Vol.2019-MUS-123 No.56
Vol.2019-SLP-127 No.56
2019/6/23

音響モデルの知識蒸留の際の正解ラベルの利用法

太刀岡 勇気[1,a)]

概要：小規模 (生徒) モデル学習時に，高精度 (教師) モデルの教師ラベルとして使う知識蒸留処理により，書き起こしを元としたハードラベルに基づく学習よりも性能が向上する．本稿では，これに加えてハードラベルを利用するため，ハードラベルの損失関数と教師ラベルの損失関数を発話ごとに確率的に選択して使う方法 (Sequence-level distillation; SD) と，それらを内挿する方法 (Sequence-level interpolation; SI) を比較し，SI の方が性能が一貫してよいことを示した．また温度パラメータとともにアニーリングを行うと，さらに性能が向上した．

1. はじめに

音声認識の性能は，音響モデルによるところが大きい．精度の高い音響モデルは，サイズが大きくなる傾向にある．計算機資源が限られている場合には，より小さいモデルが望ましい．小さい (生徒) モデルを学習する際に，高精度な大きい (教師) モデルにより得られた事後確率をソフト教師ラベルとして用いた知識蒸留処理が有効であることが知られている [1], [2]．正解書き起こしから得られたハードラベルに基づく学習と比較して，性能改善が報告されている [3]．知識蒸留の際にも，ソフトラベルに加えて，ハードラベルを付加的に使うことが有効である．本報では，付加的にハードラベルを用いた 2 手法 (Sequence-level distillation (SD) と Sequence-level interpolation (SI)) を比較する．さらに，温度パラメータを用いたアニーリングと生徒モデルの再学習の有効性を検証する．騒音下音声認識と大語彙連続音声認識のタスクにより，2 手法の性能を比較する．

2. 正解ラベルを利用した知識蒸留処理

クロスエントロピー基準の音響モデルの学習では，書き起こしに基づくハードラベル w_i が使われ，損失関数は

$$\mathcal{L}_H = -\sum_i w_i \ln s_i \tag{1}$$

となる．s_i は生徒モデルの HMM 状態 i の出力確率である．一方，知識蒸留処理では，ハードラベルではなく，教師ラベル t_i を用いた損失関数

$$\mathcal{L}_S = -\sum_i t_i \ln s_i \tag{2}$$

を使う．教師ラベル t_i は，教師モデルの最終層の出力を z_i として，

$$t_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{3}$$

のようになる．ハードラベルは正解 i に対してのみ one-hot であるが，$t_i \geq 0$ $(for^{\forall} i)$ である．

2.1 Sequence-level distillation (SD)

SD では，確率 $(1-\alpha)$ で $\mathcal{L}_S$ が，α で $\mathcal{L}_H$ が選択される．損失関数は

$$\mathcal{L}_{SD} = \sigma(\alpha - r)\mathcal{L}_H + \sigma(r - \alpha)\mathcal{L}_S \tag{4}$$

となる．σ は階段関数，r は 0 から 1 の一様乱数である．ここでは，それを発話ごとに切り替えることとした．

2.2 Sequence-level interpolation (SI)

SI は選択ではなく，α によりソフトラベルとハードラベルを内挿する．

$$\mathcal{L}_{SI} = -\sum_i [\alpha w_i + (1-\alpha)t_i] \ln s_i \tag{5}$$

2.3 温度パラメータ

式 (3) に対して，温度パラメータ T を導入することで，t_i の分布形状を変化させる．

$$t_{i,T} = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \tag{6}$$

のようにしてラベルを作る．これにより，$t_{i,1} = t_i$ で，$t_{i,T>1}$ の分布が平滑に，$t_{i,T<1}$ の分布が急峻になる．

[1] デンソーアイティーラボラトリ
東京都渋谷区渋谷 2-15-1 渋谷クロスタワー 28F 150-0002
[a)] ytachioka@d-itlab.co.jp

Vol.2019-MUS-123 No.56
Vol.2019-SLP-127 No.56
2019/6/23

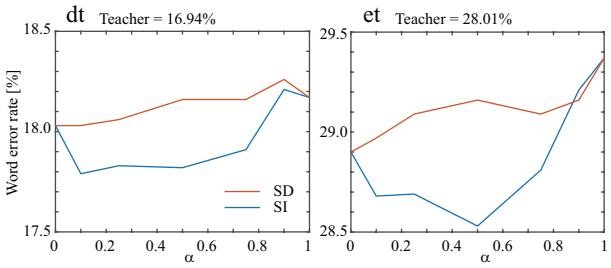

図 1　WER[%] against α of SD and SI.

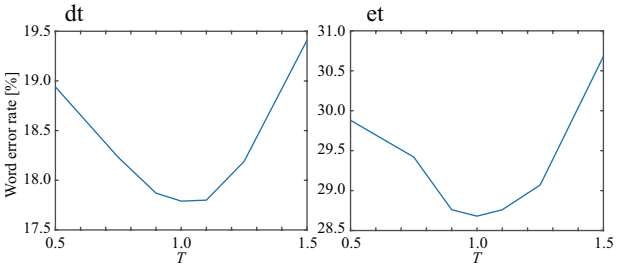

図 2　WER[%] against temperature (SI and $\alpha = 0.1$).

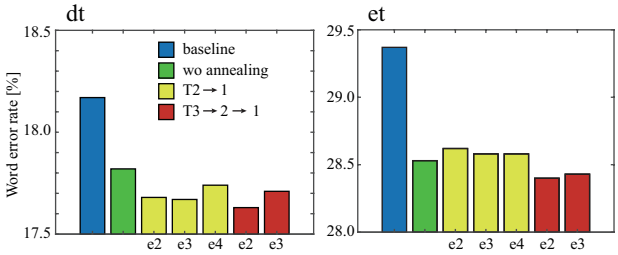

図 3　Effect of annealing (SI and $\alpha = 0.5$).

3. 実験 (騒音下音声認識)

3.1 実験設定

語彙数 5,000 の騒音下音声認識実験により，提案法の有効性を検証した．4 種の騒音環境で，開発セット (dt)，評価セット (et) に対する平均単語誤り率 (WER) で評価した．音響特徴量は，13 次元 MFCC とその動的特徴量に，線形判別分析を加えたのち，特徴量空間最尤線形回帰を施した 40 次元である．

入・出力次元は 440(前後 5 フレームを連結)，1987 次元とした．教師モデルは 30M パラメータ (2048[ノード/層]×7[層]) である．生徒モデルは 2M パラメータ (512[ノード/層]×4[層]) である．教師ラベルは $t_i < 0.01$ となる t_i は 0 とし，$\sum_i t_i = 1$ となるように再度正規化した．

3.2 結果と考察

図 1 は，教師モデルの WER と，SD/SI の各 α における生徒モデルの WER を示す．$\alpha = 1$ が生徒モデルの原性能，$\alpha = 0$ が知識蒸留の場合である．SD/SI ともに知識蒸留は有効だが，SD は $\alpha = 0$ が最良で内挿により性能が低下した．これに対し，SI は適当な α で性能が向上した．図 2 には，温度パラメータ T と WER の関係を示す．T を 1 以外にすると WER が悪化することが分かった．これに対して，図 3 のように，$T = 2$ を $e(= 2, 3, 4)$ エポック分それを繰り返して $T = 1$ とする $T \to 1$，もしくは，e エポックずつ $T = 3 \to 2 \to 1$ と下げていくと，WER が改善することが分かった．また，生徒モデルが学習済の場合，教師ラベルで再学習することで学習時間を短縮できる．再学習でも，同様の設定で 17.77%(dt), 28.42%(et) が得られ，再学習で十分なことが分かった．

4. 実験 (大語彙連続音声認識)

4.1 実験設定

日本語話し言葉コーパス (CSJ) [4] は，日本語音声認識システムを構築するために最もよく用いられてるタスクである．語彙サイズは約 7 万である．テストセットは 3 種類あり，それぞれ 10 話者による講演から構成されている．テストセット E1，E2，E3 は 22,682，23,226，14,896 単語からなる．

表 1　WER[%] on CSJ. (SI type)

	α	E1	E2	E3
teacher	-	11.71	9.12	12.74
student	-	13.73	10.15	14.45
scratch	0	13.42	10.03	14.30
retrain	0	13.39	10.04	14.40
retrain	0.1	13.41	10.06	14.42
retrain	0.25	13.50	**10.02**	14.37
retrain	0.5	**13.38**	10.09	**14.26**

上述の 40 次元の音響特徴量に対して，クロスエントロピー基準で DNN-HMM を学習した．Kaldi [5] CSJ レシピにより，DNN 音響モデルを得た．DNN の入力次元は 1,400 である (前後 17 フレームを連結)．出力次元は 9,388 である．教師モデルは，38.7M パラメータである $(1,905$[ノード/層]$\times 6$[層]$)$．生徒モデルは，6.3M パラメータである (512[ノード/層]×4[層])．

4.2 結果と考察

表 1 は，CSJ タスクによる WER を示す．前の実験によると，SI は SD を一貫して上回ったので，この節では SI について検証する．再学習とスクラッチからの学習はいずれも性能を向上させ，両者はほぼ同等の性能であった．α による内挿により性能が向上し，$\alpha = 0.5$ の場合に最良の性能を示した．傾向は前節での実験と同様である．

5. まとめ

知識蒸留において，SD と SI の 2 手法を検証し，SI の方が性能が一貫してよいことを示した．また騒音下音声認識，大語彙連続音声認識の両タスクで有効性を確認した．

温度パラメータとともにアニーリングを行うと，さらに性能が向上した．

参考文献

[1] Hinton, G. E., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *Proceedings of NIPS* (2014).
[2] Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. and Ramabhadran, B.: Efficient Knowledge Distillation from an Ensemble of Teachers, *Proceedings of INTERSPEECH*, pp. 3697–3701 (2017).
[3] Kim, Y. and Rush, A. M.: Sequence-level Knowledge Distillation, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1317–1327 (2016).
[4] Furui, S., Maekawa, K. and Isahara, H.: A Japanese national project on spontaneous speech corpus and processing technology, *Proceedings of ASR*, pp. 244–248 (2000).
[5] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Petr, M., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K.: The Kaldi Speech Recognition Toolkit, *Proceedings of ASRU*, pp. 1–4 (2011).