CONFIDENCE-AWARE MULTI-TEACHER KNOWLEDGE DISTILLATION

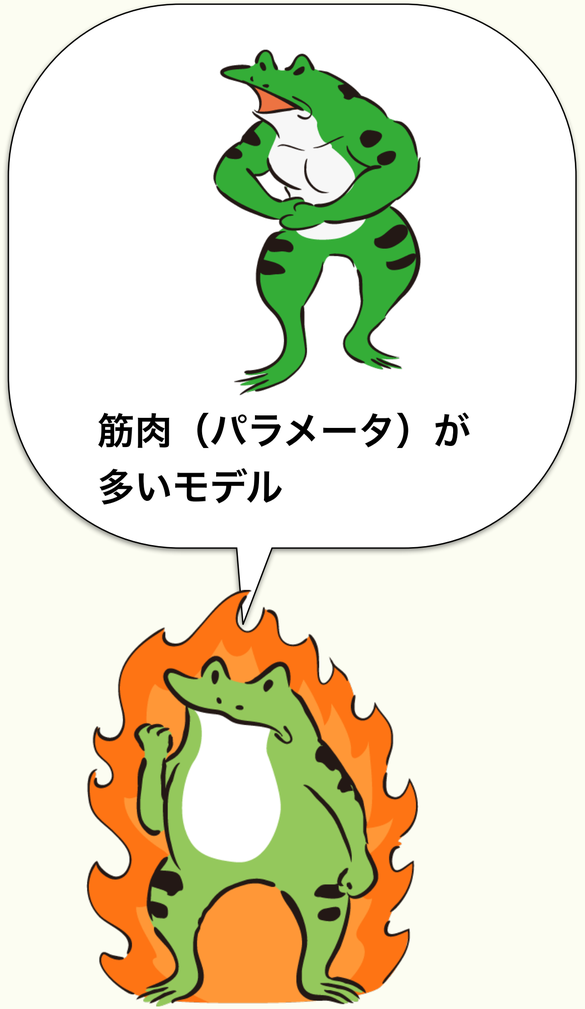
読み会@2022/05/24 守山　慧

論文情報

- 著者
  - Hailin Zhang（浙江大学）
  - Defang Chen（浙江大学）
  - Can Wang（浙江大学）
- 出典：IEEE ICASSP 2022
- 選んだ理由
  - 自分の研究に関係がありそうだったから

イントロダクション

- 巨大なネットワークは様々なタスクにおいて良いパフォーマンスを発揮している
  - これらのモデルのパラメータ数は膨大で，層の数も多くなっている
- 課題：計算時間や計算資源をたくさん必要とする
- 巨大なネットワークと同じ性能でパラメータ数を減らしたネットワークを作るために蒸留と呼ばれる手法が提案されている



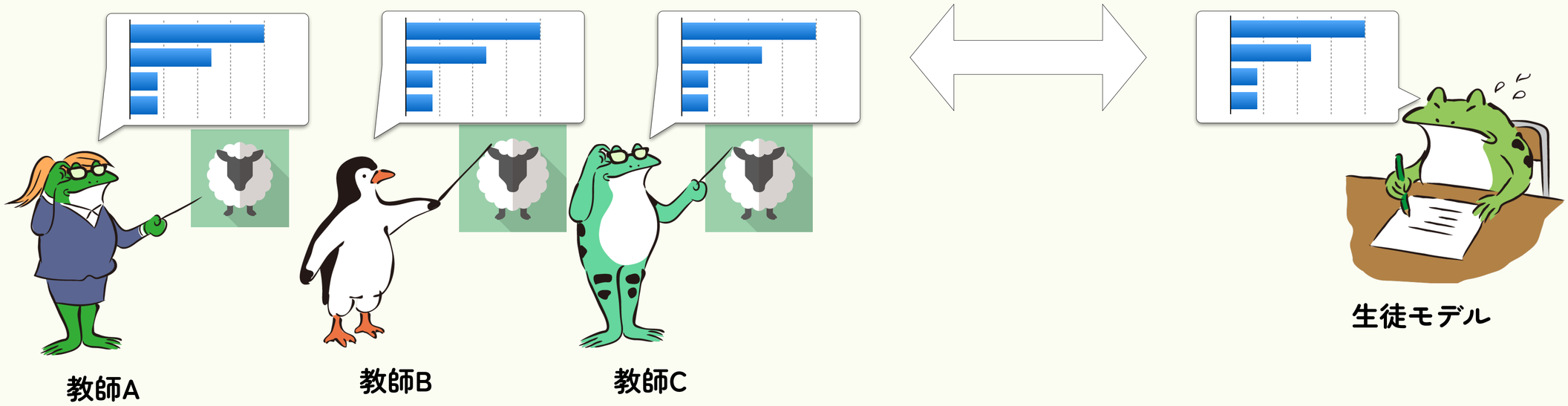
イントロダクション

- 蒸留（Knowledge Distillation，以下KD）とは
  - 生徒ネットワークの出力を教師ネットワークの出力に合わせるように学習する手法
  - 同じ入力に対して同じ確率分布を出力するような状態を目指す
- この論文では，教師ネットワークを複数用意した時の蒸留手法（MKD）を扱っている
  - MKD：Multi-Teacher Knowledge Distillation

蒸留のイメージ図

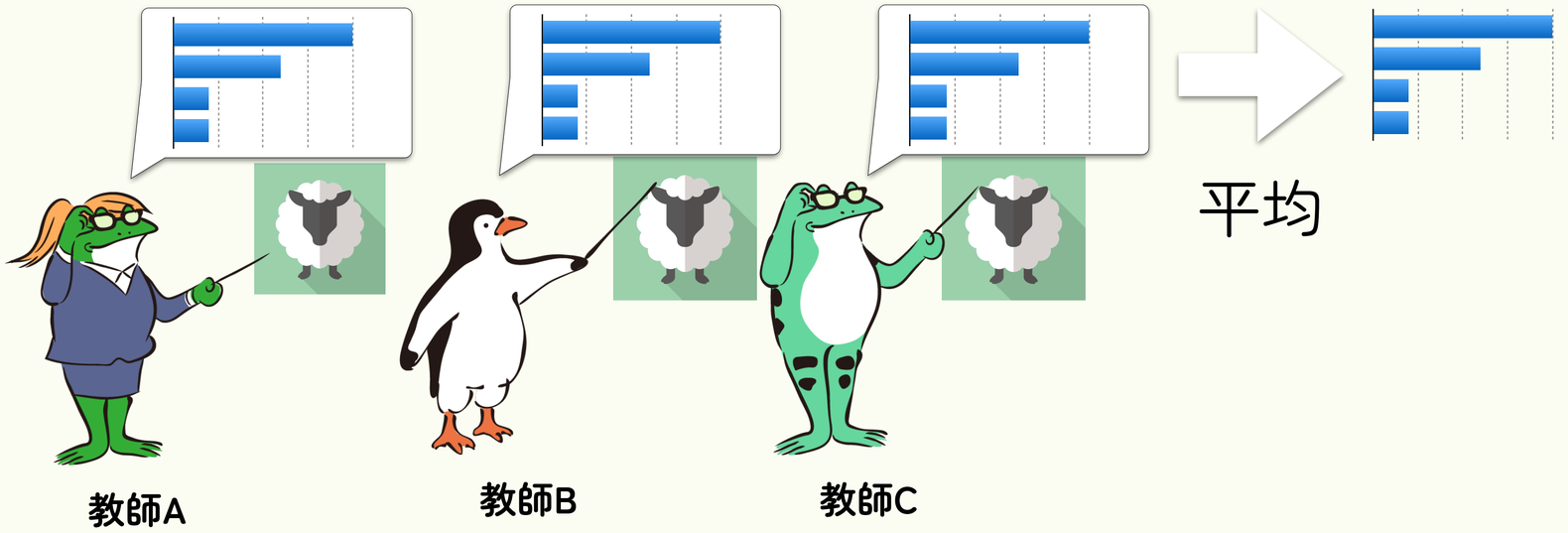
イントロダクション

- これまでのMKD
  - 複数の教師が出力した確率分布を扱うためのいくつかの手法が提案されている
  - 平均を計算する方法
  - 重み付き和を計算する方法



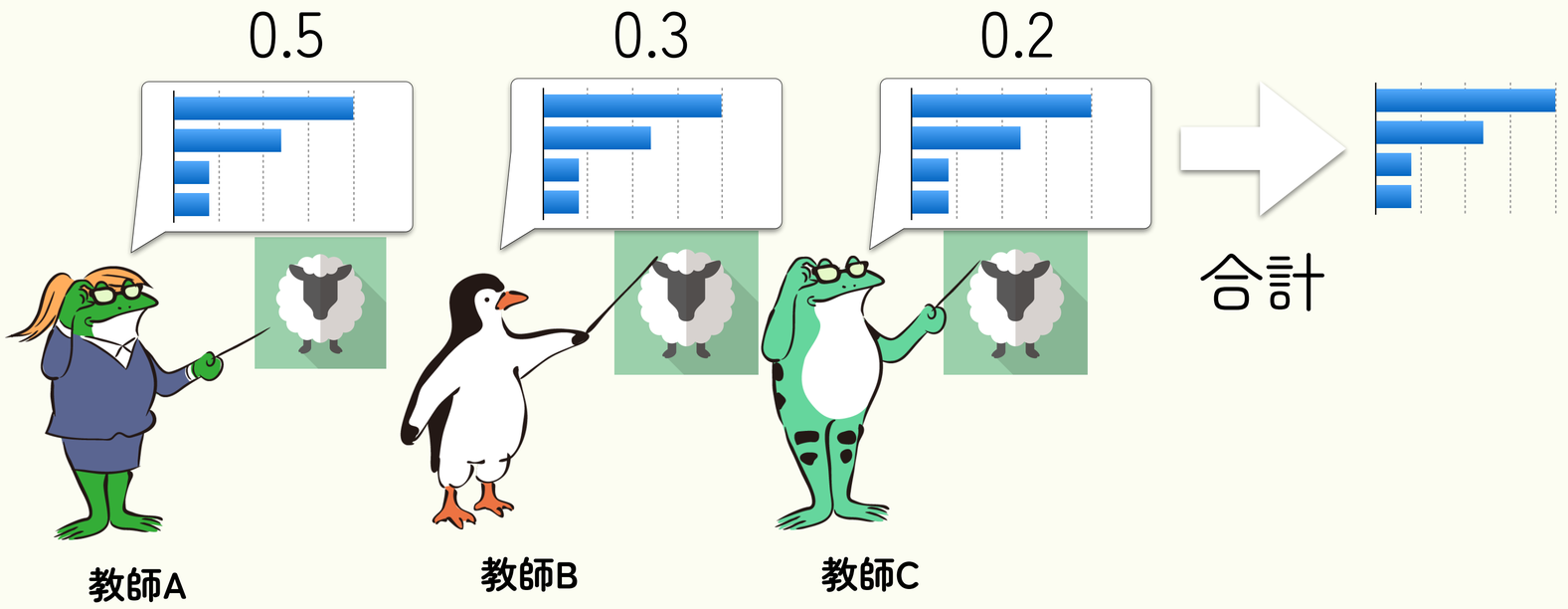
イントロダクション

- MKDの手法 1
  - 教師モデルの出力した確率分布の平均を計算し，それを使って蒸留する



イントロダクション

- MKDの手法２
  - 教師モデルが出力した確率分布の重み付き和を計算して，それを使って蒸留する
  - この時の重みは，最適化やクロスエントロピーを用いて決定される



イントロダクション

- 重み付き和を用いる手法の具体例（1,2)
  - 損失関数$\mathscr{L}(\theta)$を最小化するように学習する
  - q_iがラベルiに対する確率分布の重み付き和
  - p_iがラベルiに対する生徒モデルの出力
  - w_iにさまざまな決め方がある
    - 補完により決定する（1）
    - グリッドサーチによって決定される（2）

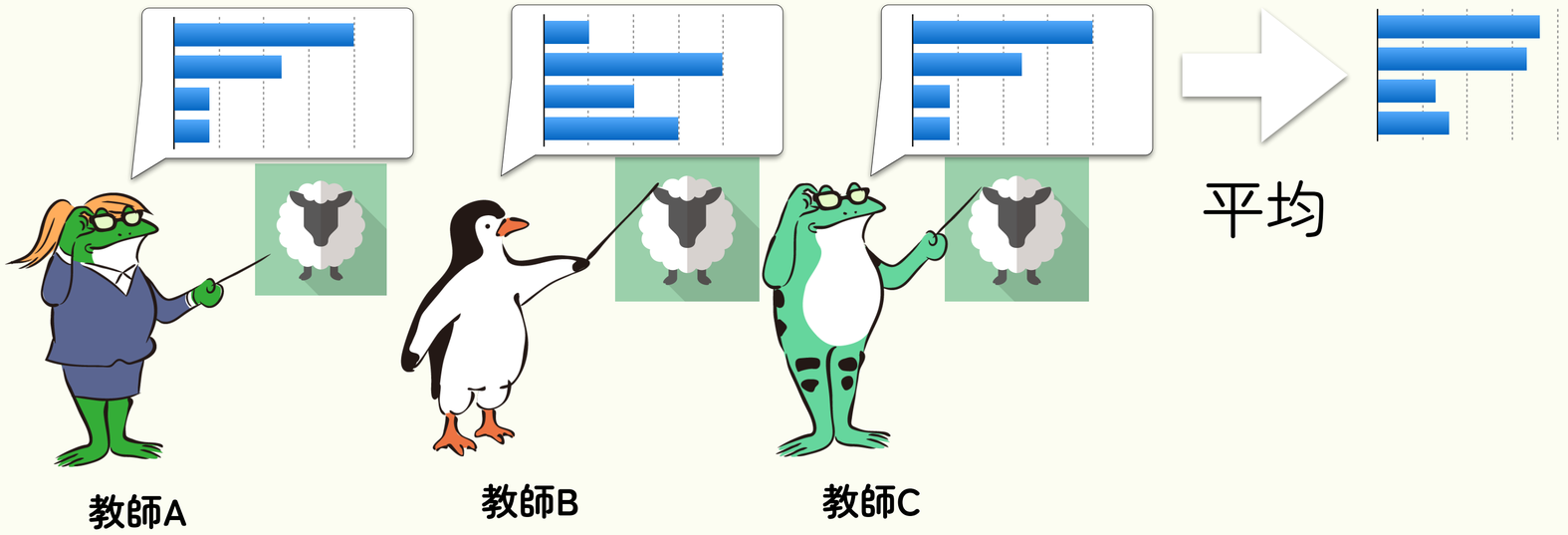
$$\mathscr{L}(\theta) = -\sum_i q_i log\left(p_i\right)$$

$$q_i = \sum_k w_k q_{i,k}$$

- (1)Fukuda et al. Efficient Knowledge Distillation from an Ensemble of Teachers interspeech 2017
- (2) Yevgen Chebotar el al. Distilling knowledge from ensembles of neural networks for speech recognition interspeech 2016

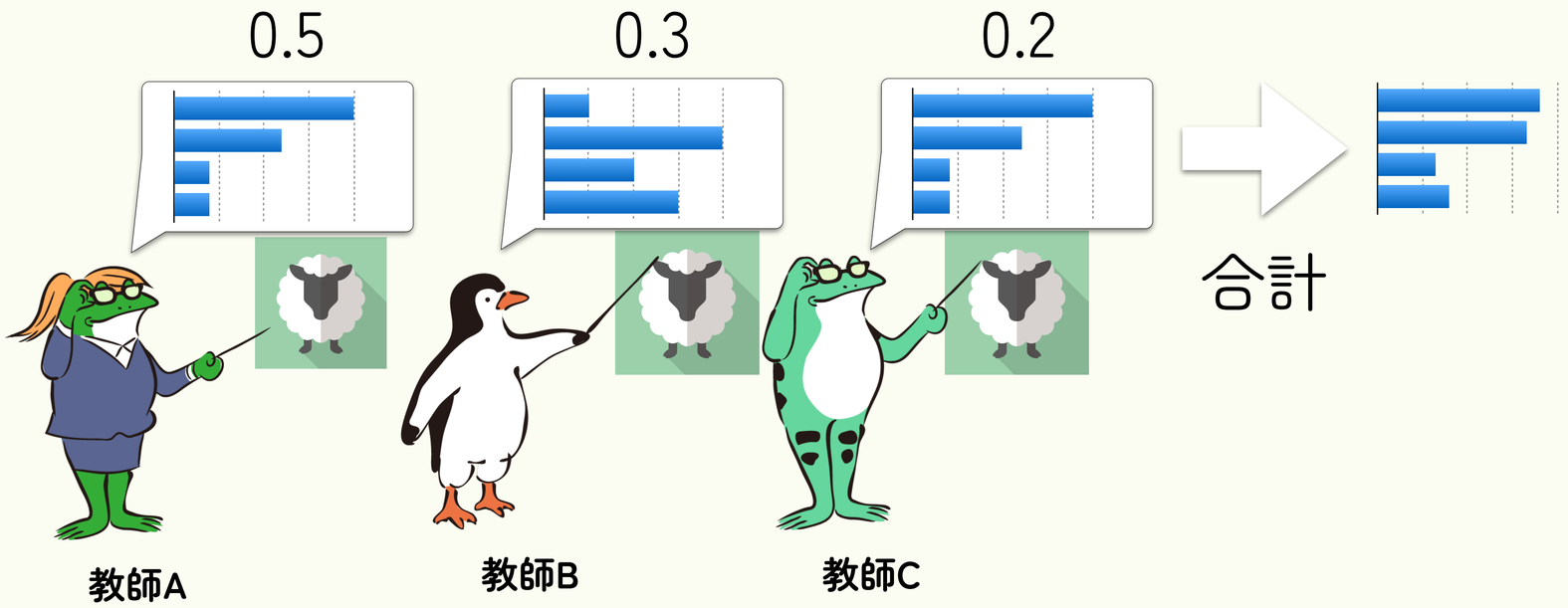
イントロダクション

- MKDの手法の課題
  - 平均を計算する場合
  - パフォーマンスの良くない教師がいるときに，影響されてしまう



イントロダクション

- MKDの手法の課題
  - 重み付き和を用いる場合
  - 良い教師モデルと悪い教師モデルの差を見分けることができない



イントロダクション

- 学習データのラベルを使って重みを適切に調整する手法を提案
  - Confidence-Aware Multi-teacher Knowledge Distillation(CA-MKD)
  - 正解ラベルを考慮した重みづけを行い，蒸留を行う
  - 教師モデルに対してのみではなく，中間表現の蒸留についてもこの重みづけは有効になる

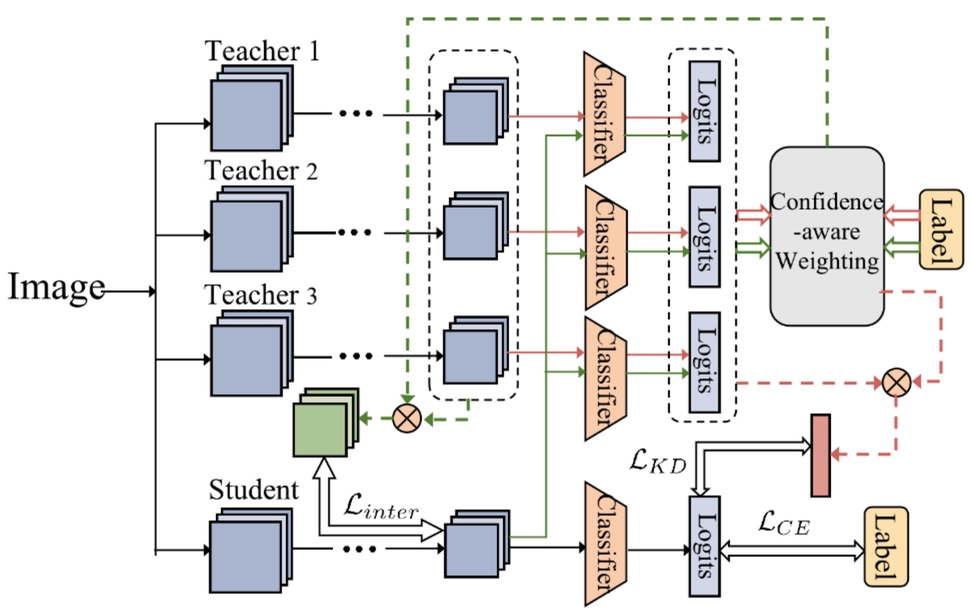


Fig. 2. An overview of our CA-MKD. The weight calculation of teacher predictions and intermediate teacher features are depicted as the red lines and green lines, respectively.

提案手法

- 3つの損失を計算する
  - 教師モデルの出力と生徒モデルの出力に対する損失：$\mathscr{L}_{KD}$
  - 教師モデルと生徒モデルの中間表現の差分の損失：$\mathscr{L}_{inter}$
  - 生徒モデルのクロスエントロピー損失：$\mathscr{L}_{CE}$

提案手法

- 教師モデルの出力と生徒モデルの出力に対する損失
- 重みの計算
  - モデルの予測結果が正解に近いほど，$\mathscr{L}^k_{CE_{KD}}$の値は0に近くなる

$$\mathscr{L}^k_{CE_{KD}} = -\sum_{c=1}^{C} y^c log\left(\sigma\left(z^c_{T_k}\right)\right)$$

C：クラス数

T_k：k番目の教師モデル

$z^c_{T_k}$：k番目の教師モデルのクラスcの出力

$\sigma(z^c)$：温度つきsoftmax関数

提案手法

- 教師モデルの出力と生徒モデルの出力に対する損失
- 重みの計算
  - 各教師モデルの予測値をもとに重みを計算
  - $\mathscr{L}^k_{CE_{KD}}$が小さいとw^k_{KD}が大きくなるように計算をしている
  - $\mathscr{L}^k_{CE_{KD}}$の値をそのまま用いると，モデルの予測があっているかどうかに関わらず，確率分布が鋭いと重みが大きくなってしまう問題がある

$$w^k_{KD} = \frac{1}{K-1}\left(1 - \frac{\exp\left(\mathscr{L}^k_{CE_{KD}}\right)}{\sum_j \exp\left(\mathscr{L}^j_{CE_{KD}}\right)}\right)$$

提案手法

- 教師モデルの出力と生徒モデルの出力に対する損失
- 生徒モデルとの損失を計算する
  - 正確な予測をしている教師モデルの予測分布に大きな重みが割り当てられている
    - 右の$\sum_{c=1}^{C}$の部分で生徒モデルと教師モデルの確率分布の類似度を計算している
    - 左の$\sum_{k=1}^{K}$で重みをつけている

$$\mathscr{L}_{KD} = -\sum_{k=1}^{K} w_{KD}^{k} \sum_{c=1}^{C} z_{T_k}^{c} log\left(\sigma\left(z_S^c\right)\right)$$

提案手法

- 教師モデルと生徒モデルの中間表現の差分の損失
  - 生徒モデルの出力を教師モデルの分類器に入力する
    - 中間地点での予測ラベルを用いる
  - この計算から損失$\mathscr{L}^k_{CE_{inter}}$を計算する

$$z_{S\to T_k} = W_{T_k} h_S$$

$$\mathscr{L}^k_{CE_{inter}} = -\sum_{c=1}^{C} y^c log\left(\sigma\left(z^c_{S\to T_k}\right)\right)$$

W_{T_K} : k番目の教師モデルの分類器

$h_S\left(= AvgPooling\left(F_S\right)\right)$: 生徒モデルの中間出力ベクトル

提案手法

- 教師モデルと生徒モデルの中間表現の差分の損失
  - 計算した損失をもとに重みを決定する
  - 生徒モデルと教師モデルの中間表現の出力の二乗和誤差を損失とする

$$w_{inter}^{k} = \frac{1}{K-1}\left(1 - \frac{\exp\left(\mathscr{L}_{CE_{inter}}^{k}\right)}{\sum_{j}\exp\left(\mathscr{L}_{CE_{inter}}^{j}\right)}\right)$$

$$\mathscr{L}_{inter} = \sum_{k=1}^{K} w_{inter}^{k} \left|\left| F_{T_k} - r\left(F_S\right) \right|\right|_2^2$$

提案手法

- 生徒モデルのクロスエントロピー損失
  - $L_{CE} = -\sum_{c=1}^{C} y^c log\left(\sigma\left(z_S^c\right)\right)$
- これらの損失を全て合計したものを生徒モデルの損失の値とする
  - $\mathscr{L} = \mathscr{L}_{CE} + \alpha\mathscr{L}_{KD} + \beta\mathscr{L}_{inter}$
  - α, βはハイパーパラメータ

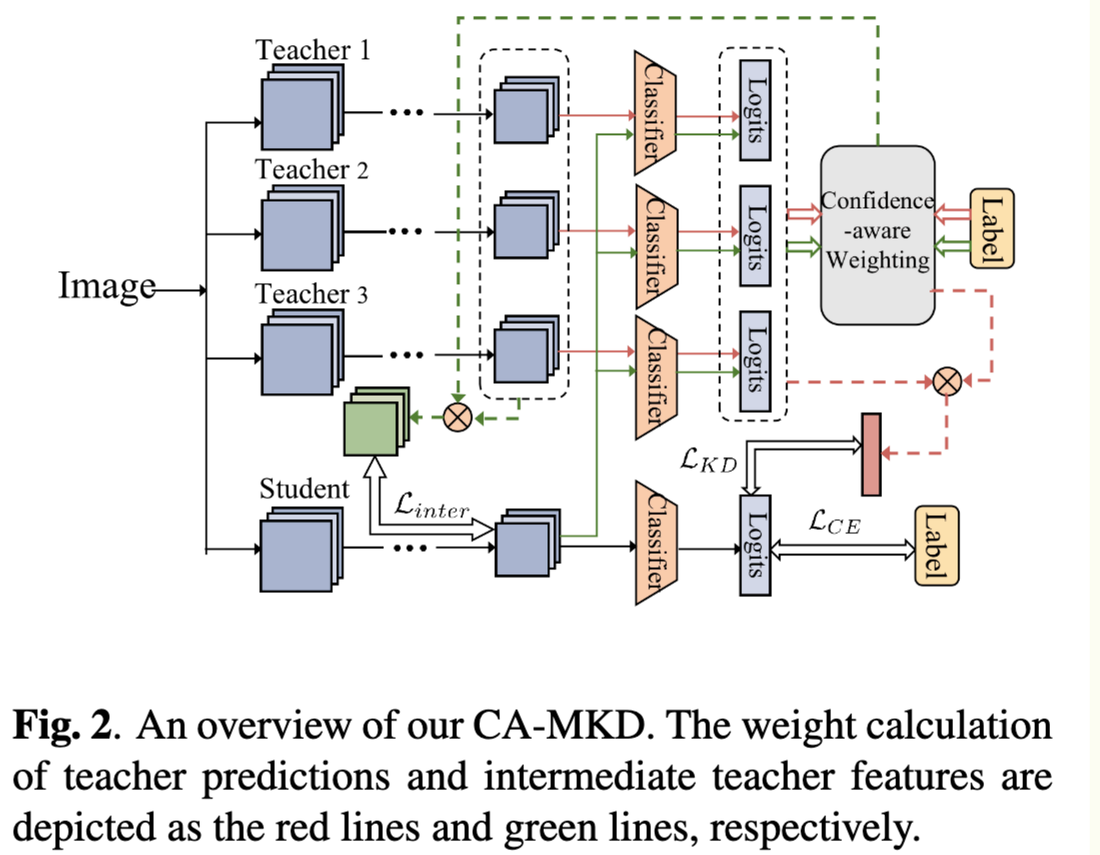


Fig. 2. An overview of our CA-MKD. The weight calculation of teacher predictions and intermediate teacher features are depicted as the red lines and green lines, respectively.

実験

- データセット：CIFAR-100
- 実験内容
  - 他のMKD手法との比較
  - KDとの比較
  - 教師モデルの数を変化させたときのパフォーマンスの比較
  - アブレーションスタディ

実験

- 他のMKD手法との比較
  - どのモデルにおいても提案手法の方がパフォーマンスが優れている

Table 1. Top-1 test accuracy of MKD methods by distilling the knowledge on multiple teachers with the same architectures.

Teacher	WRN40-2	ResNet56	VGG13	VGG13	ResNet32x4	ResNet32x4	ResNet32x4
	76.62±0.26	73.28±0.30	75.17±0.18	75.17±0.18	79.31±0.14	79.31±0.14	79.31±0.14
Ensemble	79.62	76.00	77.07	77.07	81.16	81.16	81.16
Student	ShuffleNetV1	MobileNetV2	VGG8	MobileNetV2	ResNet8x4	ShuffleNetV2	VGG8
	71.70±0.43	65.64±0.19	70.74±0.40	65.64±0.19	72.79±0.14	72.94±0.24	70.74±0.40
AVER [8]	76.30±0.25	70.21±0.10	74.07±0.23	68.91±0.35	74.99±0.24	75.87±0.19	73.26±0.39
FitNet-MKD [5]	76.59±0.17	70.69±0.56	73.97±0.22	68.48±0.07	74.86±0.21	76.09±0.13	73.27±0.19
EBKD [12]	76.61±0.14	70.91±0.22	74.10±0.27	68.24±0.82	75.59±0.15	76.41±0.12	73.60±0.22
AEKD [11]	76.34±0.24	70.47±0.15	73.78±0.03	68.39±0.50	74.75±0.28	75.95±0.20	73.11±0.27
CA-MKD	**77.94±0.31**	**71.38±0.02**	**74.30±0.16**	**69.41±0.20**	**75.90±0.13**	**77.41±0.14**	**75.26±0.32**

実験

- 他のKD手法との比較
  - 教師モデルが1つの場合の手法と比べパフォーマンスが良い
  - 複数の教師を用いることの有効性がわかる

Table 2. Top-1 test accuracy of CA-MKD compared to single-teacher knowledge distillation methods.

Teacher	WRN40-2 76.62±0.26	ResNet32x4 79.31±0.14	ResNet56 73.28±0.30
Student	ShuffleNetV1 71.70±0.19	VGG8 70.74±0.40	MobileNetV2 65.64±0.43
KD [4]	75.77±0.14	72.90±0.34	69.96±0.14
FitNet [5]	76.22±0.21	72.55±0.66	69.02±0.28
AT [6]	76.44±0.38	72.16±0.12	69.79±0.26
VID [14]	76.32±0.08	73.09±0.29	69.45±0.17
CRD [15]	76.58±0.23	73.57±0.25	71.15±0.44
CA-MKD	**77.94±0.31**	**75.26±0.13**	**71.38±0.02**

実験

- 異なる教師モデルの組み合わせによるパフォーマンス
  - 教師モデルとして用いたのはResNet8x4，ResNet20x4，ResNet32x4の３種類
  - 異なる教師モデルを用いた方がパフォーマンスがよくなる
    - モデルの構造が異なるので，知識の幅が広がっている

Table 3. Top-1 test accuracy of MKD approaches by distilling the knowledge on multiple teachers with different architectures.

VGG8	AVER	FitNet-MKD	EBKD	AEKD	CA-MKD	ResNet8x4	ResNet20x4	ResNet32x4
70.74±0.40	74.55±0.24	74.47±0.21	74.07±0.17	74.69±0.29	**75.96±0.05**	72.79	78.39	79.31

実験

- 学習時のResNet8x4，ResNet20x4，ResNet32x4の３種類の重み
  - 各点がサンプルデータを表している
- ResNet8x4のパフォーマンスは他の2つのモデルよりも良くない
- なので，ResNet8x4には大きな重みが割り当てられていない
- 提案手法の重み付けが妥当であることがわかる

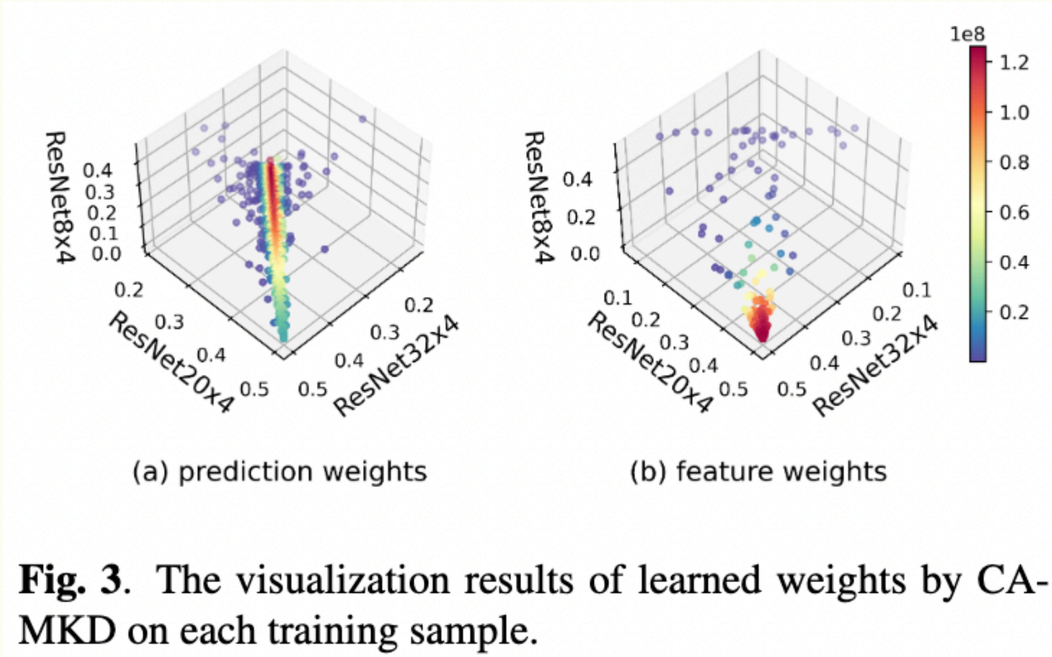


Fig. 3. The visualization results of learned weights by CA-MKD on each training sample.

実験

- 教師モデルの数を変化させたときのパフォーマンスの比較
  - 提案手法の方が他手法よりもパフォーマンスが良い
  - 他手法は教師モデルの数を増やしてもパフォーマンスの向上がない
  - 提案手法はパフォーマンスが上がっている

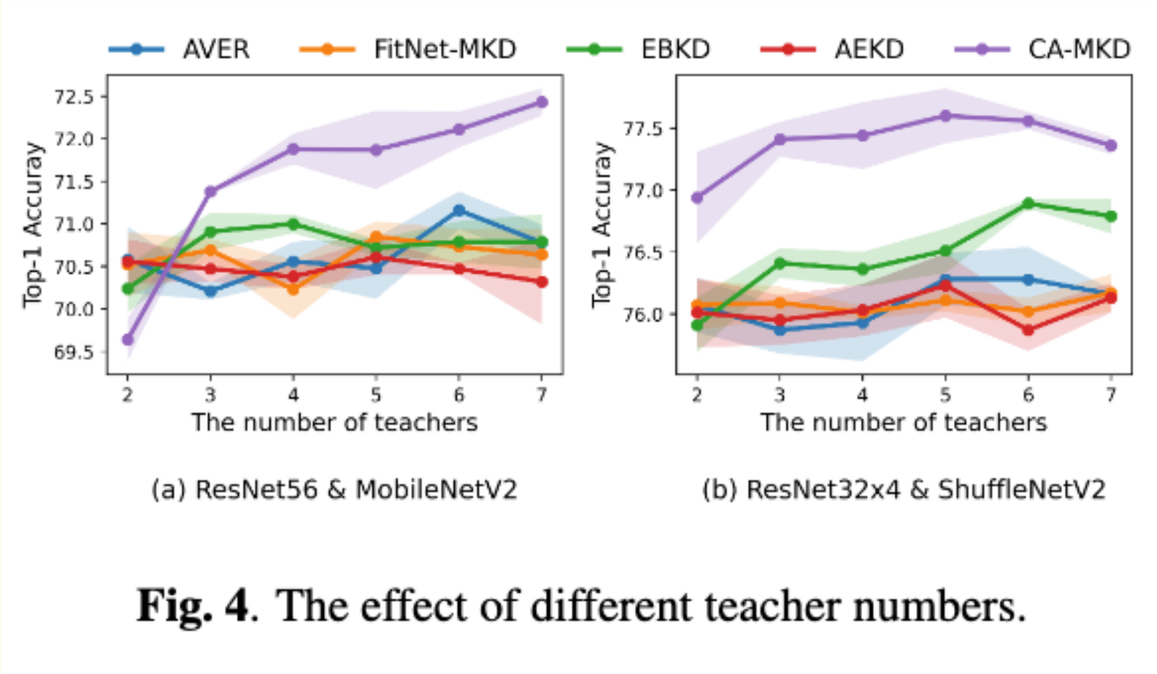


Fig. 4. The effect of different teacher numbers.

実験：アブレーションスタディ

- avg weight：教師の出力を平均して蒸留をしたモデル
- $w/o\mathcal{L}_{inter}$：中間表現の蒸留を無くした時のパフォーマンス
- w/ow^k_{inter}：中間表現の重みづけを正解ラベルに対する重みw^k_{KD}に置き換えたもの

Table 4. Ablation study with VGG13 & MobileNetV2.

avg weight	w/o $\mathcal{L}_{inter}$	w/o w^k_{inter}	CA-MKD
67.74±0.87	68.11±0.02	68.82±0.63	**69.41±0.20**

まとめ

- データの正解ラベルをもとに重みづけをする手法を提案
- モデルの出力だけではなく，中間表現に対しても同様の手法が有効であることを示した
- 既存のMKDの手法よりも良いパフォーマンスを達成した

感想

- 複数の教師モデルを用いた学習手法についてしれたのはよかった

- 自分の研究と関連性が想像できたのが嬉しかった
  - すべての参加者のパフォーマンスが悪かったり，正解ラベルが無いときは違う方法を考えないといけなさそう

- $z_{S \to T_L}$の計算過程で，生徒モデルの中間出力を教師モデルの分類器に通す気持ちがわからなかった
  - 生徒側の分類器ではダメだったのか？