Improving Group Fairness in Knowledge Distillation via Laplace Approximation of Early Exits

Edvin 24V0074 Sagar 24D0367

CS 769 Optimization in Machine Learning

2 May 2025

Overview

1. Recap from Seminar

2. Experiments And Results

3. Analysis And Future Work

Section Overview

1. Recap from Seminar

2. Experiments And Results

3. Analysis And Future Work

 Knowledge Distillation as an effective way to distill knowledge from teacher to student

- Knowledge Distillation as an effective way to distill knowledge from teacher to student
- Teacher Provides "Soft Targets"

- Knowledge Distillation as an effective way to distill knowledge from teacher to student
- Teacher Provides "Soft Targets"
- Loss: kl divergence + cross-entropy

- Knowledge Distillation as an effective way to distill knowledge from teacher to student
- Teacher Provides "Soft Targets"
- Loss: kl divergence + cross-entropy
- Student model relies on spurious correlations

- Knowledge Distillation as an effective way to distill knowledge from teacher to student
- Teacher Provides "Soft Targets"
- Loss: kl divergence + cross-entropy
- Student model relies on spurious correlations
- Student's early Layers overconfident on hard instances

- Knowledge Distillation as an effective way to distill knowledge from teacher to student
- Teacher Provides "Soft Targets"
- Loss: kl divergence + cross-entropy
- Student model relies on spurious correlations
- Student's early Layers overconfident on hard instances
- DEDIER loss

$$\mathcal{L}_{ extit{student}} = \sum_{D_{ extit{w}}} (1 - \lambda) \cdot \emph{l}_{ extit{ce}} + \lambda \cdot extstyle{ ext{wt}} \cdot \emph{l}_{ extit{kd}}$$

where $\mathbf{wt} = \exp^{\beta.\mathbf{cm}.\alpha}$ and $\mathbf{cm}(\mathbf{p}) = \mathbf{p_{max}} - \max_{\mathbf{p_k} \in \mathbf{p} - \mathbf{p_{max}}} \mathbf{p_k}$

• Two alternate approaches for estimating uncertainity in prediction in early exit layers.

- Two alternate approaches for estimating uncertainity in prediction in early exit layers.
- [Meronen et al., 2023] used Laplace approximation for bayesian posterior at exit layer.

- Two alternate approaches for estimating uncertainity in prediction in early exit layers.
- [Meronen et al., 2023] used Laplace approximation for bayesian posterior at exit layer.
- [Jazbec et al., 2024] used AVCS based on Predictive-likelihood ratio to get confidence intervals for predictions.

- Two alternate approaches for estimating uncertainity in prediction in early exit layers.
- [Meronen et al., 2023] used Laplace approximation for bayesian posterior at exit layer.
- [Jazbec et al., 2024] used AVCS based on Predictive-likelihood ratio to get confidence intervals for predictions.
- Experiment: Laplace Approximation based uncertainity estimate to reweight both the losses.

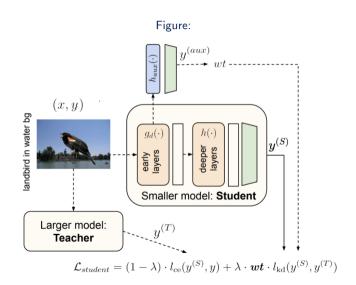


Figure:



(blond, male)







(landbird, land bg) (waterbird, land bg)





S1: oh uh-huh well no they wouldn't would they no S2: No. they wouldn't go there. Group: (contradiction X, has negation words ✔)

leaving all her money to Miss Howard? I asked in a low voice, with some curiosity. S2: I yelled at the top of my lungs. Group: (contradiction . has negation words

S1: Do you think Mrs. Inglethorp made a will

S1: so i have to find a way to supplement that S2: I need a way to add something extra. Group: (contradiction X, has negation words X) Sentence: You sound like a terrorist Group: (Toxic ✓, mention of identity X)

Sentence: She hates men because that's what her mother taught her Group: (Toxic ✓, mention of identity ✓)

Sentence: I doubt that anyone cares whether you believe it or not Group: (Toxic X, mention of identity X)

(non-blond, male) (non-blond, female) (landbird, water bg) (waterbird, water bg)

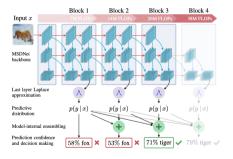
CelebA

Waterbirds

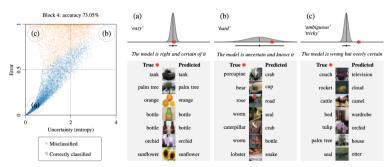
MultiNLI

CivilComments-WILDS

Figure:







• Bayesian treatment of parameters

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathsf{train}}) = \frac{p(\mathcal{D}_{\mathsf{train}} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}_{\mathsf{train}}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}} = \frac{[\mathsf{likelihood}] \times [\mathsf{prior}]}{[\mathsf{model evidence}]}$$

• Bayesian treatment of parameters

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathsf{train}}) = \frac{p(\mathcal{D}_{\mathsf{train}} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}_{\mathsf{train}}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}} = \frac{[\mathsf{likelihood}] \times [\mathsf{prior}]}{[\mathsf{model evidence}]}$$

MAP estimate can be found by maximising the unnormalised posterior:

$$\hat{m{ heta}} = rg \max_{m{ heta}} \log p(\mathcal{D}_{\mathsf{train}} \mid m{ heta}) + \log p(m{ heta})$$

• Bayesian treatment of parameters

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathsf{train}}) = \frac{p(\mathcal{D}_{\mathsf{train}} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}_{\mathsf{train}}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}} = \frac{[\mathsf{likelihood}] \times [\mathsf{prior}]}{[\mathsf{model evidence}]}$$

MAP estimate can be found by maximising the unnormalised posterior:

$$\hat{m{ heta}} = rg \max_{m{ heta}} \log p(\mathcal{D}_{\mathsf{train}} \mid m{ heta}) + \log p(m{ heta})$$

Gaussian distribution via laplace approximation

$$p(\hat{\mathbf{z}}_i \mid \mathbf{x}_i) = \mathcal{N}(\hat{\mathbf{W}}_{MAP}^{\top} \hat{\boldsymbol{\phi}}_i, (\hat{\boldsymbol{\phi}}_i^{\top} \mathbf{V} \hat{\boldsymbol{\phi}}_i) \mathbf{U})$$

 $\mathbf{V}^{-1} \otimes \mathbf{U}^{-1} = \mathbf{H}^{-1}$

• Bayesian treatment of parameters

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathsf{train}}) = \frac{p(\mathcal{D}_{\mathsf{train}} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}_{\mathsf{train}}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}} = \frac{[\mathsf{likelihood}] \times [\mathsf{prior}]}{[\mathsf{model evidence}]}$$

MAP estimate can be found by maximising the unnormalised posterior:

$$\hat{oldsymbol{ heta}} = rg \max_{oldsymbol{ heta}} \log p(\mathcal{D}_{\mathsf{train}} \mid oldsymbol{ heta}) + \log p(oldsymbol{ heta})$$

Gaussian distribution via laplace approximation

$$p(\hat{\mathbf{z}}_i \mid \mathbf{x}_i) = \mathcal{N}(\hat{\mathbf{W}}_{\mathsf{MAP}}^{\top} \hat{\boldsymbol{\phi}}_i, (\hat{\boldsymbol{\phi}}_i^{\top} \mathbf{V} \hat{\boldsymbol{\phi}}_i) \mathbf{U})$$

 $\mathbf{V}^{-1} \otimes \mathbf{U}^{-1} = \mathbf{H}^{-1}$

Samples

$$\hat{\mathbf{z}}_i^{(I)} = \hat{\mathbf{W}}_{\mathsf{MAP}}^{\top} \hat{\boldsymbol{\phi}}_i + (\hat{\boldsymbol{\phi}}_i^{\top} \mathbf{V} \hat{\boldsymbol{\phi}}_i)^{\frac{1}{2}} (\mathbf{L} \mathbf{g}^{(I)})$$

 $\mathbf{g}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and \mathbf{L} is the Cholesky factor of \mathbf{U}



Section Overview

1. Recap from Seminar

2. Experiments And Results

3. Analysis And Future Work

• The MultiNLI dataset [Williams et al., 2018] was used

Figure:



(blond, male)







(landbird, land bg) (waterbird, land bg)







S1: so i have to find a way to supplement that S2: I need a way to add something extra. Group: (contradiction X, has negation words X)

S1: oh uh-huh well no they wouldn't would they no S2: No. they wouldn't go there. Group: (contradiction X, has negation words ✔)

S1: Do you think Mrs. Inglethorp made a will leaving all her money to Miss Howard? I asked in a low voice, with some curiosity. S2: I yelled at the top of my lungs. Group: (contradiction ✓, has negation words

Sentence: You sound like a terrorist Group: (Toxic ✓, mention of identity X)

Sentence: She hates men because that's what her mother taught her Group: (Toxic ✓, mention of identity ✓)

Sentence: I doubt that anyone cares whether you believe it or not Group: (Toxic X, mention of identity X)

(non-blond, male) (non-blond, female) (landbird, water bg) (waterbird, water bg)

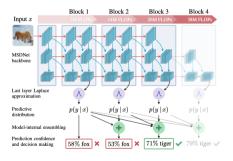
CelebA

Waterbirds

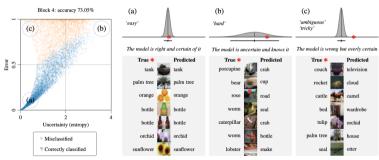
MultiNLI

CivilComments-WILDS

Figure:







Section Overview

1. Recap from Seminal

2. Experiments And Results

3. Analysis And Future Work

References



Meronen, L., Trapp, M., Pilzer, A., Yang, L., and Solin, A. (2023). Fixing Overconfidence in Dynamic Neural Networks.

Version Number: 4.

Williams, A., Nangia, N., and Bowman, S. R. (2018).

A broad-coverage challenge corpus for sentence understanding through inference.

In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.