音響モデルの知識蒸留における正解ラベルと温度パラメータの利用

Knowledge Distillation of Acoustic Models with Reference Labels and Temperature Parameter

太刀岡勇気　Yuuki Tachioka

デンソーアイティーラボラトリ　Denso IT laboratory

1　はじめに

小規模 (生徒) モデル学習時に，高精度 (教師) モデルのソフトラベルを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 [1]．本稿では，ハードラベルも付加的に使う方法について，2 種のハードラベルの利用法と温度パラメータについて検討した．

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 $\mathcal{L}_H = -\sum_i w_i \ln s_i$ 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 $\mathcal{L}_S = -\sum_i t_i \ln s_i$ を使う．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．

Sequence-level distillation (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

$$\mathcal{L}_{SD} = \sigma(\alpha - r)\mathcal{L}_H + \sigma(r - \alpha)\mathcal{L}_S \qquad (1)$$

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．

Sequence-level interpolation (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．

$$\mathcal{L}_{SI} = -\sum_i [\alpha w_i + (1-\alpha)t_i] \ln s_i \qquad (2)$$

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．教師モデルの最終層の出力を z_i として，$t_{i,T} = \exp(z_i/T)/\sum_j \exp(z_j/T)$ 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．

3　実験 (第 4 回 CHiME チャレンジ 1ch トラック)

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，特徴量空間最尤線形回帰を施したものである．

入・出力次元は 440，1987 とした．教師モデルは 30M ラメータ (2048 × 7 層) である．生徒モデルは 2M パラメータ (512 × 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．

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．

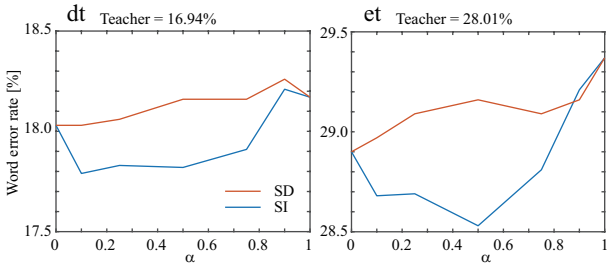

図 1　WER[%] against α of SD and SI.

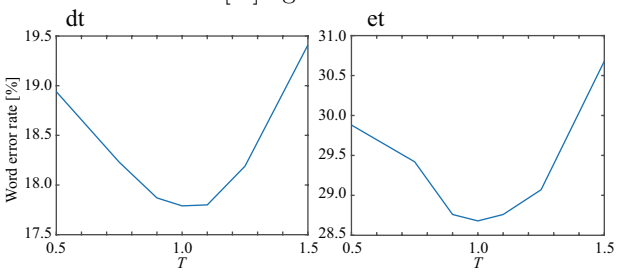

図 2　WER[%] against temperature (SI and $\alpha = 0.1$).

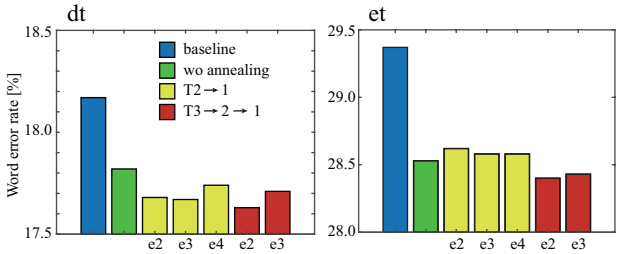

図 3　Effect of annealing (SI and $\alpha = 0.5$).

4　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．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] Y. Kim and A.M. Rush, “Sequence-level Knowledge Distillation”, Proc. of EMNLP, 1317–1327, 2016