# ML Calibration

2023. 04. 22 @ Deepest Season13 Hosting

Speaker: 박승원 (https://swpark.me)

# About me

- ML Engineer @ Moloco (22.09 – now)
- Majored Physics & CSE @ SNU (17.03 – 22.08)
- Deepest (Season 5 – 9, 12 – now)

- Research Intern @ MARG & Supertone (21.10 – 22.07)
- Symbiote (21.03 – 07)
- AI Scientist @ maum.ai ~~MINDs Lab~~ (18.06 – 21.02)

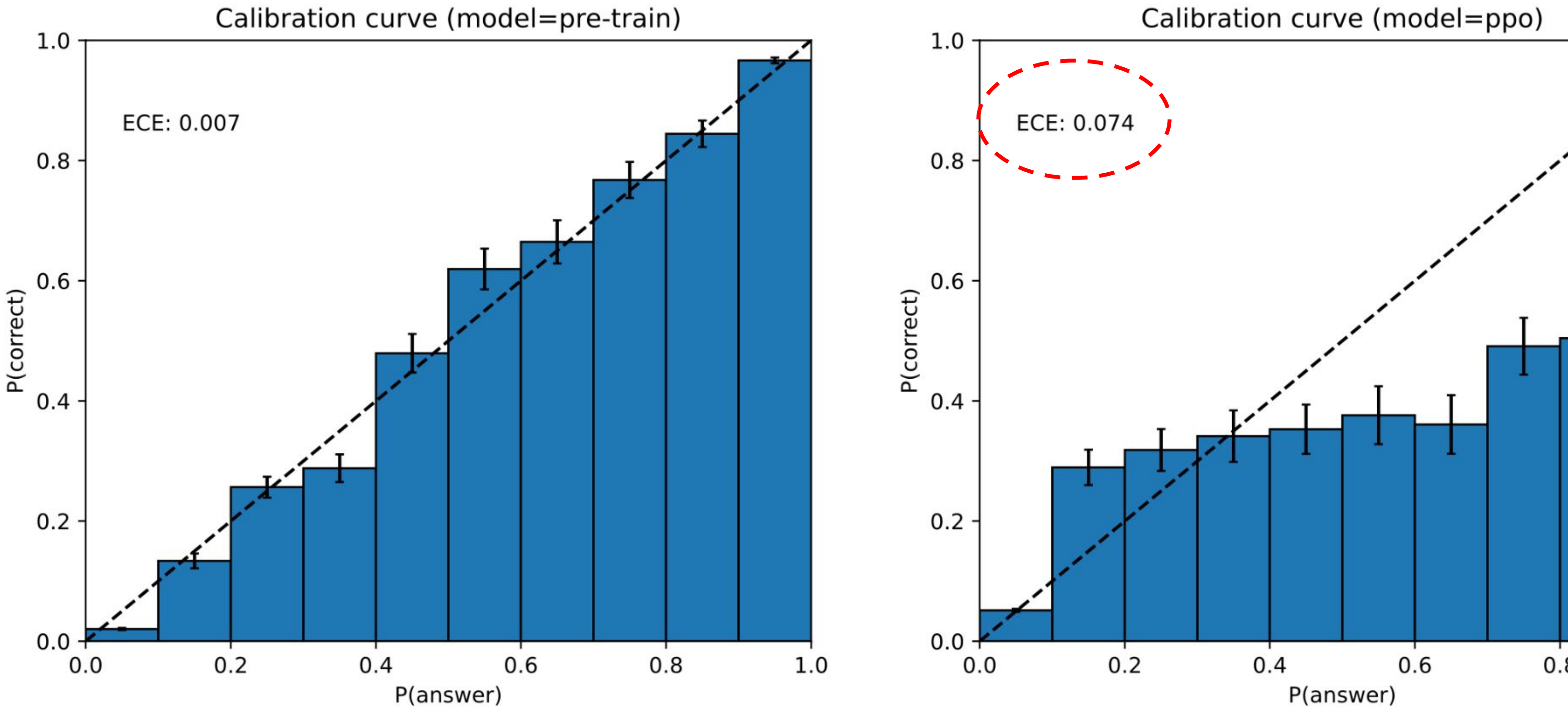



**Figure 8.** Left: Calibration plot of the pre-trained GPT-4 model on a subset of the MMLU dataset. On the x-axis are bins according to the model's confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The dotted diagonal line represents perfect calibration. Right: Calibration plot of the post-trained GPT-4 model on the same subset of MMLU. The post-training hurts calibration significantly.

# Let’s think about confidence of model prediction

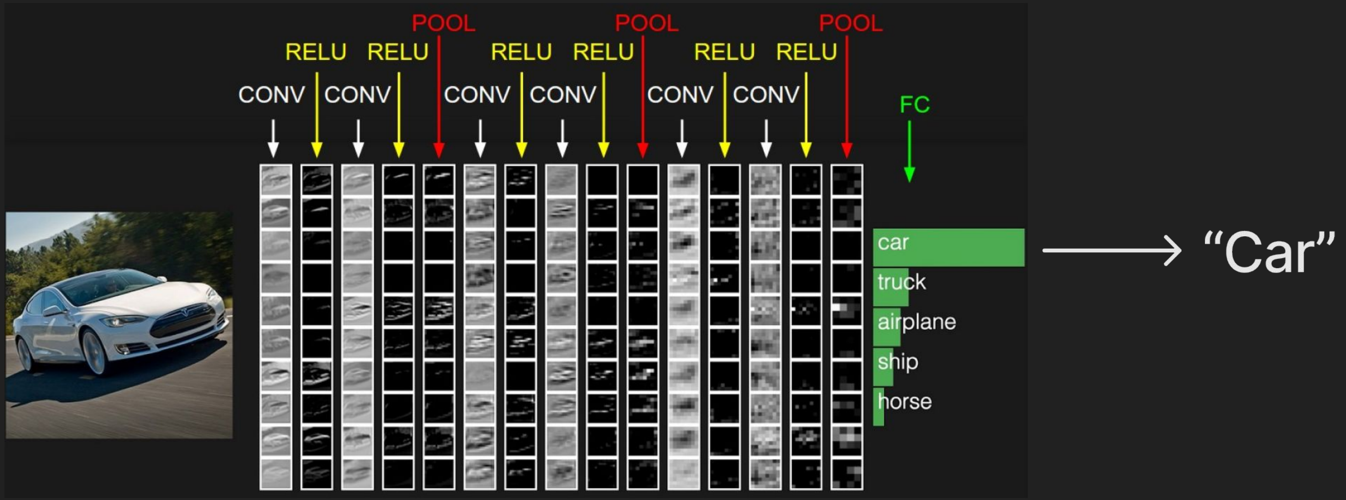
강수확률 p=30%? *“argmax(1-0.3, 0.3) = 0 이니까 비 안 오겠지…”*

그런데, 정말 확률이 30%일까?
비슷하게 예보됐던 날 100개를 모아보면, 정말 100일 중 30일에 비가 왔을까?

| 지역 | 22일(토) | | 23일(일) | | 24일(월) | | 25일(화) | | 26일(수) | | 27일(목) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 오전 | 오후 | 오전 | 오후 | 오전 | 오후 | 오전 | 오후 | 오전 | 오후 | |
| 서울 인천 경기도 | 30% | 30% | 30% | 30% | 30% | 40% | 80% | 40% | 20% | 10% | 10% |
| 강원도 영서 | 20% | 20% | 20% | 20% | 30% | 40% | 80% | 40% | 20% | 20% | 20% |
| 강원도 영동 | 20% | 20% | 30% | 20% | 30% | 30% | 80% | 40% | 40% | 20% | 10% |

# We mostly use only *confidence.argmax()*

Only the ordering of the scores contributes to the final prediction & evaluation

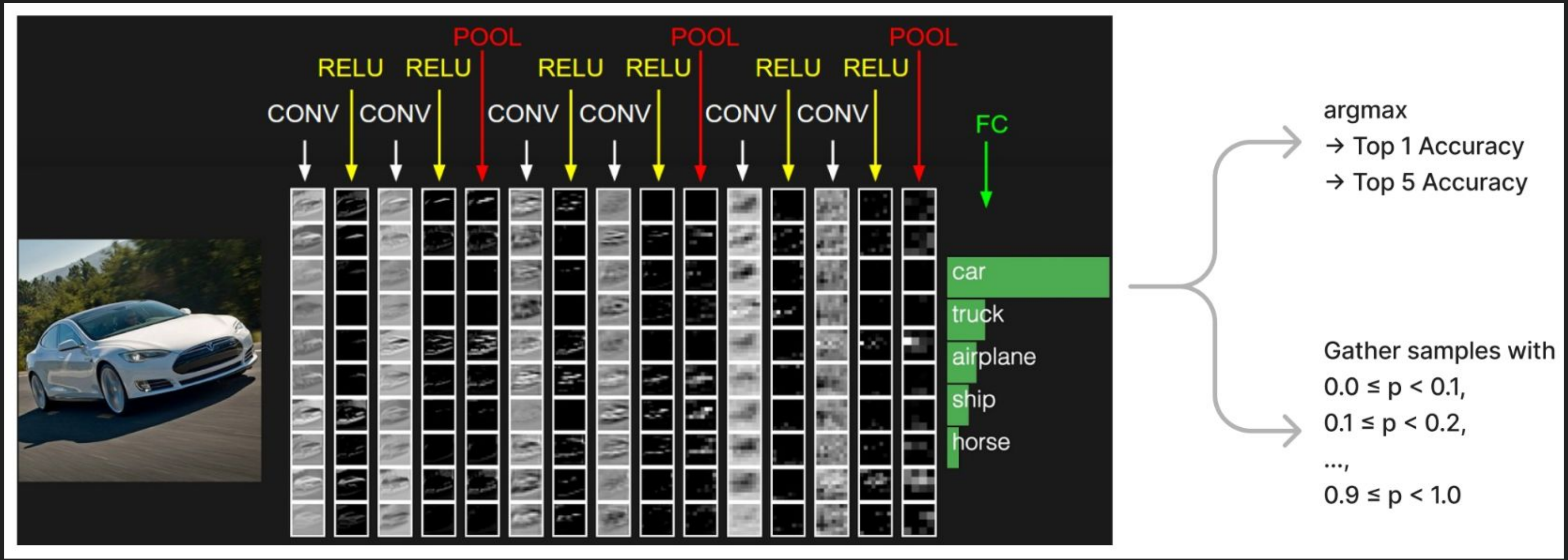


# But, truthful confidence also matters

- Cost-sensitive classification
  - Moloco: uses expectation(E) value to form optimal bidding price for ad auction
  - Insurance company: also uses E.
- Any situation where uncertainty matters – to be cautious when $p < threshold$
  - Healthcare: to reject low-quality/OOD inputs.
  - ChatGPT: when to say “Sorry, as an AI language model, I can’t …”
  - Self-driving cars
  - …

# Defining “calibration”

# How to evaluate Confidence



# Defining “calibration”

Calibration Error = |Confidence – Frequency|

Example)

Suppose that we have 100 samples with precipitation p=30%

- actual frequency = 30%: model had perfect calibration
- actual frequency < 30%: model was over-confident
- actual frequency > 30%: model was under-confident

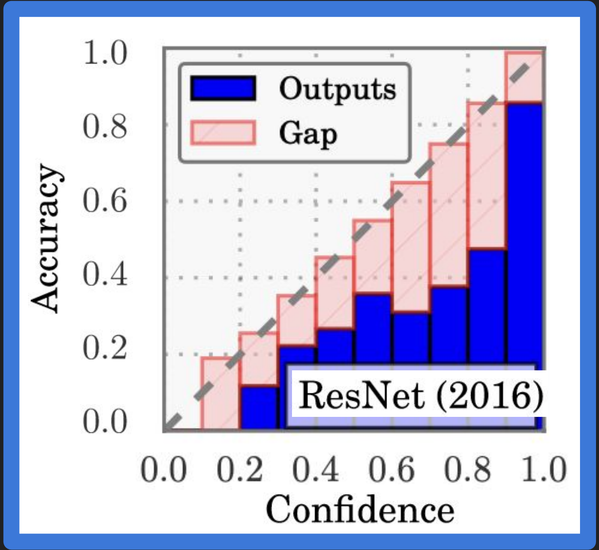
Today's most important slide

# Reliability diagram & ECE

- Bin(group) the data by model output (confidence) interval
- For each bin: compute "actual frequency" & compare with ideal value

Popular option: ECE (Expected Calibration Error). // 사실 정의하기 나름…

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|$$

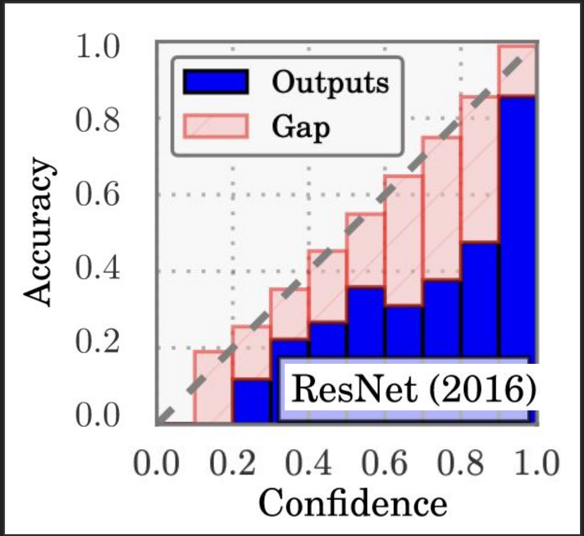



Quiz: Upper bound of ECE? / what if *M*=1?

# Reliability diagram & ECE

- ECE can't be evaluated on each data; must be binned.
- If multi-class *(K>2)*, we may calculate ECE for each category
- Properties
  - Perfect calibration does not imply accurate prediction
  - $0 \leq \text{ECE} \leq 1$
  - What if *M*=1? When would we want to do that?

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|$$



# Modern NNs are overconfident

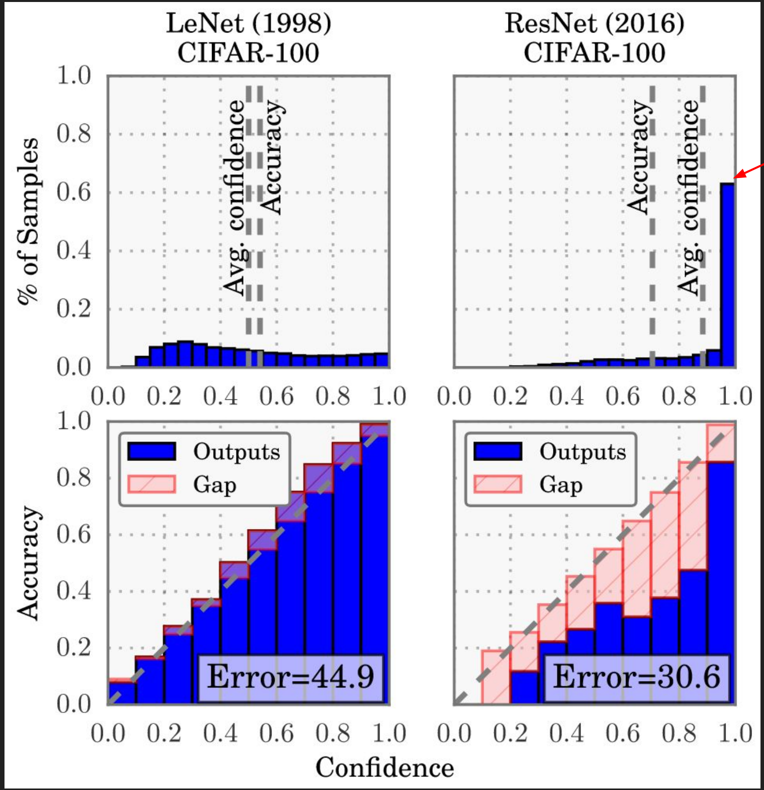
# On Calibration of Modern Neural Networks

Chuan Guo [*1] Geoff Pleiss [*1] Yu Sun [*1] Kilian Q. Weinberger [1]

(ICML 2017)


## Abstract

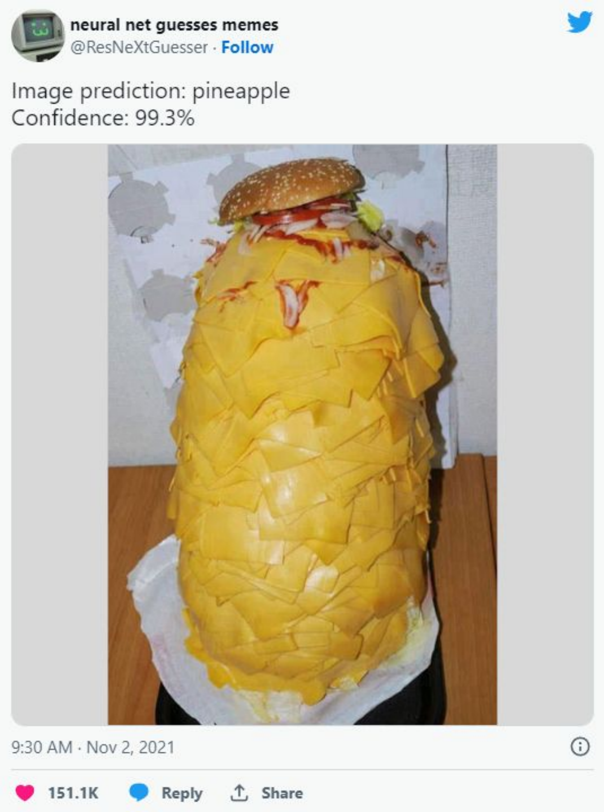
Confidence calibration – the problem of predicting probability estimates representative of the true correctness likelihood – is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated.


OMG…

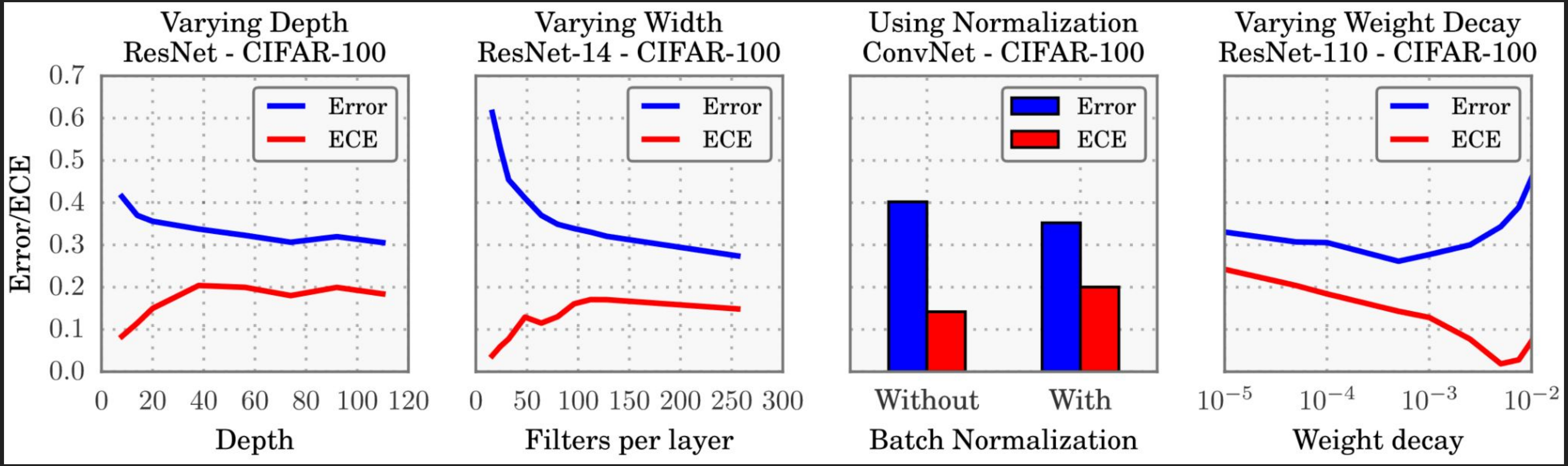
# Um… okay.


neural net guesses memes
@ResNeXtGuesser · Follow
Image prediction: ping-pong ball
Confidence: 99.99%
Submission by @Minish900
3:53 PM · Nov 10, 2021
110K
Reply
Share


neural net guesses memes
@ResNeXtGuesser · Follow
Image prediction: pineapple
Confidence: 99.3%
9:30 AM · Nov 2, 2021
151.1K
Reply
Share

# What causes miscalibration?

Turns out, the modern NN techniques have been harming calibration 😱
(this paper is empirical; so there's no deep analysis here)



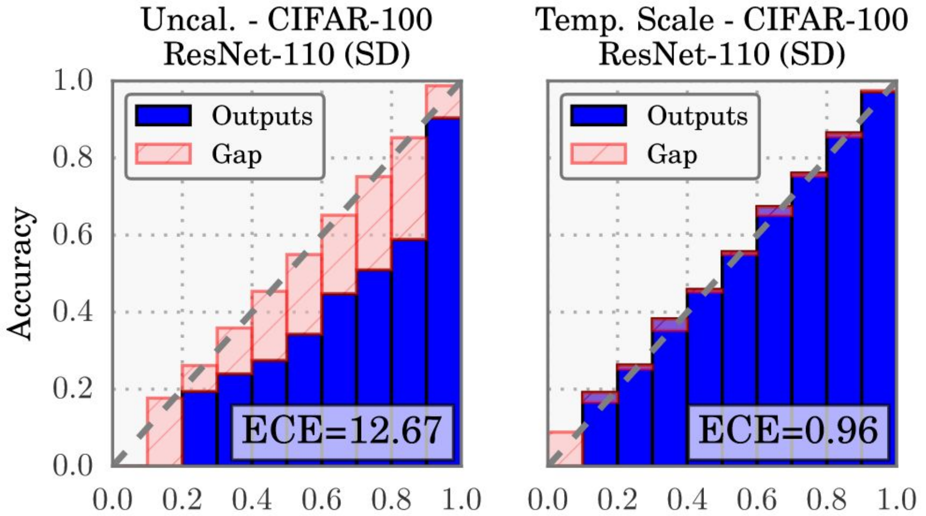
# How can we fix it?

Post-hoc calibration / Model regularization

# Temperature scaling: A post-hoc calibration

Divide all logits (values before softmax) by constant T (>0).

- With T>1, we can ‘flatten’ some overconfident predictions
- How to find optimal T? → optimize NLL on validation set!




Note.

- The ‘temperature’ here is identical to that of knowledge distillation.
- TS does not change ordering; thus, accuracy remains unchanged.

# Label Smoothing

**When Does Label Smoothing Help?**

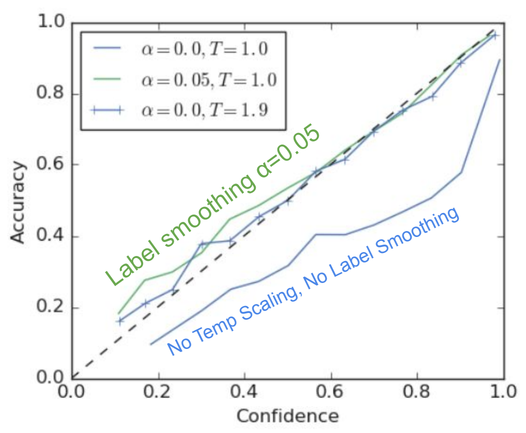
Rafael Müller,* Simon Kornblith, Geoffrey Hinton

(NeurIPS 2019)

"Resolving mis-calibration" = "Handling overconfidence" = "Label smoothing"

Table 3: Expected calibration error (ECE) on different architectures/datasets.

| Data set | Architecture | Baseline ECE (T=1.0, $\alpha = 0.0$) | Temp. scaling ECE / T ($\alpha = 0.0$) | Label smoothing ECE / $\alpha$ (T=1.0) |
|---|---|---|---|---|
| CIFAR-100 | ResNet-56 | 0.150 | 0.021 / 1.9 | 0.024 / 0.05 |
| ImageNet | Inception-v4 | 0.071 | 0.022 / 1.4 | 0.035 / 0.1 |
| EN-DE | Transformer | 0.056 | 0.018 / 1.13 | 0.019 / 0.1 |




(ResNet-56, CIFAR-100)

# Label Smoothing

Caveat: Mixed use of T.S. & L.S. damages both ECE & ‘accuracy’.

w/o Label Smoothing w/ Label Smoothing

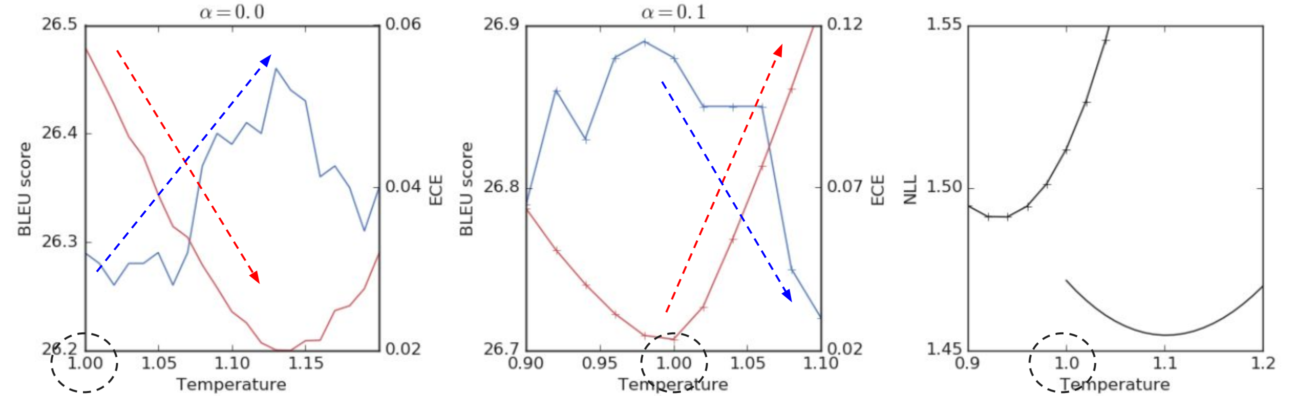
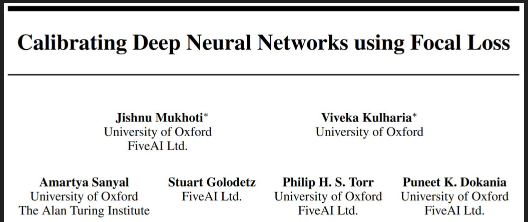



Figure 4: Effect of calibration of Transformer upon BLEU score (blue lines) and NLL (red lines). Curves without markers reflect networks trained without label smoothing while curves with markers represent networks with label smoothing.

# Focal Loss


**Calibrating Deep Neural Networks using Focal Loss**


**Jishnu Mukhoti*** University of Oxford FiveAI Ltd.
**Viveka Kulharia*** University of Oxford
**Amartya Sanyal** University of Oxford The Alan Turing Institute
**Stuart Golodetz** FiveAI Ltd.
**Philip H. S. Torr** University of Oxford FiveAI Ltd.
**Puneet K. Dokania** University of Oxford FiveAI Ltd.


(NeurIPS 2020)

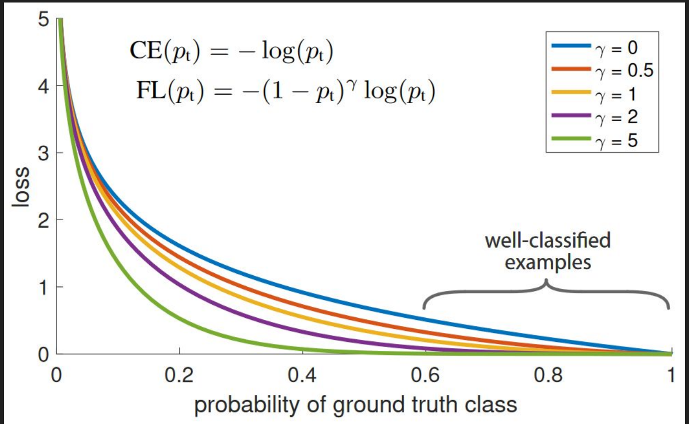
(Yes, it's a concept derived from *RetinaNet* for Dense Object Detection!)

"Focus on learning hard samples" = "Prevent overconfidence"

$$\mathcal{L}_{\text{CE}} = -\log p$$

$$\mathcal{L}_{\text{Focal}} = -(1-p)^{\gamma} \log p$$

- With CE, loss is non-trivial even when *p*>0.5
  - Even after achieving 100% accuracy, optimizer can still reduce loss by making model overconfident.
  - Let's assign smaller loss on easy samples.




Quiz: CE(0.9), FL(0.9) = ? (no calculators!)

# Label Smoothing & Focal Loss – with equations

L.S. = encourage larger sum(log p) of confidence output

$$\mathcal{L}_{\mathrm{CE}}\left(q^{\mathrm{LS}}, \hat{p}\right) = (1-\epsilon)\mathcal{L}_{\mathrm{CE}}(q, \hat{p}) + \boxed{\epsilon\mathcal{L}_{\mathrm{CE}}(U, \hat{p})}$$

($U$: uniform distribution)

F.L. = encourage larger entropy of confidence output

$$\mathcal{L}_f \geq \mathrm{KL}(q||\hat{p}) + \underbrace{\mathbb{H}[q]}_{constant} \boxed{-\gamma\mathbb{H}[\hat{p}]}.$$

(proof: Appendix B)

Both values are minimal when $p=U$ since -log($p$), -$p$log($p$) is convex downward.
→ prevent overconfidence.

# Inverse Focal Loss


Rethinking Calibration of Deep Neural Networks:
Do Not Be Afraid of Overconfidence

Deng-Bao Wang,[1,2] Lei Feng,[3] Min-Ling Zhang[1,2*]

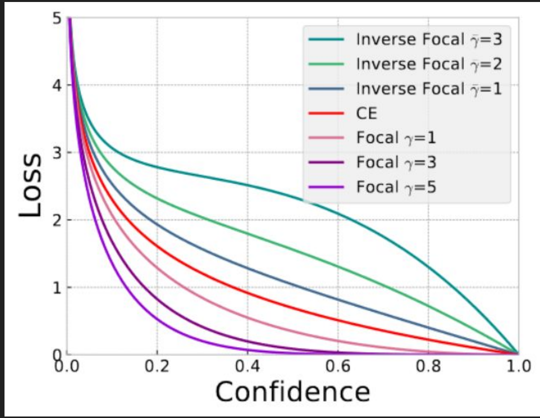
(NeurIPS 2021)

Is overconfidence really an issue?

Regularizing model to produce less-confident results might result in mixing up easy/hard samples.
→ Less distinguishable, worse ECE after T.S.

<u>"From *Calibrated* to *Calibratable*"</u>
Let's amplify the overconfidence (higher loss on easy) so that easy/hard samples are more distinguishable.
→ Better ECE after T.S.

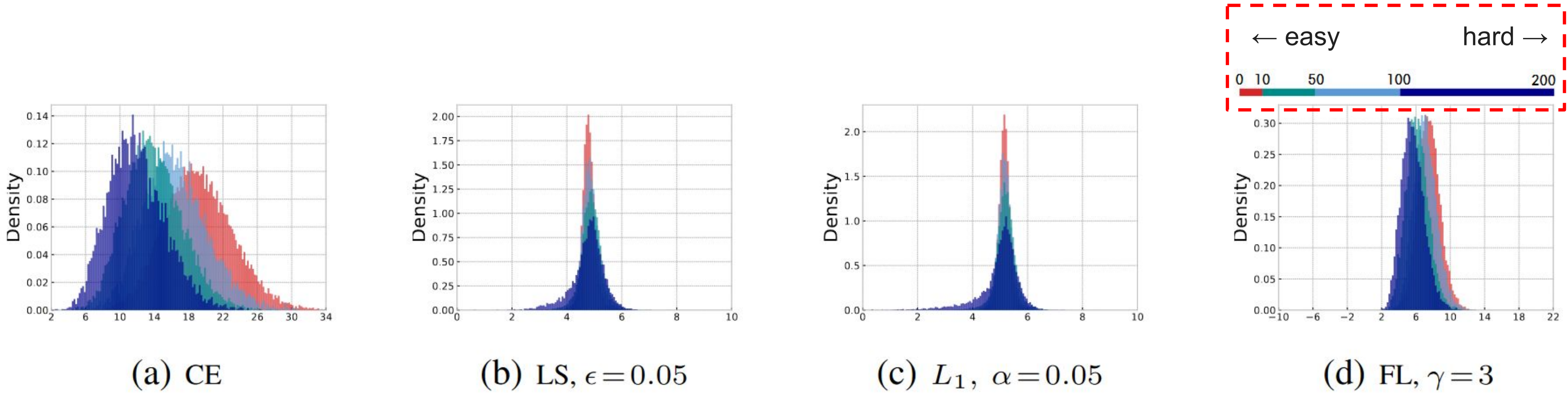
$$\mathcal{L}_{\text{Inv. Focal}} = -(1+p)^{\gamma} \log p$$




Disclaimer: I should mention that the inverse focal loss itself is NOT this work's main contribution.

# Inverse Focal Loss – more distinguishable samples

Def. *learned epoch*: at what epoch does the sample get correctly classified?

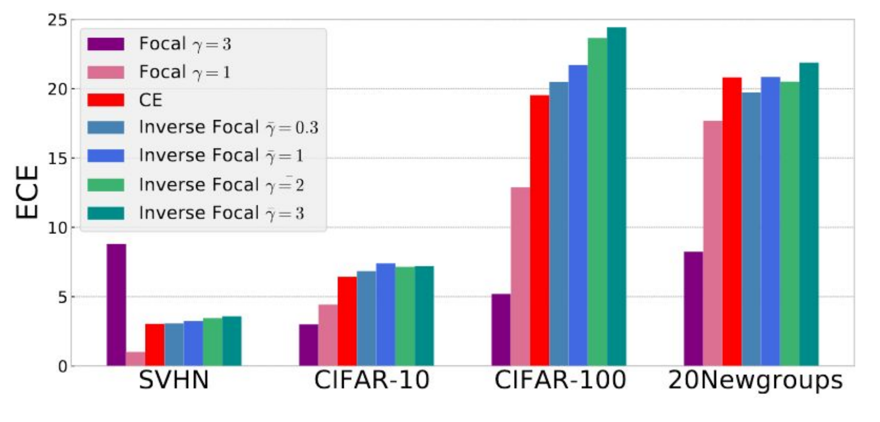



Label Smoothing

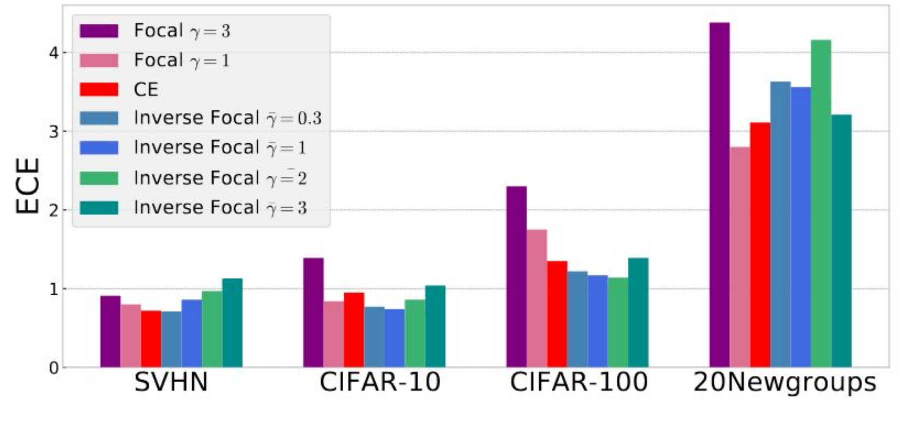
(other regularization)

Focal Loss

# Inverse Focal Loss – Better ECE after T.S.




(c) ECE without post-hoc calibration




(e) ECE after TS

# Wrapping up

# Takeaways

- Modern NNs are widely miscalibrated & overconfident.
    - Higher accuracy does not lead to good calibration
- Calibration can be quantified with ECE & visualized with Reliability diagram
- To resolve miscalibration:
    - Temperature scaling as a post-hoc calibration
    - Model regularizations (label smoothing, focal loss) to prevent overconfidence
- But, model regularization can hurt ability to distinguish easy/hard samples.

# References


- [1] “On Calibration of Modern Neural Networks”, Guo et al. [link]
- [2] “When Does Label Smoothing Help?”, Muller et al. [link]
- [3] “Calibrating Deep Neural Networks using Focal Loss”, Mukhoti et al. [link] [blog]
- [4] “Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence”, Wang et al. [link] [OpenReview]

- “Introduction to Uncertainty in Deep Learning”, Balaji Lakshminarayanan [link]
- Paper review of [4] at dsba.korea.ac.kr, 박경찬 [link]