Knowledge distillation:
A good teacher is patient and consistent

読み会@20221025

楊明哲

論文情報と選択理由

- 論文情報

Lucas Beyer* Xiaohua Zhai* Amélie Royer*† Larisa Markeeva*‡ Rohan Anil Alexander Kolesnikov*
Google Research, Brain Team
{lbeyer,xzhai,akolesnikov}@google.com

- 選択理由
  - 知識蒸留の問題設定の教師と生徒の関係は,
    機械教示と類似しているため

CV技術発展しているけど，実社会で使いづらい

- 背景：大規模モデルは性能が高いが，計算コスト，設備コストが高すぎる
- →現実的には小さなモデルの方が使い勝手が良く重要

- 大規模モデルの能力を活用するために，性能維持したままモデルを軽量化することが求められている

KDは「教師と生徒のモデルマッチング」と解釈

- モデルの軽量化には2つの手法が有名
- 1. Model pruning
  - モデルの一部をカットすることで，軽量化をおこなう．
  - モデルの構造を変更できない e.g. ResNet → MobileNet
- 2. Knowledge distillation
  - 教師（大きめ）と生徒（小さめ）を設定し．生徒は教師と同じ能力を目指す

知識蒸留は，入力データと学習時間が大事

- ここでは，モデルを軽量化するために，知識蒸留を利用
- 知識蒸留において何が大事なのかを見つける

- 1. 生徒と教師は同じ入力を用いることが大事
- 2. 入力データはデータ拡張するほど良い
- 3. 学習時間は多い方が良い

問題設定・実験設定

- 評価指標に画像のクラス分類精度を用いる
- 大規模モデルを教師と設定し，分類精度を落とさずに，モデルの軽量化を目指す.
  - 教師：BiTの事前学習モデル
  - 生徒：ResNet-50
- データセットは5種類(小中規模4つ，大規模1つ)
  - クラス数は37-1000, データサイズは1010-1281167

教師と生徒の出力分布を近づける

- 知識蒸留の学習に用いる損失でKLダイバージェンス用いる

$$KL(p_t||p_s) = \sum_{i \in C} [-p_{t,i} \log p_{s,i} + p_{t,i} \log p_{t,i}]$$

- p_t, p_sはそれぞれ教師の予測，生徒の予測
- モデルの出力自体は，温度付きソフトマックス

ハイパラ設定とか実験の小技たち

- 最適化手法としてAdamを用いる．ハイパラはデフォルト値
- スケジューラはCosine decayを用いる
- 学習安定のため，勾配の大きさを1.0に制限
- バッチサイズを512に設定．ただしImageNetは4096

Mixupによってデータ拡張をする

- 2つの訓練データのペアを混合して新しくデータを作成[1]
- 訓練データがミックスされるだけでなく，**ラベルもミックス**
- $X = \lambda X_1 + (1-\lambda)X_2, y = \lambda y_1 + (1-\lambda)y_2$
- $\lambda \in [0,1]$から一様にサンプリングする(本家は$\lambda \sim Beta(\alpha, \alpha)$)

https://arxiv.org/abs/1710.09412

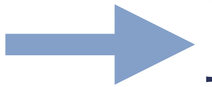
画像の前処理

- データの前処理でinception-style cropを用いる
- 1. 元データの画像サイズの(0.08-1)でランダムにクリップ
- 2. クリップした後ランダムなアスペクト比に変換
- 3. 変換したあと固定サイズ(224x224)にリサイズ

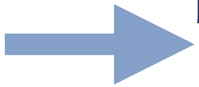
https://arxiv.org/abs/1409.4842

主張のおさらい

- 1. 生徒と教師は同じ入力を用いることが大事
- 2. 入力データはデータ拡張するほど良い
- 3. 学習時間は多い方が良い



Consistent teacher



Patient teacher

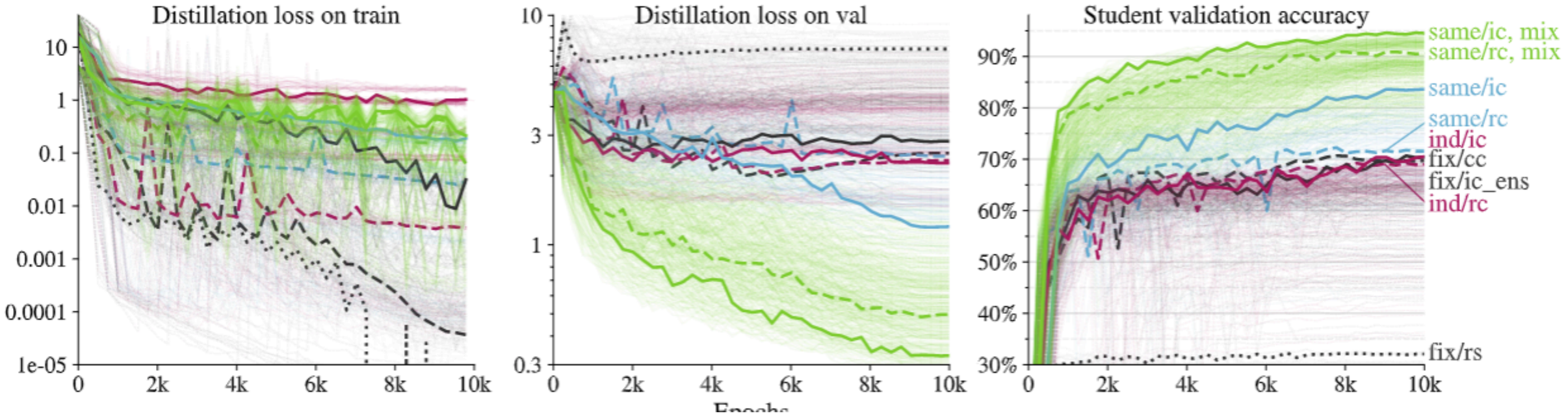
- ロバスト性を見るために小中規模データセット4つを利用
- 交絡因子を排除するためにさまざまなハイパラで実験

Consistent teacher (一貫した教師)についてみる

- 教師，生徒の設定を4つ用意する.
- Fixed teacher：教師の出力が固定されている
- Independent noise：教師，生徒で異なる入力
- Consistent teaching：教師と生徒で同じ入力
- Function matching：Consistent teaching＋データ拡張

- 下二つが一貫した教師設定を表している.

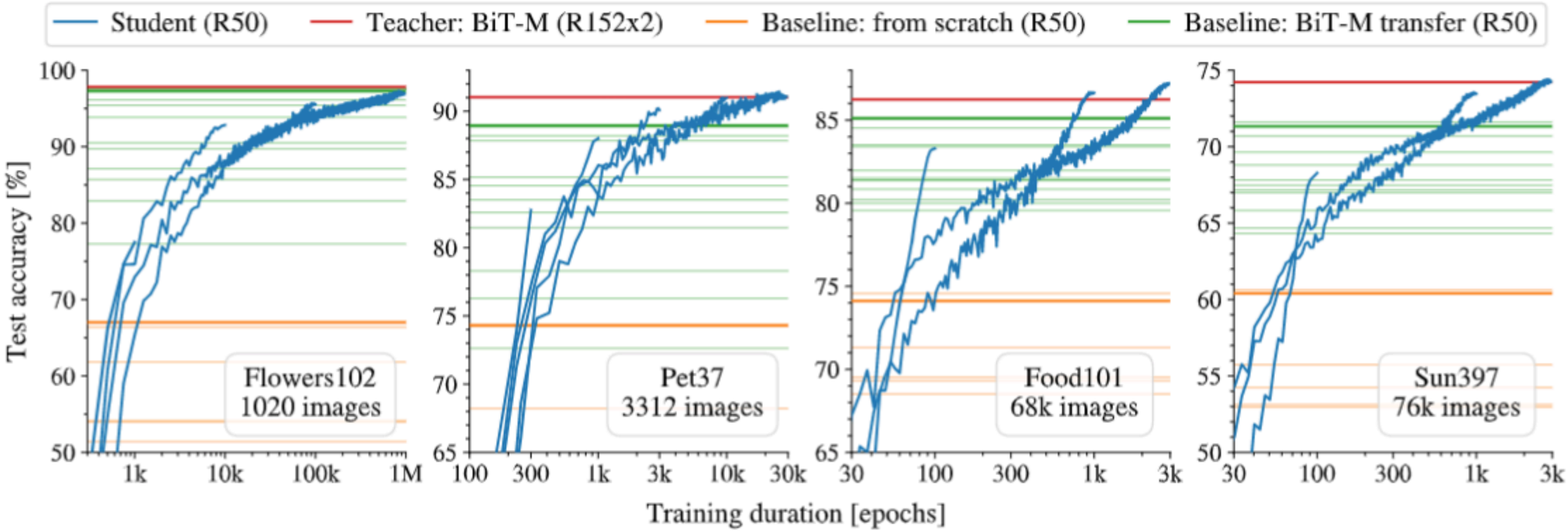
一貫した教師は精度が高い. 他は過学習を起こす



Patient teacher(忍耐強い教師)についてみる

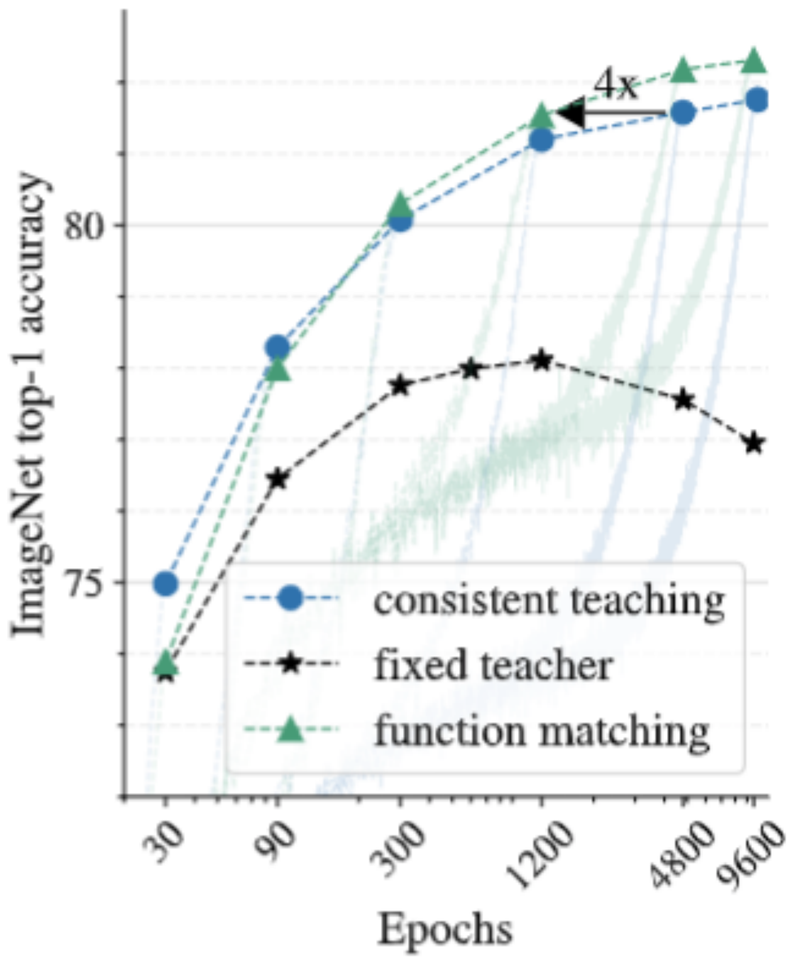
- 一般的な教師あり学習の場合，ラベルに対し，画像が大きく歪む可能性がある.
- ここでは，知識蒸留=教師と生徒の関数マッチングと解釈
  - 一貫して同じ入力を与えるなら，入力自体は歪んでもいい
    → データ拡張が有効なのでは?
- データ拡張を行い，大きなエポック数で学習するとより生徒モデルの性能が向上することを検証する

過学習を回避しながら，性能が向上する



- 赤い線（教師）と同等の性能に到達
- 緑線（転移学習）より最終的には性能が良い
- オレンジ（0から学習）より性能がいい → 知識蒸留の効果がある

大規模データセット（ImageNet）で知識蒸留



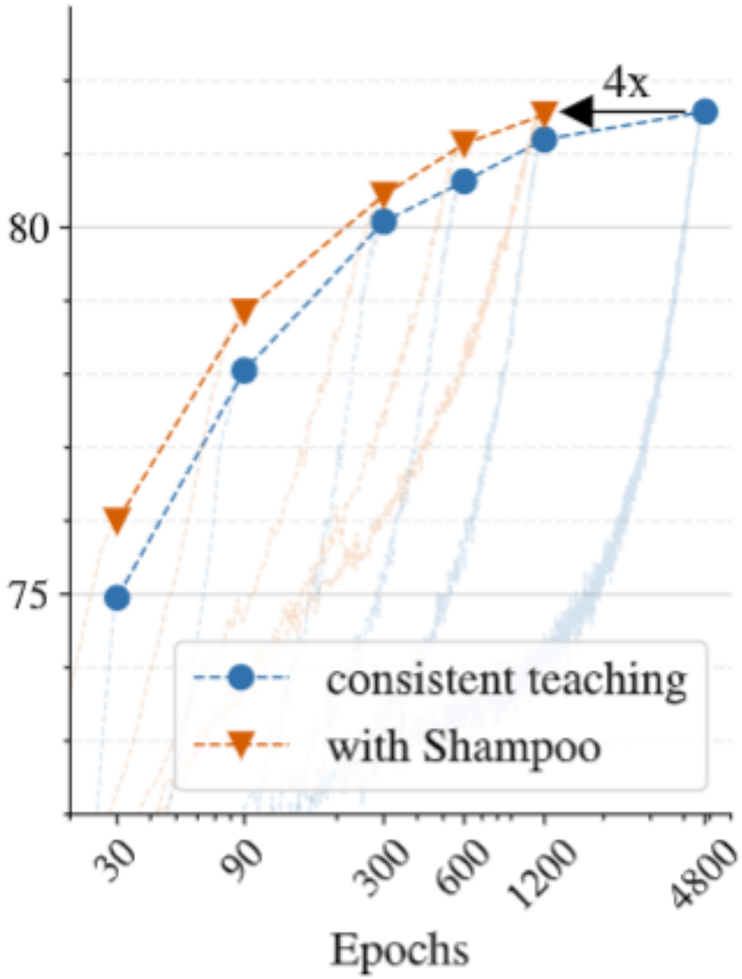
- 一貫した教師は過学習していない
- またデータ拡張すると少ないエポック数で高い精度を達成
  - → iteration数は同じ??

異なる解像度でも知識蒸留がうまくいく

Experiment	300	1200	4800	9600
T224 → S224	80.30	81.54	82.18	82.31
T224 → S160	78.17	79.61	N/A	80.49
T384 → S224	80.46	81.82	82.33	82.64

- 教師が高い解像度→生徒が低い解像度の方が性能が良くなる

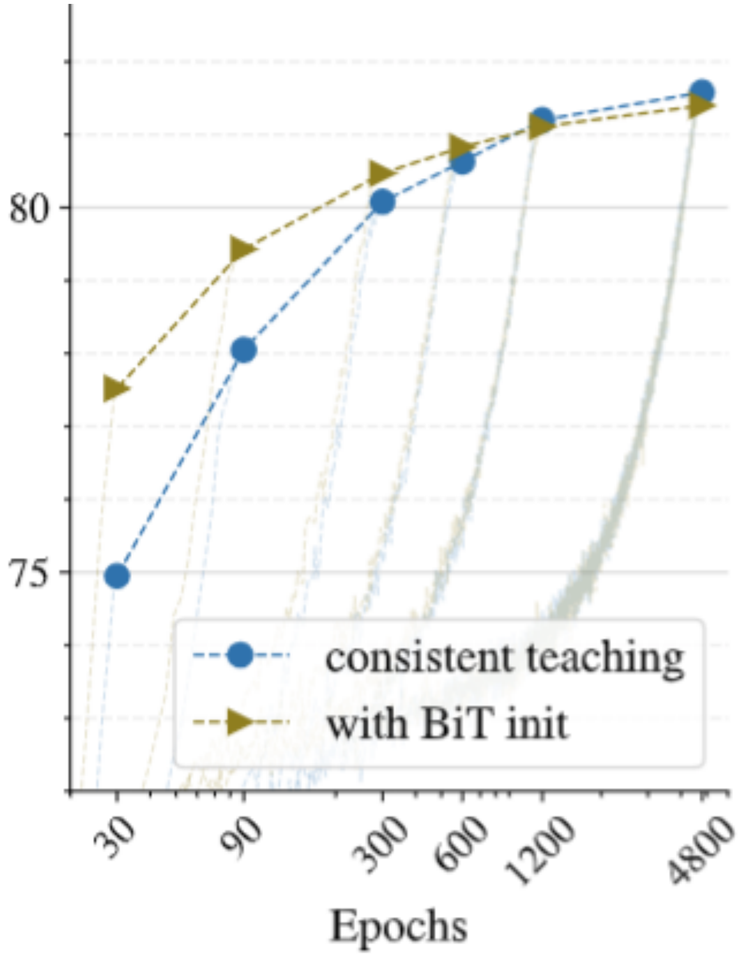
Shampooを使って学習したら高速になった



- データ拡張によって性能向上したが，その分学習時間が伸びてしまう
- Adamからより強力な最適化手法であるShampooを使ってみる.
- →学習速度が4倍になった！

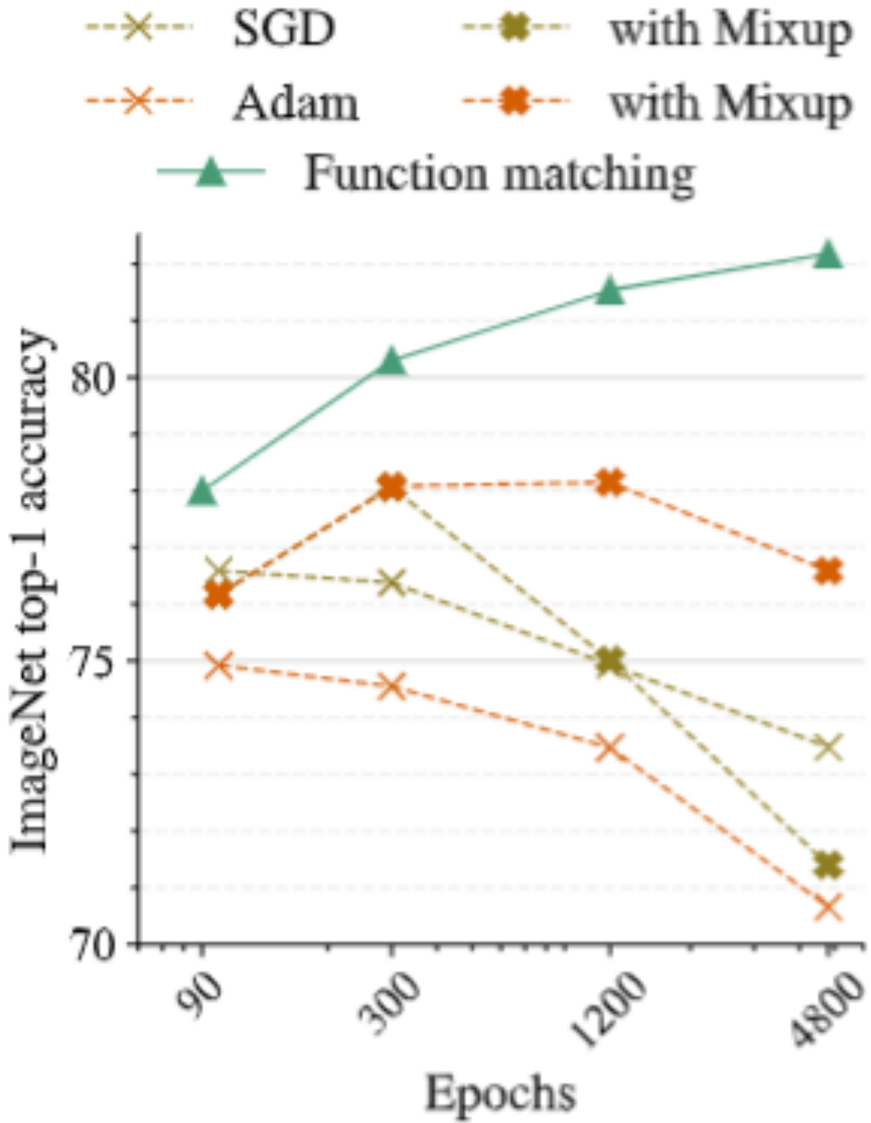
https://arxiv.org/abs/1802.09568

転移学習つかっても知識蒸留いいんじゃない?



- Patient teacherのとき，転移学習有効だったから知識蒸留+転移学習もいいのでは?
- 学習初期は転移学習よかったけど，最終的には逆転

知識蒸留はやっぱり有効



- 知識蒸留って無くてもいいってことはない？
- 知識蒸留がないと過学習を起こす

まとめと感想

- 知識蒸留で重要な要素として,
- 1. 教師と生徒の入力が同じ
- 2. データ拡張をたくさんする
- 3. 学習エポック数を増やす　　ことについて示した.

- 知識蒸留の新手法を提案したわけではないが，既存のモデルを用いて，軽量モデルがSoTAを取れるかもしれないロマンを感じれた