Human-in-the-loop による設計空間の絞り込み法を導入した知識転移グラフの探索

Exploring knowledge transfer graphs
by introducing design space refinement method with human-in-the-loop

岩田 幸　南 蒼馬　平川 翼　山下 隆義　藤吉 弘亘
Sachi Iwata　Soma Minami　Tsubasa Hirakawa　Takayoshi Yamashita　Hironobu Fujiyoshi

中部大学
Chubu University

Deep collaborative learning is a method of transferring knowledge between multiple networks. Knowledge transfer graph has been proposed as deep collaborative learning that makes a rich in diversity of knowledge transfer. However, designing a knowledge transfer graph is difficult due to many combinations, so it is not clear the trend for highly accurate knowledge transfer graphs. To address this problem, we propose a method for designing search space with human-in-the-loop for knowledge transfer graph. We analyze the trend of graphs and designing graphs with high accuracy based on the acquired results. The experimental results with CIFAR-100 show that the search space explored by the proposed method is better than that of deep mutual learning. We confirmed that the accuracy of the best knowledge transfer graph in the search space is better than that of using the asynchronous successive halving algorithm.

1. はじめに

一般社会において，教師から生徒へ知識を転移するだけではなく，生徒同士で知識を転移し合うことにより学力を向上することができる．深層学習のネットワークモデルの学習においても，教師ラベルのみを用いて学習した場合と比べて，ネットワーク間での知識転移を取り入れた手法の方がネットワークの性能が向上することが知られている [Hinton 15][Zhang 18][Minami 20]．

ネットワーク間の知識転移の代表的な手法として，Knowledge Distillation (KD) [Hinton 15] や Deep Mutual Learning (DML) [Zhang 18] がある．KD はパラメータ数が大きな Teacher ネットワークからパラメータ数が小さな Student ネットワークへ知識を伝達することにより，Student ネットワークの性能を向上することができる．DML は複数のネットワークを同時に学習し，その際互いに知識を伝達することにより，ネットワークの性能を向上することができる．ネットワーク間の知識転移の手法は，ネットワークを圧縮するという目的だけではなく，ネットワークの性能向上を目的とした研究でも検討されている [Zhang 18][Minami 20]．本稿では，KD や DML などの複数のネットワーク間で知識を伝達する学習法のことを，ネットワーク間の共同学習という．

ネットワーク間の共同学習を大規模化した知識転移グラフ [Minami 20] がある．知識転移グラフでは，KD や DML をグラフ構造で表現することによって，KD や DML を内包した学習方法を表現することができる．また，知識の転移をゲート関数を用いて制御することにより，不要な知識の伝達を抑制することができる．知識転移グラフでは各ネットワークをノードで表現し，ノード間の知識の伝達をエッジとして有向グラフで表現する．ノードとエッジを自動最適化することにより，最適な共同学習法を獲得することができる．しかし，知識転移グラフの探索空間は膨大であるため，精度の高いグラフに共通する傾向はまだ定かでない．

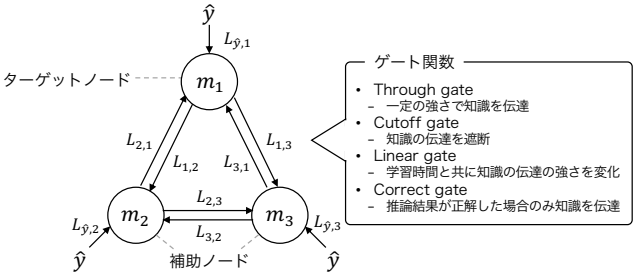


図 1: 知識転移グラフ

本研究では，認識精度の高いグラフにはどのようなノードやエッジが用いられるかを分析して，得られた傾向を基に知識転移グラフの設計空間の絞り込みを行い，最適な共同学習法を獲得することを目的とする．ここで設計空間とは，最適な知識転移グラフを見つけるために人が設計する探索空間のことを指す．まず，知識転移グラフをランダムサーチして，認識精度の高いグラフに共通する傾向を分析する．その次に，見つかった傾向を基に人が知識転移グラフの設計空間を限定する．これを十分に設計空間が狭くなるまで繰り返す．最終的に絞り込んだ設計空間を全探索することにより，最適なグラフを獲得する．絞り込んだ設計空間と DML を比較し，提案手法によって DML と比べて精度の高いグラフの多い設計空間を獲得できるか確認する．さらに，最終的に絞り込んだ設計空間の全探索によって発見した最適なグラフと，従来の探索手法である Asynchronous Successive Halving Algorithm (ASHA) [Li 20] を使用した探索結果を精度比較する．

2. 関連研究

複数のネットワーク間で知識を伝達する共同学習は，知識の伝達が一方向なもの，相互に伝達し合うもの，2 つを組み合わせたものの 3 つに大別される．

連絡先:
岩田 幸：isachi@mprg.cs.chubu.ac.jp
南 蒼馬：minami@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

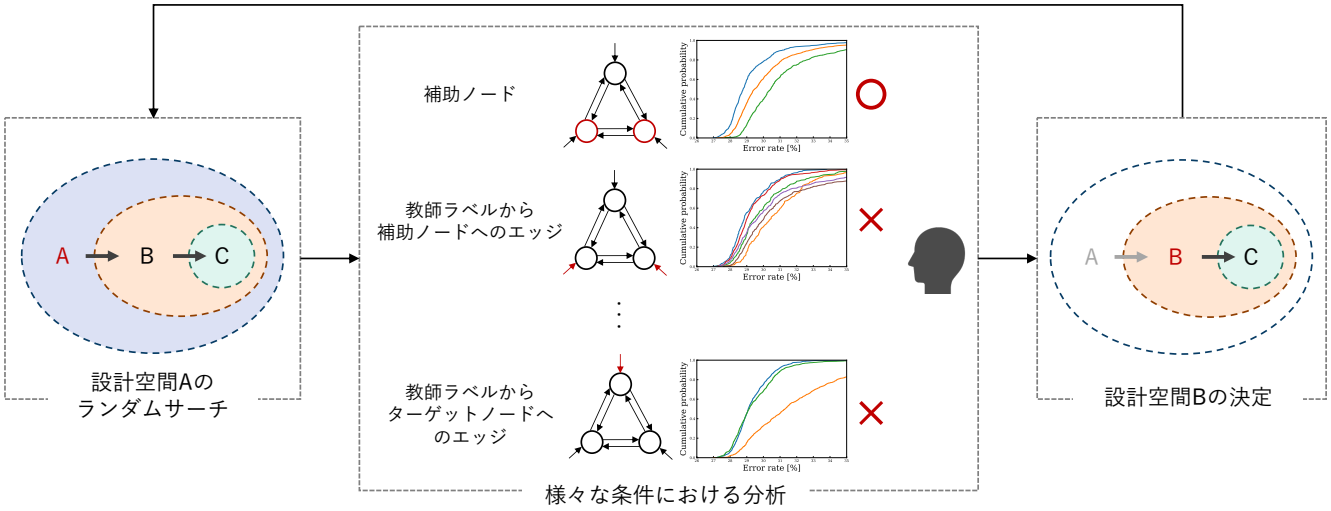


図 2: 提案手法の流れ

知識の伝達方向が一方向である代表的な手法として，Knowledge Distillation (KD) [Hinton 15] がある．KD は，学習済みでパラメータ数が大きな Teacher ネットワークの知識を使用して，未学習でパラメータ数が小さな Student ネットワークを学習する手法のことを指す．

知識の伝達方向が双方向である代表的な手法として，Deep Mutual Learning (DML) [Zhang 18] がある．DML は，Teacher ネットワークを使用せず Student ネットワーク間で知識を伝達しあうことにより，Student の認識精度を向上させる手法のことを指す．DML の学習方法は，KD のようなネットワークの大小といった関係だけではなく，小さなネットワーク間で知識を伝達することや，3 つ以上のネットワーク間で知識を伝達することが可能である．

一方向と双方向を組み合わせた知識転移をする手法として，知識転移グラフ [Minami 20] がある．図 1 に知識転移グラフを示す．m はネットワーク，L は損失関数，$\hat{y}$ は教師ラベルである．知識転移グラフは，各ネットワークをノードで表現し，ノード間の知識の伝達をエッジとして表現する．知識転移グラフのハイパーパラメータは，ネットワークとゲート関数である．ゲート関数には，学習中に知識を伝達する強さを一定にする Through gate，学習中に知識を伝達を遮断する Cutoff gate，学習中に知識を伝達する強さに変化を加える Linear gate，学習中に推論結果が正しい場合のみ知識を伝達する Correct gate がある．精度を向上させたいノードをターゲットノードとして，ターゲットノードをサポートするノードを補助ノードとする．これらのハイパーパラメータを自動最適化し，最適なグラフを獲得することで，多様な知識の伝達方法の獲得が可能となり，ターゲットノードの認識精度を向上させることが可能となる．ただし，ノード数を 3 とした知識転移グラフでは，ノードやエッジといったハイパーパラメータの組み合わせは 1,179,648 通りとなり，全探索は現実的な時間で計算不可能である．

3. 提案手法

知識転移グラフの探索空間は膨大であるため，精度が高いグラフに共通する傾向はまだ定かでない．そこで，高精度なグラフに共通するハイパーパラメータを分析し，設計空間を Human-in-the-looop で段階的に狭める手法を提案する．

3.1 評価指標

設計空間の評価には，Empirical Distribution Function (EDF) [Radosavovic 19] を使用する．EDF はエラー率が閾値 e より低いグラフの割合を表す指標のことである．式 (1) に EDF を示す．ここで，n は探索で得られたグラフの数，e は閾値，e_i は i 番目のグラフのエラー率を示す．エラー率が低いグラフが多く含まれる設計空間であるほど，良い設計空間である．設計空間を絞り込む際には，EDF をグラフにプロットし，人が見て判断する．この指標を用いて設計空間が良いか悪いかを判断するには，十分な探索回数が必要である．EDF をスカラーで比較する方法の 1 つとして，Area Under the Curve (AUC) が挙げられる．しかし，AUC では極端に良い設計空間と悪い設計空間が重なった設計空間を，良い設計空間と判断することは不可能であるため，人が認識精度の高いグラフが多く含まれる設計空間であるかを判断する．

$$F(e) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[e_i < e] \tag{1}$$

3.2 提案手法の流れ

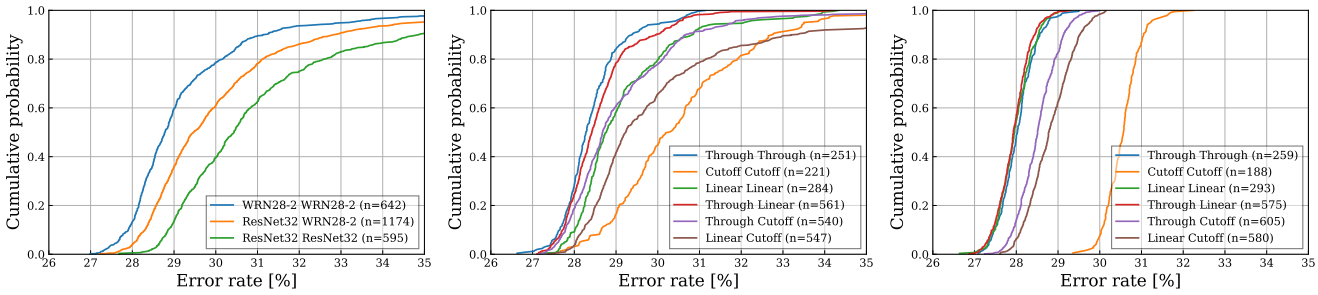
提案手法は精度が高いグラフに共通するハイパーパラメータに着目して設計空間を絞り込む．図 2 に提案手法の流れを示す．設計空間 A から絞込み，設計空間 B を得る．これを組み合わせ数が十分少なくなるまで繰り返して設計空間 X を獲得する．この探索方法は設計空間を段階的に絞り込むため，ハイパーパラメータは次第に限定され，局所最適解に陥りにくい．提案手法は以下の Step 1 から Step 4 を用いて設計空間を段階的に絞り込み，知識転移グラフを探索する．

Step 1. 知識転移グラフの設計空間をランダムサーチする．

Step 2. 補助ノード，教師ラベルから補助ノードへのエッジ，教師ラベルからターゲットノードへのエッジ，補助ノードからターゲットノードへのエッジ，Pre-trained モデルの有無の 5 条件で EDF を使用した分析を行う．

Step 3. Step 2 の分析を基に精度の高いグラフに共通するハイパーパラメータを選択して，設計空間を限定する．

Step 4. 設計空間が十分に小さくなるまで Step 1 から Step 3 を繰り返し，限定した設計空間 X に対して全探索を行う．



(a) 設計空間 A → B　(b) 設計空間 B → C　(c) 設計空間 C → X

図 3: 各設計空間の EDF

表 1: 設計空間

設計空間	制約	組み合わせ数
A	なし	39,366
B	$+m_2 = m_3 =$ WRN28_2	9,842
C	$+L_{\hat{y},2} = L_{\hat{y},3} =$ Through	1,094
X	$+L_{2,1} = L_{3,1} =$ Through or Linear	486

4. 評価実験

知識転移グラフの設計空間を段階的に絞り込み，提案手法により限定した設計空間 X と DML を比較する．さらに，ASHA を使用した探索結果と精度比較して，提案手法の有効性を示す．

4.1 実験条件

一般物体認識用データセットである CIFAR-100 [Krizhevsky 09] を使用して実験を行う．CIFAR-100 は Train 用に 50,000 枚，Test 用に 10,000 枚の計 60,000 枚で構成される．Train 用のデータセット 50,000 枚のうち，訓練用に 40,000 枚，検証用に 10,000 枚ランダムに振り分けて使用する．最適化アルゴリズムは，確率的勾配降下法 (SGD) に Nesterov momentum を適応させる．初期の学習率は 0.1，Momentum は 0.9，バッチサイズは 64 とする．学習率のスケジューラは最終エポックで 0 となる半周期のコサイン関数を使用する．設計空間 A〜C の探索時のエポック数は 100 であるが，設計空間 X の探索時のエポック数は 200 である．ゲート関数は，Through gate，Cutoff gate と Linear gate を使用する．ネットワークモデルは，ResNet32 [He 15] と WRN28-2 [Zagoruyko 16] を使用する．設計空間 A〜C の探索回数は各 2,500 回であり，設計空間 X は全探索を 5 回行う．

知識転移グラフはノードとエッジを自動最適化するため，常に高精度なグラフが獲得できるとは限らない．認識精度低下例の 1 つとして，教師ラベルの知識がターゲットノードに伝達されない場合が挙げられる．本実験ではターゲットノードの認識精度が 2%以下のグラフは分析に使用しない．

4.2 設計空間の絞り込み

設計空間の絞り込みを行う．まず，設計空間 A を探索する．探索に使用するハイパーパラメータに制約はない．設計空間 A を探索して，最も効果の高い補助ノードについて分析を行う．図 3(a) に補助ノードについての EDF を示す．横軸は閾値であり，縦軸はエラー率が閾値以下となるグラフの割合を示す．n は探索によって獲得したグラフの数である．補助ノードに WRN28-2 を 2 つ使用すると高精度なグラフを多く獲得できることが確認できる．このことから，認識精度の高いモデルは表現力があり，ターゲットノードの認識精度向上に貢献すると考えられる．次に，設計空間 B を探索する．設計空間 B は，設計空間 A の探索で最も効果が高い WRN28-2 (補助ノード) を用いた設計空間である．設計空間 B を探索して，最も効果の高い教師ラベルから補助ノードへのエッジで分析を行う．図 3(b) に教師ラベルから補助ノードへのエッジの EDF を示す．教師ラベルから補助ノードへ Through gate を使用すると高精度なグラフを多く獲得できることが確認できる．次に，設計空間 C を探索する．設計空間 C は，設計空間 B の探索で最も効果が高い Through gate (教師ラベルから補助ノードへ) を用いた設計空間である．設計空間 C を探索して，最も効果の高い補助ノードからターゲットノードへのエッジで分析を行う．図 3(c) に補助ノードからターゲットノードへのエッジの EDF を示す．補助ノードからターゲットノードへは Through gate もしくは Linear gate を使用すると高精度なグラフを多く獲得できることが確認できる．このことから，学習を通して教師ラベルを用いて学習した各補助ノードの知識はターゲットノードの認識精度向上に繋がると考えられる．設計空間 C の探索で最も効果が高い Through gate もしくは Linear gate (補助ノードからターゲットノードへ) を用いた設計空間を設計空間 X とする．

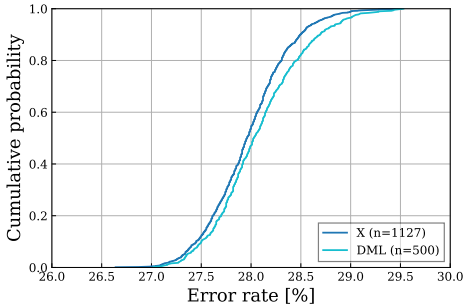
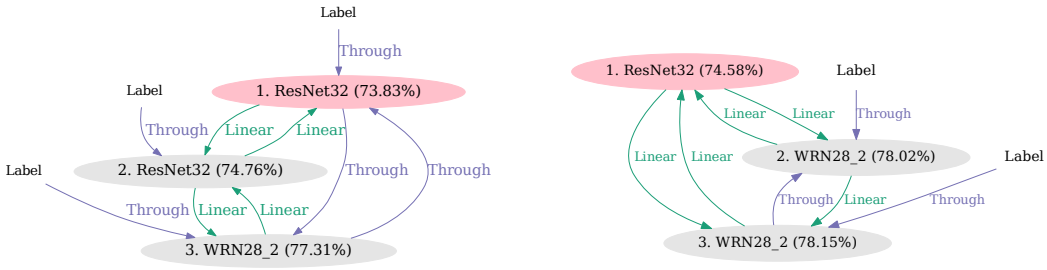


図 4: 設計空間 X と DML の比較

表 1 に提案手法により絞り込んだ設計空間を示す．設計空間の絞り込みによって獲得した設計空間 X の組み合わせ数は初期の設計空間である設計空間 A の組み合わせ数の 1/81 であり，設計空間 A を全探索する際と比較して設計空間 X は全探索の探索回数を大幅に削減できる．

さらに，設計空間の絞り込みによって獲得した設計空間 X と DML を比較する．図 4 に設計空間 X と DML の比較を示



(a) ASHA (設計空間 A)　　(b) Ours (設計空間 X)

図 5: 獲得した Top1 のグラフ

表 2: ASHA との比較 [%]

探索方法	DML	Optimized
ASHA (設計空間 A)	73.66 ± 0.35	73.96 ± 0.16
Ours (設計空間 X)	74.06 ± 0.26	$\mathbf{74.52} \pm 0.19$

す．ここで，DML とはエッジを全て Through gate にしたグラフのことである．DML に比べて精度の高いグラフが多く含まれる設計空間を獲得することができた．

4.3　ASHA を使用した探索との比較

ASHA を使用した探索によって得られた最適なグラフと設計空間 X を全探索することによって得られた最適なグラフの認識精度の比較を行う．ASHA とは，1, 2, 4, $\cdots$, 2^n エポック目において，それまでに獲得したグラフのうち認識精度の 50% 以下であれば，学習を打ち切り，次の探索に移る手法である．ASHA を使用した探索における探索回数は [Minami 20] と同様に 1,500 回である．表 2 に ASHA を使用した探索で獲得した Top1 のグラフと設計空間 X を全探索して獲得した Top1 のグラフを各 5 回学習した平均精度と標準偏差を示す．ここで，“Optimized” は各探索によって獲得した Top1 のグラフのことを指す．エッジを全て Through gate にした DML に比べて高精度なグラフを獲得できた．ASHA を使用して探索した結果と比較しても，高精度なグラフを獲得することができた．

図 5 に探索によって獲得したグラフを示す．赤色のノードはターゲットノードを表し，括弧内の数値は各ノードの認識精度である．“Label” は教師ラベルを指す．エッジが表記されていないのは Cutoff gate が選択された箇所である．図 5(a) に設計空間 A を ASHA を用いた探索によって獲得した最適な知識転移グラフを示す．獲得したグラフから，学習初期はノード間の知識の伝達の強さを抑えるグラフが高い認識精度であることが確認できる．図 5(b) に設計空間 X の全探索によって獲得した最適な知識転移グラフを示す．獲得したグラフから，学習初期は補助ノードを学習し，次第にターゲットノードへ知識転移をするグラフが高い認識精度であることが確認できる．

5.　おわりに

本稿では，知識転移グラフの設計空間を段階的に狭める手法を提案した．設計空間を段階的に狭めることにより，DML と比較して精度の高いグラフが多く含まれる設計空間を獲得した．提案手法によって獲得した設計空間 X は設計空間 A を全探索する場合と比べて少ない探索回数で認識精度の高いグラフの獲得が可能となった．また，ASHA で探索した場合と比べて高精度なグラフを獲得した．今後の予定としては，ベイズ最適化を用いた知識転移グラフの探索と傾向調査が挙げられる．

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

[Hinton 15] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, in *Neural Information Processing Systems Deep Learning Workshop* (2015)

[Zhang 18] Zhang, Y., Xiang, T., Hospedales, T. M. and Lu, H.: Deep Mutual Learning, in *IEEE Conference on Computer Vision and Pattern Recognition* (2018)

[Minami 20] Minami, S., Hirakawa T., Yamashita, T. and Fujiyoshi, H.: Knowledge Transfer Graph for Deep Collaborative Learning, in *Asian Conference on Computer Vision* (2020)

[Li 20] Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B. and Talwalkar, A.: A System for Massively Parallel Hyperparameter Tuning, in *Proceedings of Machine Learning and Systems* (2020)

[Radosavovic 19] Radosavovic, I., Johnson, J., Xie, S., Lo, W. and Dollár, P.: On Network Design Spaces for Visual Recognition, in *International Conference on Computer Vision* (2019)

[Krizhevsky 09] Krizhevsky, A. and Hinton, G.: Learning multiple layers of features from tiny images, Citeseer (2009)

[He 15] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, in *IEEE Conference on Computer Vision and Pattern Recognition* (2015)

[Zagoruyko 16] Zagoruyko, S. and Komodakis, N.: Wide Residual Networks, in *British Machine Vision Conference* (2016)