動，$f(\theta, x)$ は重み θ をもつネットワークの x に対する事後確率である．

一貫性損失は，画像変換等により異なる摂動 ζ_1, ζ_2 をそれぞれ入力に付与して，同一のネットワークに入力した時の事後確率間の誤差により求める．一貫性損失の値が小さいほど，ネットワークが摂動に対して頑健である事を表す．ネットワークが摂動に対して頑健になるように学習させることで，類似した入力データに対してより抽象的な不変性を獲得する．一貫性に基づいて摂動を付与することで正則化された分類モデルは，ラベルありのデータ点とラベルなしのデータ点の距離を最小化し，ラベルありのデータ点から識別境界を遠ざけることが確認されている [Tarvainen 17].

3.3 安定化損失

各ネットワーク i に対する安定化損失 $\mathcal{L}_{sta}$ を式 (3) に示す．

$$\mathcal{L}^i_{sta}(x) = \begin{cases} 0 & (\boldsymbol{A} = \varnothing) \\ ||f(\theta^i, x) - f(\theta^s, x)||^2 & (\text{otherwise}) \end{cases} \tag{3}$$

ここで，$\boldsymbol{A} = \{i \mid \mathcal{R}(f(\theta^i, x+\zeta_1), f(\theta^i, x+\zeta_2)) = \text{True}\}$ であり，$\mathcal{R}(f(\theta^i, x+\zeta_1), f(\theta^i, x+\zeta_2))$ は入力 x に対しネットワーク i の知識が優れているかの真偽値である．優れた知識については次の節に詳細を示す．

安定化損失は，一貫性損失を知識の指標として，最も一貫性損失が小さいネットワーク θ^s から他の全てのネットワークに知識を転移する．転移先のネットワークを θ^i とすると，知識は，ネットワーク θ^i の事後確率を，ネットワーク θ^s の事後確率に近づけるように学習することで転移させる．

3.4 優れた知識

優れた知識は，ある入力データに対して，異なる摂動を付与したときの 2 つの推測値 $f(\theta^i, x+\zeta_1), f(\theta^i, x+\zeta_2)$ において，以下の 2 つの条件を満たすネットワークが持つ．各例を図 3 に示す．

- 双方の推測値の推測クラスが一致する (例：A, B)
- 双方の推測値がともに閾値を越える (例：B, C)

図 3 中の例では，B が条件を 2 つとも満たしているため，B のような出力においてネットワークが優れた知識を持っている．

また，ネットワークが優れた知識を持つかの真偽値 $\mathcal{R}^i_x$ を式 (4) に示す．

$$\mathcal{R}^i_x = \{\mathcal{P}^i_{x+\zeta_1} = \mathcal{P}^i_{x+\zeta_2}\}_1 \& (\{\mathcal{M}^i_{x+\zeta_1} > \xi\}_1 || \{\mathcal{M}^i_{x+\zeta_2} > \xi\}_1) \tag{4}$$

ここで，$\mathcal{M}^i_x$ は $\max(f(\theta_i, x))$ である．また，$\mathcal{P}$ は推測クラス，ξ は閾値である．$\{\}_1$ は括弧内の条件が真のときは 1，偽のときは 0 の値をとる．& の左側が 1 つ目の条件，右側が 2 つ目の条件を表す．

表 1: RandAugment に用いる変換方法

関数名	詳細
AutoContrast	RGB ごとに各画素値の最小値と最大値が 0, 255 になるように正規化する．
Brightness	M の大きさに従い，明度を大きくする．
Color	M に従い，色調を調整する．M が小さいほどグレースケール画像に近づく．
Contrast	M に従い，コントラストを調整する．M が大きいほど階調が大きくなる．
CutoutAbs	Cutout [Terrance 17] を行う．矩形の大きさは M の大きさに従う．
Equalize	RGB ごとにヒストグラムを平坦化する．
Identity	変換を行わず，元画像をそのまま返す．
Invert	ネガポジ変換を行う．
Posterize	M の大きさに従い，量子化 bit 数を小さくする．
Rotate	画像を回転する．回転角は M の大きさに従う．
Sharpness	M に従い，先鋭化を行う．
ShearX(Y)	せん断を行う．せん断角は M に従う．
Solarize	M に従って閾値を決定し，閾値より大きい画素値をもつ画素に対してネガポジ反転を行う．
SolarizeAdd	M に従った定数値を各画素に加算し，閾値を 128 とする Solarize を行う．
TranslateX(Y)	M に従って画素数を決定し，水平 (垂直) 方向に平行移動を行う．

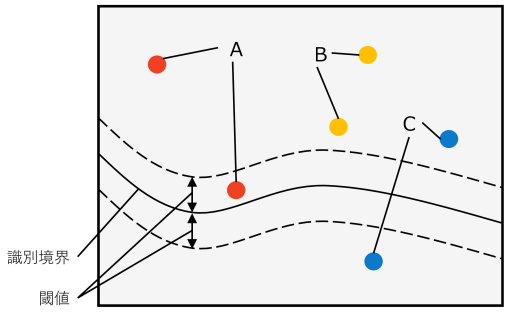


図 3: $f(\theta^i, x+\zeta_1), f(\theta^i, x+\zeta_2)$ の 2 点における各例

3.5 学習方法

提案手法におけるネットワークの更新方法を Algorithm 1 に示す．提案手法は，入力データのラベルの有無によって処理が異なり，ラベルありデータに対して教師あり損失 $\mathcal{L}_{cls}$，すべてのデータに対して一貫性損失 $\mathcal{L}_{con}$，ラベルなしデータに対して安定化損失 $\mathcal{L}_{sta}$ を求める．

Algorithm 1 半教師あり学習における Refined Consitency の学習方法

Require: ラベルありデータとラベルなしデータを含むバッチ $\mathcal{B}$，ネットワーク数 N，独立したネットワーク $f(\theta^1), f(\theta^2), \ldots, f(\theta^N)$

for 各バッチ $\mathcal{B}$ **do**
  for 各ネットワーク $f(\theta^i)$ in $\{f(\theta^1), f(\theta^2), \ldots, f(\theta^N)\}$ **do**
    ラベルありデータ x について $\mathcal{L}^i_{cls}$ を求める
    式 (2) に従い，すべてのデータ x について $\mathcal{L}^i_{con}$ を求める
  end for
  for 各ラベルなしデータ x **do**
    for 各ネットワーク $f(\theta^i)$ in $\{f(\theta^1), f(\theta^2), \ldots, f(\theta^N)\}$ **do**
      式 (4) に従い，$\mathcal{R}^i_x$ を求める
    end for
    for 各ネットワーク $f(\theta^i)$ in $\{f(\theta^1), f(\theta^2), \ldots, f(\theta^N)\}$ **do**
      式 (3) に従い，$\mathcal{L}^i_{con}$ を用いて $\mathcal{L}^i_{sta}$ を求める
    end for
  end for
  for 各ネットワーク $f(\theta^i)$ in $\{f(\theta^1), f(\theta^2), \ldots, f(\theta^N)\}$ **do**
    式 (1) に従い，$\mathcal{L}^i$ を求める
    $\mathcal{L}^i$ により確率的勾配降下法で θ^i を更新
  end for
end for

4. 評価実験

半教師あり学習における提案手法の有効性を確認するため，提案手法と Multiple Student で精度を比較する．データセットに CIFAR-100 [Krizhevsky 09]，ベースモデルに 13 層の CNN を使用する．CIFAR-100 は，クラス数が 100 であり，学習用に 50,000 枚，推論用に 10,000 枚の画像で構成される．本実験では，学習用画像のうち 10,000 枚をラベルありデータ，他 40,000 枚をラベルなしデータとする．学習回数は 350 epoch，ネットワーク数は 4，8 とする．データ拡張として，左右反転，ランダムな画素数平行移動したのちに反射パディングを行う．

また，近年の一貫性に基づいた半教師あり学習において，RandAugment などのより大きな摂動を付与するデータ拡張手法を扱うことで精度が向上することが確認されている [Xie 20] [Sohn 20]．そこで，Multiple Student および提案手法に対し，RandAugment により大きな摂動を付与することができるデータ拡張を適用し，RandAugment の有無による精度変化を評価する．

4.1 RandAugment

RandAugment は，あらかじめ複数種の単純な画像変換方法を用意しておき，2 つのパラメータ N, M により画像変換を制御することで，学習条件に適したデータ増幅を行う手法である．学習時は，複数種の画像変換方法から変換方法を N 個サンプリングし，各画像変換を強さ M(Magnitude) で適用する．画像の一定枚数ごとに変換方法を N に従ってサンプリングするため，変換の多様性によって大きな摂動を付与することができる．本研究において，RandAugment に用いる変換方法の種類を表 1 に示す．本実験では，N を 1，M を 2 として適用する．

4.2 従来手法との精度比較

提案手法および Multiple Student の精度，各手法に対して RandAugment(RA) を適用した精度を表 2 に示す．各結果は 5 回試行した時の正解率の平均値と標準偏差である．

表 2: CIFAR-100 を用いた精度比較 [%]

手法	拡張	ネットワーク数	
		4	8
Multiple Student	-	66.70±0.15	66.93±0.09
	RA	66.43±0.14	66.34±0.10
提案手法	-	66.96±0.12	67.24±0.06
	RA	66.96±0.04	**68.18**±0.10

表 2 から，Multiple Student と提案手法でネットワーク数を 4 から 8 に上げた場合，平均精度が上昇し，標準偏差が低下している．このことから，ネットワーク数を上げることにより，知識を転移する頻度が上昇していることが確認できる．提案手法は Multiple Student と比べて，ネットワーク数 4 で 0.26 pt，ネットワーク数 8 で 0.31 pt 精度が高い．よって，提案手法における知識転移は有効であると言える．さらに，RandAugment と提案手法を組み合わせることで，その効果は最大となり 68.18% まで向上した．

4.3 損失の制限による精度比較

一貫性損失 $\mathcal{L}_{con}$ と安定化損失 $\mathcal{L}_{sta}$ の有効性を評価する．評価方法として，式 1 の損失関数において $\mathcal{L}_{con}, \mathcal{L}_{sta}$ の重み λ_1, λ_2 を 0 とすることで，ネットワークの更新に用いる損失を制限する．そして，制限する場合と制限しない場合で精度がどれだけ低下するかを評価し，一貫性損失と安定化損失がそれぞれ精度向上に対して有効である事を確認する．ネットワーク数 8 の場合に各損失を使用した精度を表 3 に示す．

表 3: 各損失の使用による精度比較 [%]

Loss		Accuracy
一貫性損失 $\mathcal{L}_{con}$	安定化損失 $\mathcal{L}_{sta}$	
		60.67±0.19
	✓	67.01±0.27
✓		62.44±0.16
✓	✓	67.24±0.06

表 3 から，$\mathcal{L}_{con}$ のみの使用と $\mathcal{L}_{sta}$ のみの使用により精度が低下していることが確認できる．また，$\mathcal{L}_{con}$ と $\mathcal{L}_{sta}$ の双方を使用しない場合に大幅に精度が低下することから，$\mathcal{L}_{con}$ と $\mathcal{L}_{sta}$ を組み合わせることで，さらに一貫性が向上していることが確認できる．

5. おわりに

本研究では，1 度に全てのネットワーク間で知識を転移する Refined Consistency を提案した．評価実験では，提案手法が MS に比べて高精度であることを確認し，提案手法による知識転移が有効であることを示した．また，一貫性損失と安定下損失を組み合わせることで，さらに一貫性が向上していることを確認した．今後は，知識転移グラフへの応用を行う予定である．

6. 謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP18002) の結果得られたものである．

参考文献

[Zhanghan 19] Zhanghan, K., *et al.*: Dual Student: Breaking the Limits of the Teacher in Semi-supervised Learning, *ICCV*, pp. 6728–6736 (2019).

[Cubuk 19] Cubuk, E., *et al.*: RandAugment: Practical automated data augmentation with a reduced search space, *NIPS*, Vol. 33, pp. 18613–18624 (2019).

[Xie 20] Xie, Q., *et al.*: Unsupervised Data Augmentation for Consistency Training, *NIPS*, Vol. 33, pp. 6256–6268 (2020).

[Sohn 20] Sohn, K., *et al.*: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, *NIPS*, Vol. 33, pp. 596–608 (2020).

[Laine 17] Laine, S., *et al.*: Temporal Ensembling for Semi-Supervised Learning, *ICLR* (2017).

[Srivastava 14] Srivastava, N., *et al.*: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *JMLR*, Vol. 15, No. 56, pp. 1929–1958 (2014).

[Tarvainen 17] Tarvainen, A., *et al*: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *NIPS*, Vol. 30, pp. 1195–1204(2017).

[Radosavovic 18] Radosavovic, I., *et al.*: Data Distillation: Towards Omni-Supervised Learning, *CVPR*, pp. 4119–4128 (2018).

[Verma 19] Verma, V., *et al.*: Interpolation Consistency Training for Semi-Supervised Learning, *IJCAI*, pp. 3635–3641 (2019).

[Suzuki 20] Suzuki, T. and Sato, I.: Adversarial Transformations for Semi-Supervised Learning, *AAAI*, Vol. 34, No. 04, pp. 5916–5923 (2020).

[Zhang 18] Zhang, H., *et al.*: mixup: Beyond Empirical Risk Minimization, *ICLR* (2018).

[Krizhevsky 09] Krizhevsky, A. and Hinton, G.: Learning multiple layers of features from tiny images, *Master's thesis, Department of Computer Science, University of Toronto* (2009).

[Terrance 17] Terrance, D. and Graham, W.: Improved Regularization of Convolutional Neural Networks with Cutout, *arXiv prepront, arXiv:1708.04552* (2017).