2023년 5월 12일 DMQA 연구실 오픈 세미나

On Calibration of Deep Neural Networks

고려대학교 산업경영공학부 배진수

발표자

❖ 배진수

- 건국대학교 수학과 학부 졸업
- 고려대학교 산업경영공학과 대학원 재학
- 고려대학교 DMQA 연구실 (지도교수: 김성범)
- 박사과정 2년 차
- wlstn215@korea.ac.kr

❖ 연구분야

- Safe Semi-Supervised Learning Using a Bayesian Neural Network
- Improving Calibration of Deep Neural Networks

세미나 내용

❖ On Calibration of Deep Neural Networks

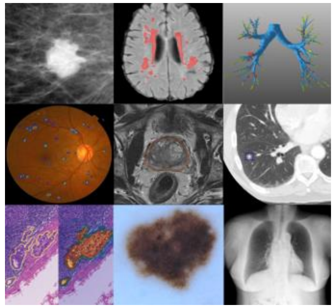
1. Calibration
2. Calibration of (Modern) Deep Neural Networks
3. Improving Calibration of Deep Neural Networks

1. Calibration

Calibration

❖ 딥러닝 발전 → 여러 분야에서 우수한 성능

❖ 높은 정확도를 가지고 있으면 실제 산업에 바로 사용할 수 있을까?



Calibration

- ❖ 딥러닝 발전 → 여러 분야에서 우수한 성능
- ❖ 높은 정확도를 가지고 있으면 실제 산업에 바로 사용할 수 있을까?

오작동 기회 비용이 큰 분야

[이슈톡] '자율주행'이라더니...테슬라, 전복 트럭도 못 피하고 정면 충돌

입력 2020-06-03 06:51 | 수정 2020-06-03 09:22

Calibration

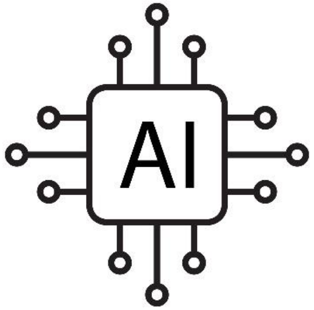
❖ 높은 정확도를 가지고 있으면 실제 산업에 바로 사용할 수 있을까?

“정상일 확률: 53%”

“질병(A)일 확률: 3%”

“질병(B)일 확률: 23%”

“질병(C)일 확률: 21%”



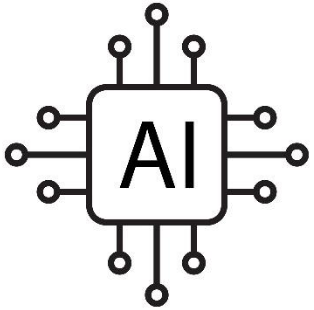
모델 1: 정확도 95%

“정상일 확률: 35%”

“질병(A)일 확률: 21%”

“질병(B)일 확률: 23%”

“질병(C)일 확률: 21%”



모델 2: 정확도 95%

질병 A, 질병 B, 질병 C, 정상

Calibration

❖ 높은 정확도를 가지고 있으면 실제 산업에 바로 사용할 수 있을까? → <u>정확도 이외에도 고려해야 할 부분이 있다</u>

- 확신에 찬 상태로 틀리는 모델 vs 불확실한 상태로 틀리는 모델

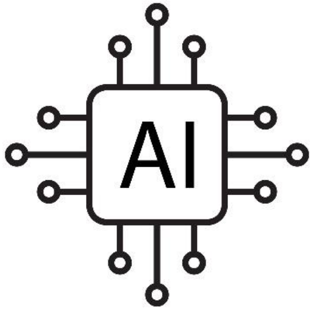
모델 1보다 모델 2를 더 신뢰할 수 있다.

"정상일 확률: 53%"

"질병(A)일 확률: 3%"

"질병(B)일 확률: 23%"

"질병(C)일 확률: 21%"



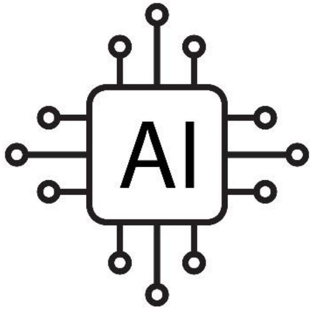
모델 1: 정확도 95%

"정상일 확률: 35%"

"질병(A)일 확률: 21%"

"질병(B)일 확률: 23%"

"질병(C)일 확률: 21%"



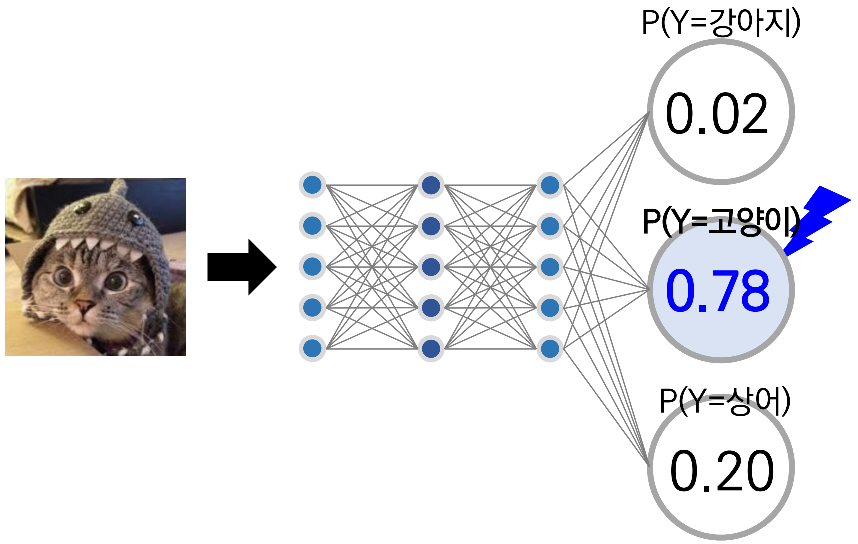
모델 2: 정확도 95%

질병 A, 질병 B, 질병 C, 정상

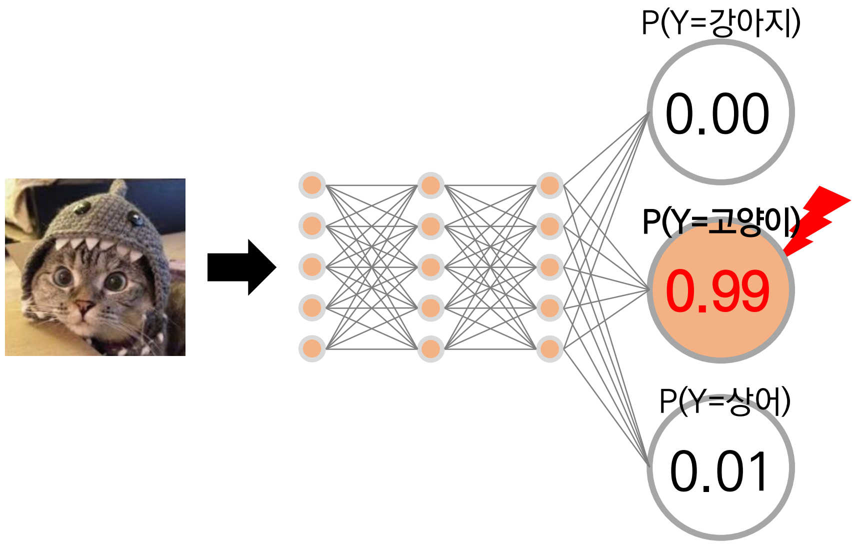
Calibration

❖ 신뢰할 수 있는 좋은 모델은 구체적으로 어떤 특성을 가질까? (이미지 분류 모델 기준)

모델 1: 정확도 80%



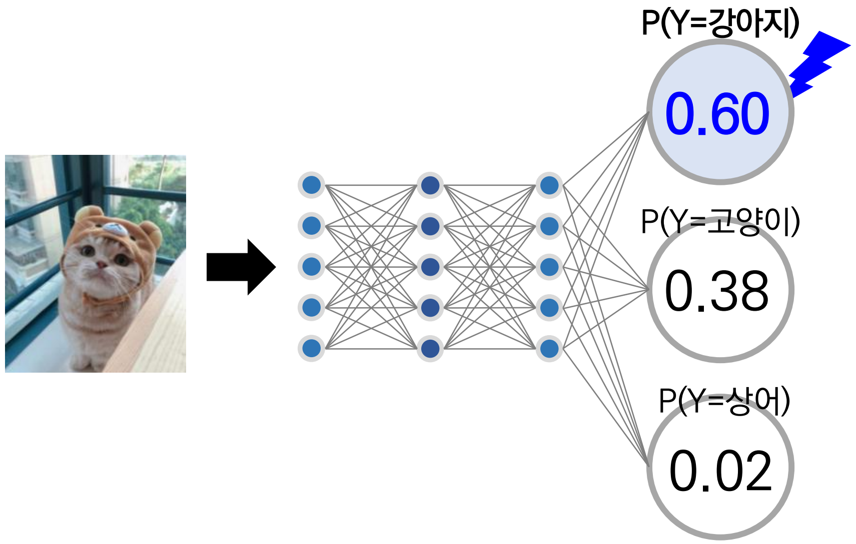
모델 2: 정확도 80%



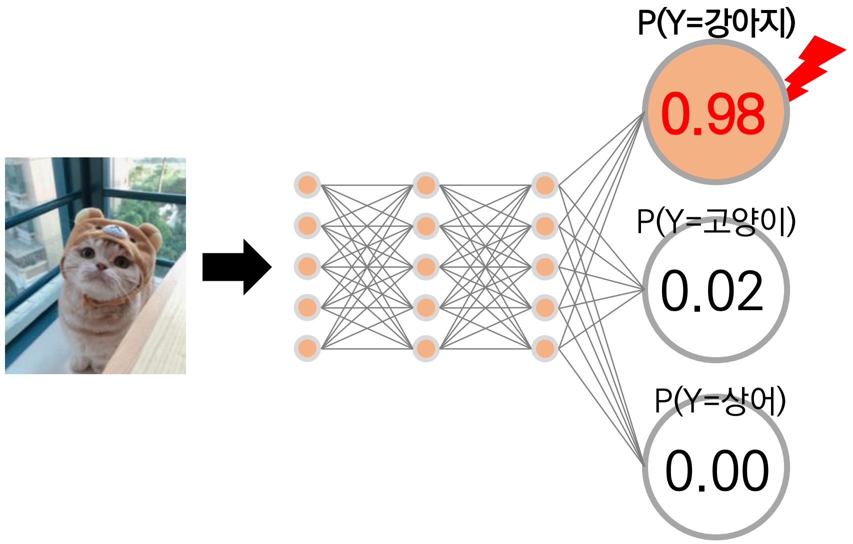
Calibration

❖ 신뢰할 수 있는 좋은 모델은 구체적으로 어떤 특성을 가질까? (이미지 분류 모델 기준)

모델 1: 정확도 80%



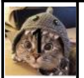
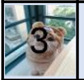
모델 2: 정확도 80%



Calibration

❖ 신뢰할 수 있는 좋은 모델은 구체적으로 어떤 특성을 가질까? (이미지 분류 모델 기준)

모델 1: 정확도 80%

Idx	P(Y=강아지)	P(Y=고양이)	P(Y=상어)	정답	예측	예측 확률
 1	0.1	0.75	0.15	고양이	고양이	75%
 2	0.1	0.8	0.1	고양이	고양이	80%
 3	0.6	0.38	0.02	고양이	강아지	60%
 4	0.1	0.8	0.1	고양이	고양이	80%
 5	0.01	0.99	0.00	고양이	고양이	99%

정확도=80% ≈ 확률값=78.8% $= \frac{75+80+60+80+99}{5}$

모델 2: 정확도 80%

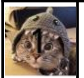
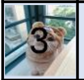
Idx	P(Y=강아지)	P(Y=고양이)	P(Y=상어)	정답	예측	예측 확률
 1	0.02	0.98	0.00	고양이	고양이	98%
 2	0.03	0.97	0.00	고양이	고양이	97%
 3	0.95	0.02	0.03	고양이	강아지	95%
 4	0.00	0.95	0.05	고양이	고양이	95%
 5	0.01	0.99	0.00	고양이	고양이	99%

정확도=80% ≪ 확률값=96.8% $= \frac{98+97+95+95+99}{5}$

Calibration

❖ 신뢰할 수 있는 좋은 모델은 구체적으로 어떤 특성을 가질까? ➔ 예측 모델의 확률 결과와 정확도가 유사함

모델 1: 정확도 80%

Idx	P(Y=강아지)	P(Y=고양이)	P(Y=상어)	정답	예측	예측 확률
 1	0.1	0.75	0.15	고양이	고양이	75%
 2	0.1	0.8	0.1	고양이	고양이	80%
 3	0.6	0.38	0.02	고양이	강아지	60%
 4	0.1	0.8	0.1	고양이	고양이	80%
 5	0.01	0.99	0.00	고양이	고양이	99%

정확도=80% ≈ 확률값=78.8% $= \frac{75+80+60+80+99}{5}$

모델 2: 정확도 80%

Idx	P(Y=강아지)	P(Y=고양이)	P(Y=상어)	정답	예측	예측 확률
 1	0.02	0.98	0.00	고양이	고양이	98%
 2	0.03	0.97	0.00	고양이	고양이	97%
 3	0.95	0.02	0.03	고양이	강아지	95%
 4	0.00	0.95	0.05	고양이	고양이	95%
 5	0.01	0.99	0.00	고양이	고양이	99%

정확도=80% ≪ 확률값=96.8% $= \frac{98+97+95+95+99}{5}$

Calibration

❖ 신뢰할 수 있는 좋은 모델은 구체적으로 어떤 특성을 가질까? ➔ 예측 모델의 확률 결과와 정확도가 유사함

Perfect-Calibrated Predictive Model

$$\downarrow \quad \uparrow$$

$$P\left(\hat{Y} = Y \middle| \hat{P} = p\right) = p, \forall p \in [0,1]$$

Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

$$\forall p \in [0,1] \;\rightarrow\; \left|\underbrace{P\left(\hat{Y} = Y \middle| \hat{P} = p\right)}_{\text{정답과 예측이 똑같은 경우 (=예측 모델의 정확도)}} - \underbrace{p}_{\text{0~1 사이의 확률}}\right|$$

Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

- M= 데이터 내에 집단 개수, n= 전체 데이터 개수, $|B_m|$= 집단 B_m의 데이터 개수

$$\mathrm{E}_{\hat{\mathrm{P}}}\left[\left|\underbrace{P\left(\hat{Y}=Y \middle| \hat{P}=p\right)}_{\substack{\text{정답과 예측이 똑같은 경우} \\ \text{(=예측 모델의 정확도)}}} - \underbrace{p}_{\text{0~1 사이의 확률}}\right|\right] \approx \Sigma_{m=1}^{M} \frac{|\mathrm{B_m}|}{\mathrm{n}} \left|\underbrace{\mathrm{acc}(\mathrm{B_m})}_{\substack{\text{집단 } \mathrm{B_m}\text{에 대한} \\ \text{예측 모델 정확도}}} - \underbrace{\mathrm{conf}(\mathrm{B_m})}_{\substack{\text{집단 } \mathrm{B_m}\text{에 대한} \\ \text{모델 확률값들의 평균}}}\right|$$

Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

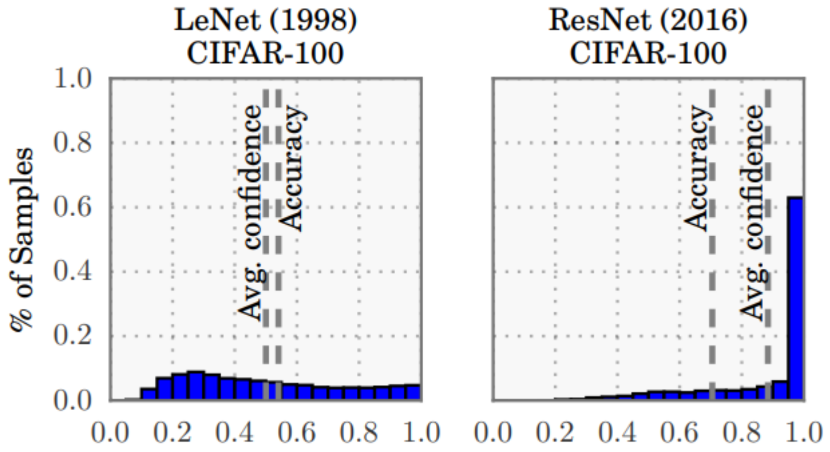
- M= 데이터 내에 집단 개수, n= 전체 데이터 개수, $|B_m|$= 집단 B_m의 데이터 개수
- D= 전체 데이터셋

$$\Sigma_{m=1}^{M} \frac{|\mathrm{B_m}|}{\mathrm{n}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})|$$

예시) M=1

$$|\mathrm{acc}(\mathrm{D}) - \mathrm{conf}(\mathrm{D})|$$

Confidence Histograms



Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

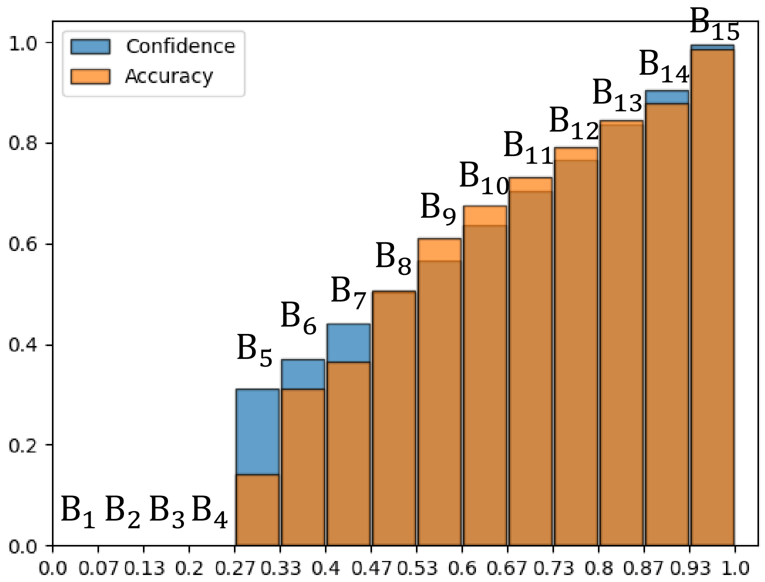
- M= 데이터 내에 집단 개수, n= 전체 데이터 개수, $|B_m|$= 집단 B_m의 데이터 개수

$$\Sigma_{m=1}^{M} \frac{|\mathrm{B_m}|}{\mathrm{n}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})|$$

예시) M=15

$$\Sigma_{m=1}^{15} \frac{|\mathrm{B_m}|}{\mathrm{n}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})|$$

Reliability Diagrams



Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

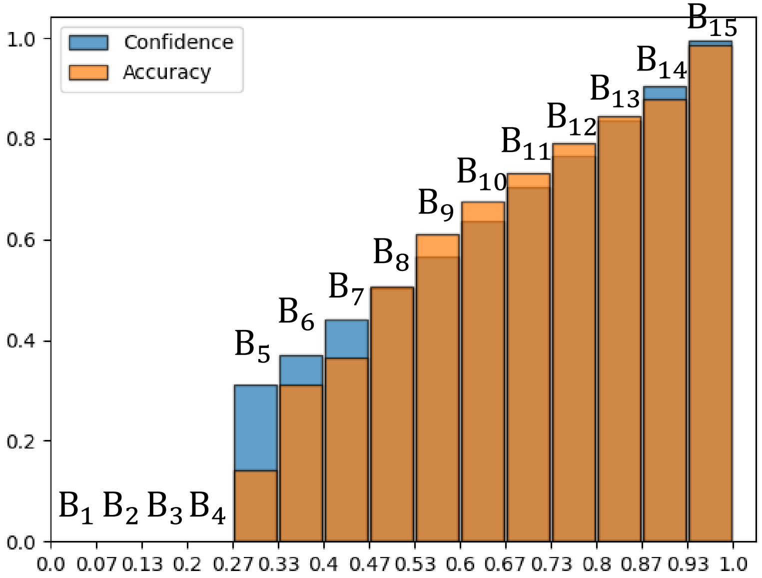
- Expected Calibration Error (ECE) = $\Sigma_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$

$$\Sigma_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

예시) M=15

$$\Sigma_{m=1}^{15} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

Reliability Diagrams



To estimate the expected accuracy from finite samples, we group predictions into M interval bins (each of size $1/M$) and calculate the accuracy of each bin. Let B_m be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The accuracy of B_m is

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i),$$

where $\hat{y}_i$ and y_i are the predicted and true class labels for sample i. Basic probability tells us that $\mathrm{acc}(B_m)$ is an unbiased and consistent estimator of $\mathbb{P}(\hat{Y} = Y \mid \hat{P} \in I_m)$. We define the average confidence within bin B_m as

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

where $\hat{p}_i$ is the confidence for sample i. $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ approximate the left-hand and right-hand sides of (1) respectively for bin B_m. Therefore, a perfectly calibrated model will have $\mathrm{acc}(B_m) = \mathrm{conf}(B_m)$ for all $m \in \{1, \ldots, M\}$. Note that reliability diagrams do not display the proportion of samples in a given bin, and thus cannot be used to estimate how many samples are calibrated.

Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

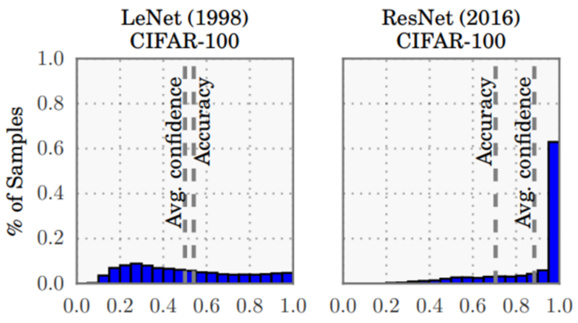
Calibration

❖ 신뢰도 관점에서 모델 성능을 어떻게 평가할까? ➔ 예측 모델의 확률 결과와 정확도의 차이가 얼마인지 확인

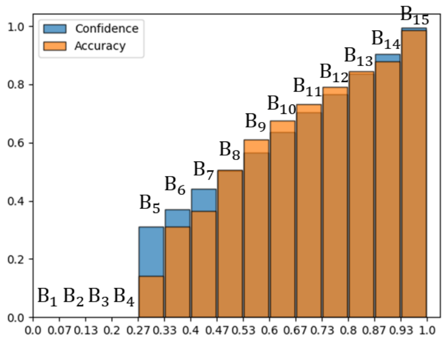
- Testing Data에 대한 ECE 지표, Reliability Diagrams, Confidence Histograms 확인
- Corrupted Testing Data에 대한 ECE 지표, Reliability Diagrams, Confidence Histograms 확인

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

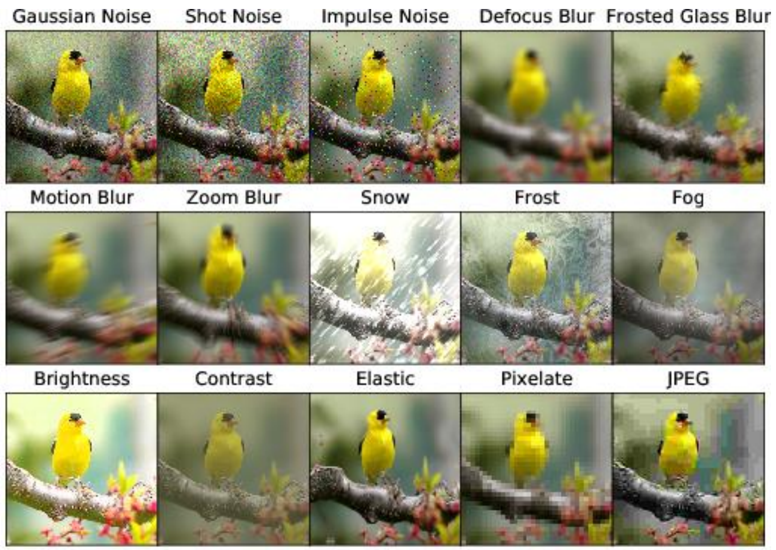
Confidence Histograms



Reliability Diagrams



Noisy(Corrupted) Data



Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.
Hendrycks, Dan, et al. "Augmix: A simple data processing method to improve robustness and uncertainty." ICLR, 2020.

2. Calibration of (Modern) Deep Neural Networks

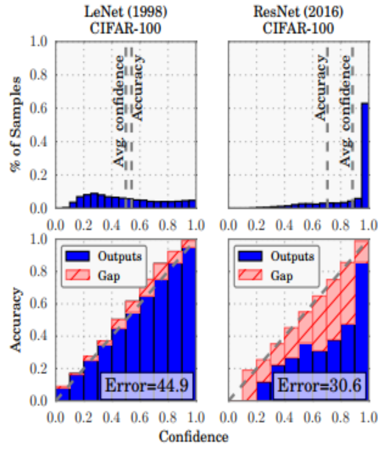
✓ PMLR 2017, 3739회 인용

On Calibration of Modern Neural Networks

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

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. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Our analysis and experiments not only offer insights into neural network learning, but also provide a simple and straightforward recipe for practical settings: on most datasets, *temperature scaling* – a single-parameter variant of Platt Scaling – is surprisingly effective at calibrating predictions.



✓ NeurIPS 2021, 117회 인용

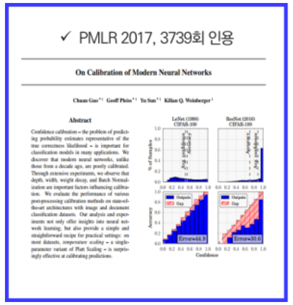
Revisiting the Calibration of Modern Neural Networks

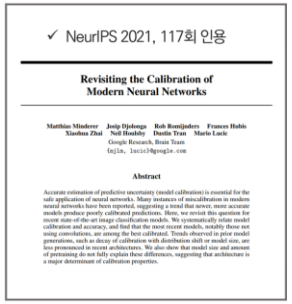
Matthias Minderer Josip Djolonga Rob Romijnders Frances Hubis
Xiaohua Zhai Neil Houlsby Dustin Tran Mario Lucic
Google Research, Brain Team
{mjlm, lucic}@google.com

Abstract

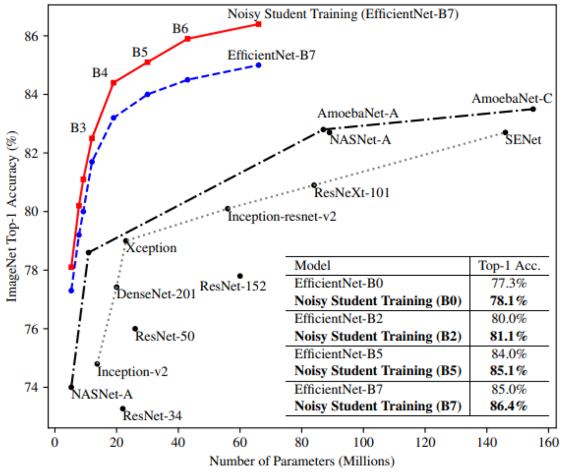
Accurate estimation of predictive uncertainty (model calibration) is essential for the safe application of neural networks. Many instances of miscalibration in modern neural networks have been reported, suggesting a trend that newer, more accurate models produce poorly calibrated predictions. Here, we revisit this question for recent state-of-the-art image classification models. We systematically relate model calibration and accuracy, and find that the most recent models, notably those not using convolutions, are among the best calibrated. Trends observed in prior model generations, such as decay of calibration with distribution shift or model size, are less pronounced in recent architectures. We also show that model size and amount of pretraining do not fully explain these differences, suggesting that architecture is a major determinant of calibration properties.

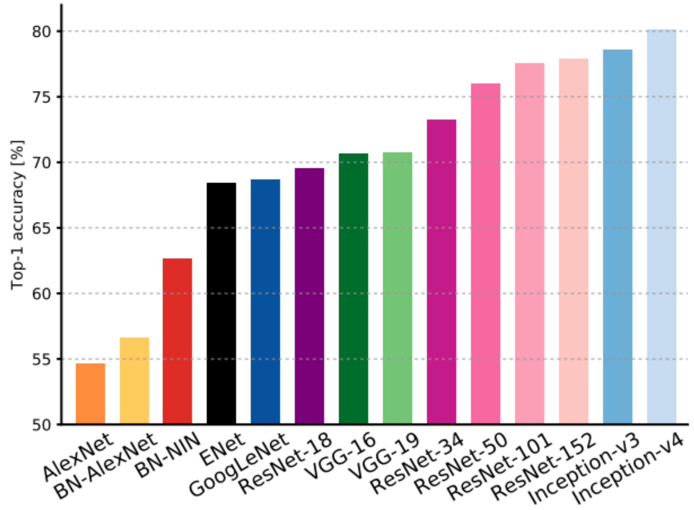
Calibration of (Modern) Deep Neural Networks

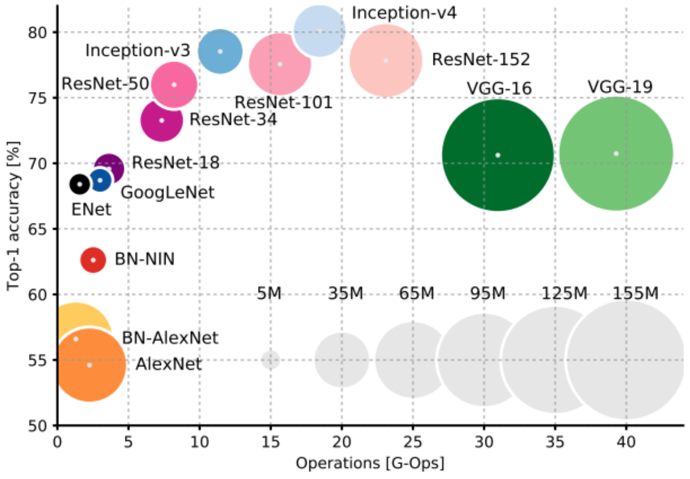




- ❖ 현대에서 주로 사용하고 있는 딥러닝 모델들의 공통점: <u>크고 넓고 무겁고 높은 정확도를 가짐</u>
- ❖ 크고 넓고 무거운 모델이 오히려 일반화 성능이 뛰어난 경향이 있음 <u>(학습 데이터 개수가 적더라도 유리)</u>

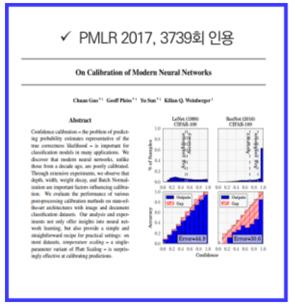


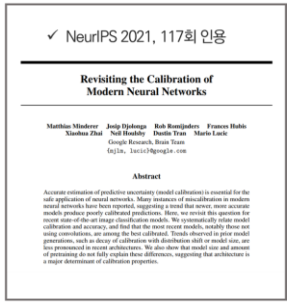




Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. In ICLR, 2017.
Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
https://culurciello.medium.com/analysis-of-deep-neural-networks-dcf398e71aae

Calibration of (Modern) Deep Neural Networks

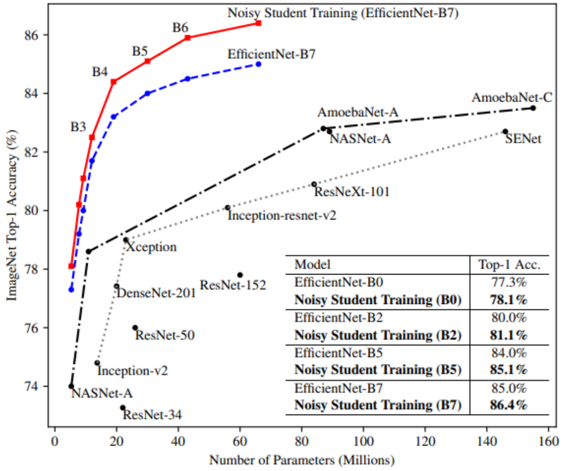


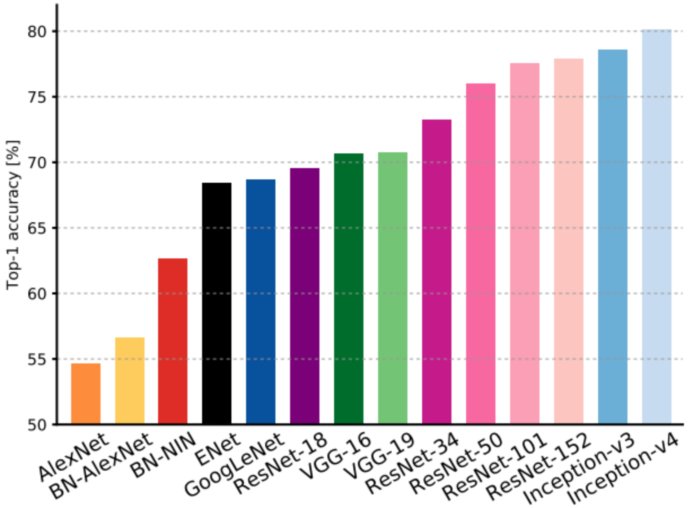


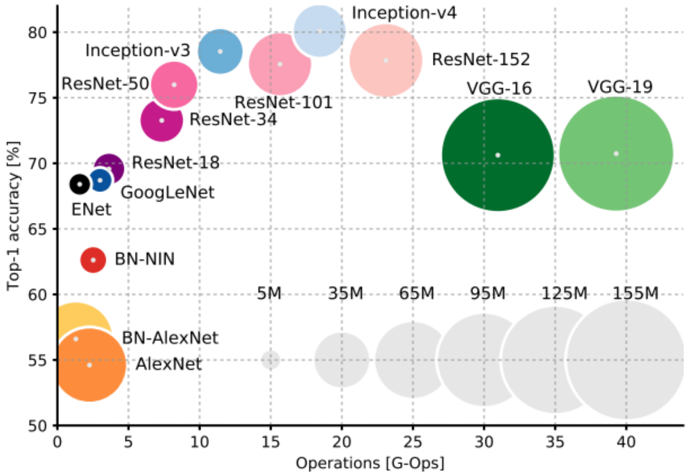
❖ 현대에서 주로 사용하고 있는 딥러닝 모델들의 공통점: <u>크고 넓고 무겁고 높은 정확도를 가짐</u>

❖ 크고 넓고 무거운 모델이 오히려 일반화 성능이 뛰어난 경향이 있음 <u>(학습 데이터 개수가 적더라도 유리)</u>

➢ <u>정확도와 신뢰도가 함께 향상되고 있을까? 정확도만 올라가고 신뢰도는 떨어지고 있을까?</u>

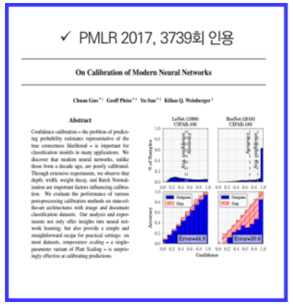


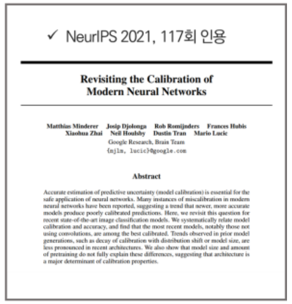




Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. In ICLR, 2017.
Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
https://culurciello.medium.com/analysis-of-deep-neural-networks-dcf398e71aae

Calibration of (Modern) Deep Neural Networks





❖ 현대에서 주로 사용하고 있는 딥러닝 모델들의 공통점: <u>크고 넓고 무겁고 높은 정확도를 가짐 → 신뢰도는 떨어지고 있는 중</u>

- ResNet Depth, 합성곱 연산 필터의 개수, Batch Normalization, Weight Decay가 신뢰도와 연관되어 있음

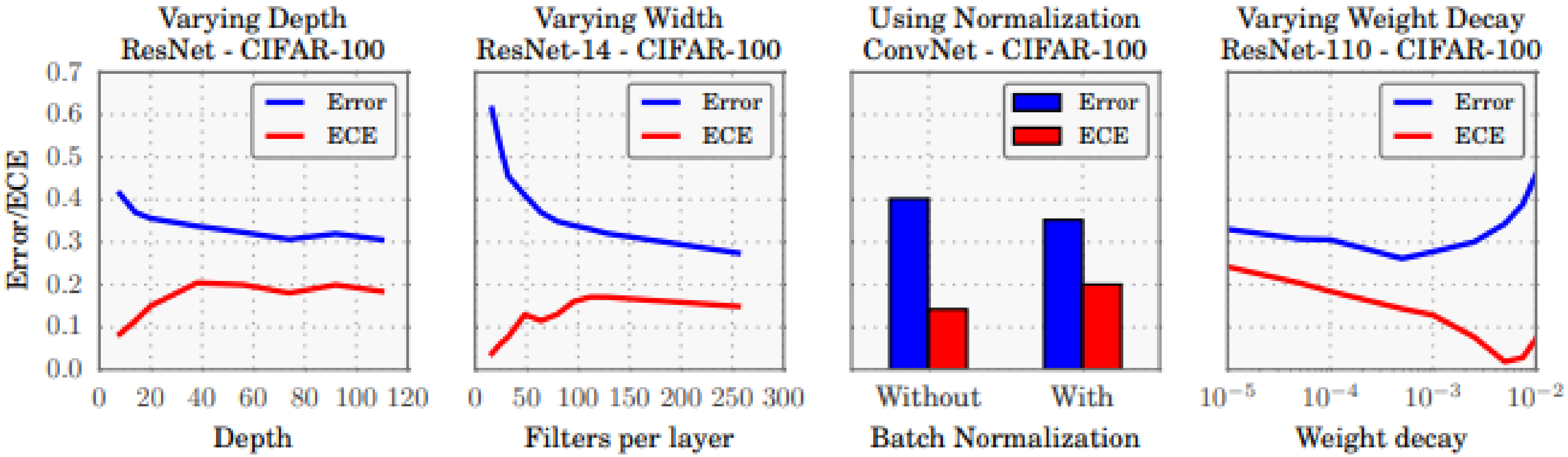
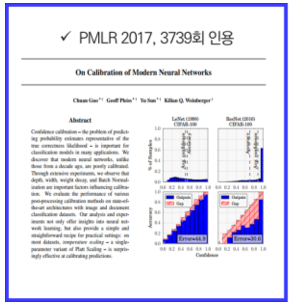
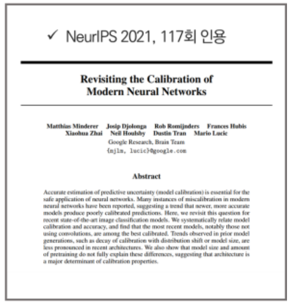


Figure 2. The effect of network depth (far left), width (middle left), Batch Normalization (middle right), and weight decay (far right) on miscalibration, as measured by ECE (lower is better).

Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

Calibration of (Modern) Deep Neural Networks





❖ 현대에서 주로 사용하고 있는 딥러닝 모델들의 공통점: <u>크고 넓고 무겁고 높은 정확도를 가짐 → 신뢰도는 떨어지고 있는 중</u>

- 학습 데이터에 대한 정확도가 100% 임에도 Cross-Entropy 기준 학습이 계속 이루어지면 Overconfidence 발생 → 신뢰도 하락

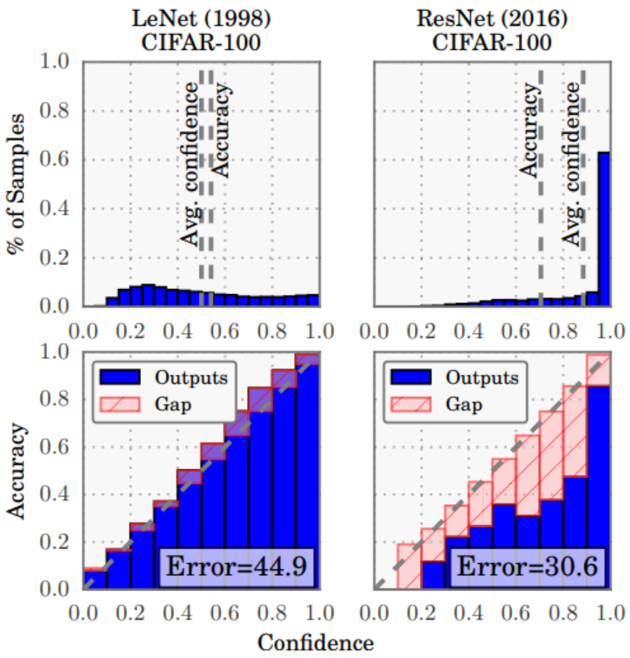
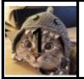


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

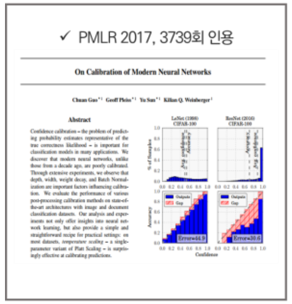
		LeNet			ResNet		
Idx	정답	P(Y=강아지)	P(Y=고양이)	P(Y=상어)	P(Y=강아지)	P(Y=고양이)	P(Y=상어)
 1	고양이	0.1	0.75	0.15	0.00	1.00	0.00
 2	고양이	0.1	0.8	0.1	0.00	1.00	0.00
 3	고양이	0.1	0.8	0.1	0.00	1.00	0.00
…	…	…	…	…	…	…	…
 1M	강아지	0.96	0.02	0.02	1.00	0.00	0.00

Training Accuracy: 100%

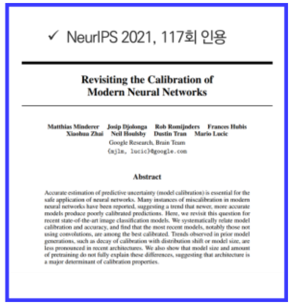
Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

Calibration of (Modern) Deep Neural Networks

✓ PMLR 2017, 3739회 인용



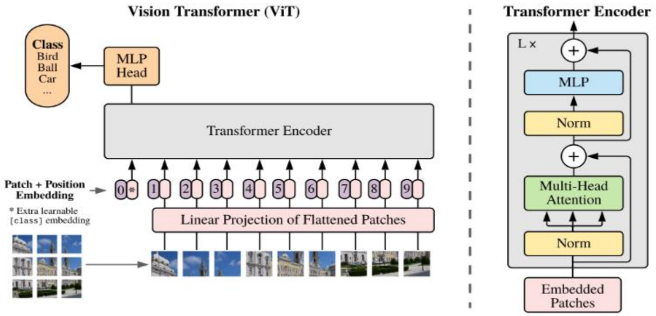
✓ NeurIPS 2021, 117회 인용



❖ 어떤 구조를 가진 이미지 분류 모델이 <u>더 좋은 신뢰도를 가지고 있을까?</u>

- Convolutional VS Non-Convolutional [Vision Transformer(2021), MLP-Mixer(2021), ⋯, etc.]

Vision Transformer



MLP-Mixer

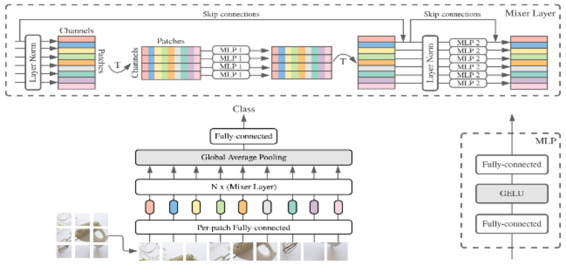
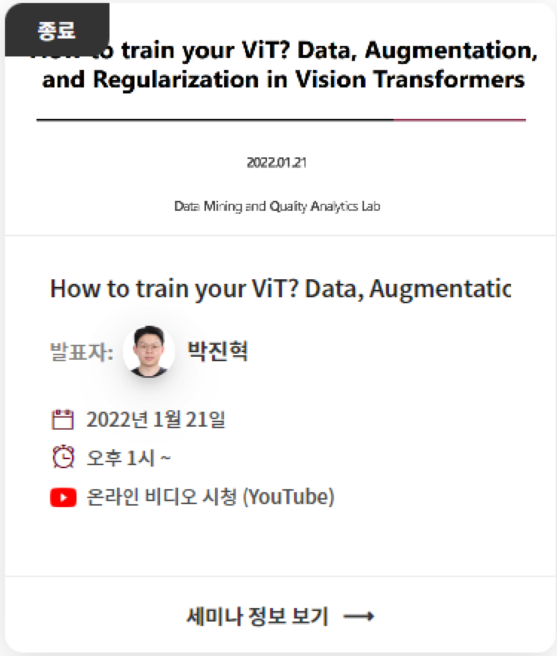
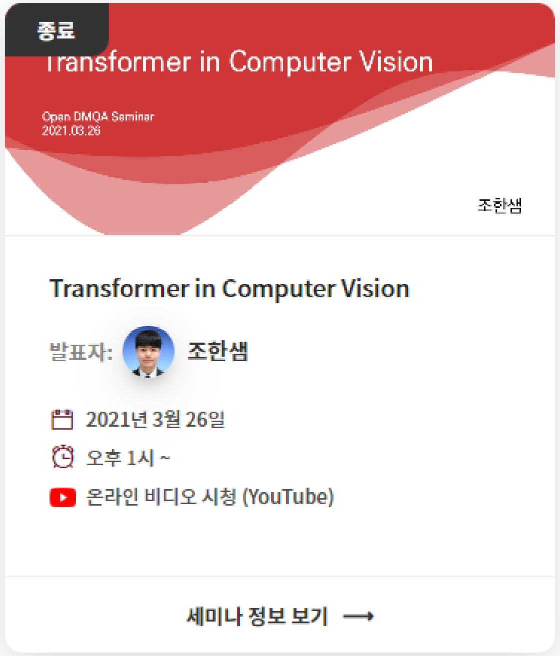


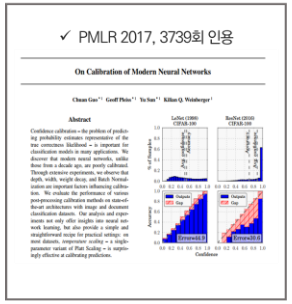
Figure 1: MLP-Mixer consists of per-patch linear embeddings, Mixer layers, and a classifier head. Mixer layers contain one token-mixing MLP and one channel-mixing MLP, each consisting of two fully-connected layers and a GELU nonlinearity. Other components include: skip-connections, dropout, and layer norm on the channels.

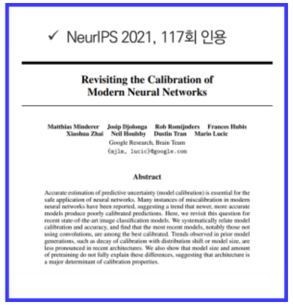




Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR, 2021.
Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." Advances in neural information processing systems, 2021.

Calibration of (Modern) Deep Neural Networks





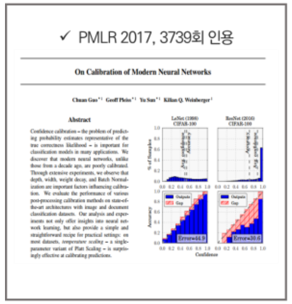
❖ 어떤 구조를 가진 이미지 분류 모델이 더 좋은 신뢰도를 가지고 있을까?

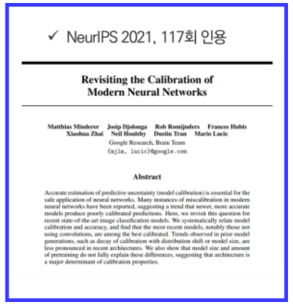
- Convolutional VS Non-Convolutional [Vision Transformer(2021), MLP-Mixer(2021), ⋯, etc.]

1. MLP-Mixer (Tolstikhin et al., 2021) is based exclusively on multi-layer perceptrons (MLPs) and is pre-trained on large supervised datasets.
2. ViT (Dosovitskiy et al., 2021) processes images with a transformer architecture originally designed for language (Vaswani et al., 2017) and is also pre-trained on large supervised datasets.
3. BiT (Kolesnikov et al., 2020) is a ResNet-based architecture (He et al., 2016). It is also pre-trained on large supervised datasets.
4. ResNext-WSL (Mahajan et al., 2018) is based on the ResNeXt architecture and trained with weak supervision from billions of hashtags on social media images.
5. SimCLR (Chen et al., 2020) is a ResNet, pretrained with an unsupervised contrastive loss.
6. CLIP (Radford et al., 2021) is pretrained on raw text and imagery using a contrastive loss.
7. AlexNet (Krizhevsky et al., 2012; Krizhevsky, 2014) was the first convolutional neural network to win the ImageNet challenge.

Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

Calibration of (Modern) Deep Neural Networks





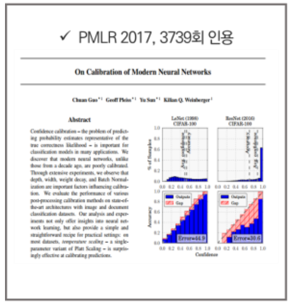
❖ 어떤 구조를 가진 이미지 분류 모델이 더 좋은 신뢰도를 가지고 있을까?

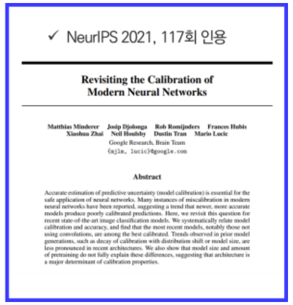
- Convolutional VS Non-Convolutional [Vision Transformer(2021), MLP-Mixer(2021), …, etc.]
- 사전학습의 영향이 있진 않았을까?

1. MLP-Mixer (Tolstikhin et al., 2021) is based exclusively on multi-layer perceptrons (MLPs) and is pre-trained on large supervised datasets.
2. ViT (Dosovitskiy et al., 2021) processes images with a transformer architecture originally designed for language (Vaswani et al., 2017) and is also pre-trained on large supervised datasets.
3. BiT (Kolesnikov et al., 2020) is a ResNet-based architecture (He et al., 2016). It is also pre-trained on large supervised datasets.
4. ResNext-WSL (Mahajan et al., 2018) is based on the ResNeXt architecture and trained with weak supervision from billions of hashtags on social media images.
5. SimCLR (Chen et al., 2020) is a ResNet, pretrained with an unsupervised contrastive loss.
6. CLIP (Radford et al., 2021) is pretrained on raw text and imagery using a contrastive loss.
7. AlexNet (Krizhevsky et al., 2012; Krizhevsky, 2014) was the first convolutional neural network to win the ImageNet challenge.

Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

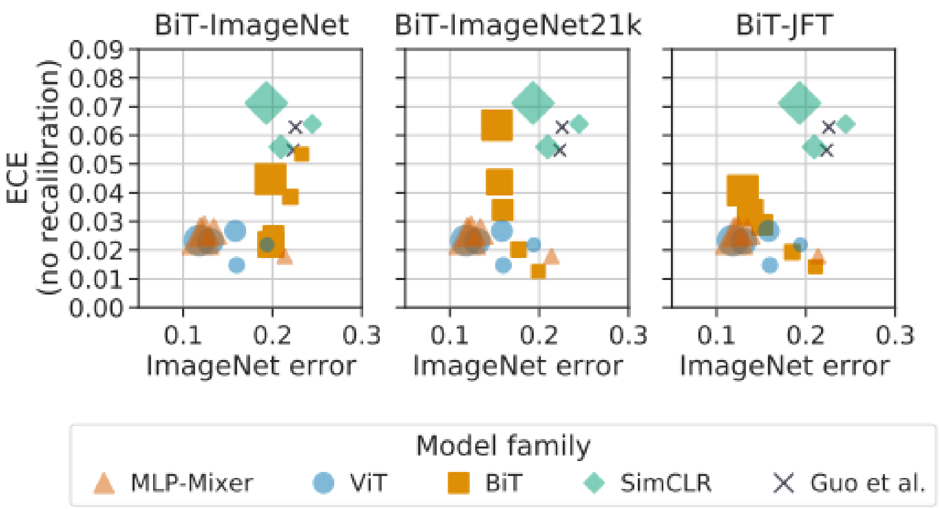
Calibration of (Modern) Deep Neural Networks





❖ 어떤 구조를 가진 이미지 분류 모델이 <u>더 좋은 신뢰도를 가지고 있을까?</u>

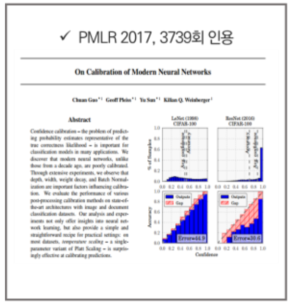
- Convolutional VS Non-Convolutional [Vision Transformer(2021), MLP-Mixer(2021), ⋯, etc.]
- 사전학습의 영향이 있진 않았을까? ➔No. 무거운 모델 기준 데이터를 많이 학습할수록 정확도는 향상되었지만, 신뢰도 성능은 변화 없었음

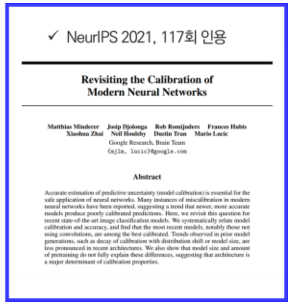


- ImageNet: 1.3M Images
- ImageNet-21k: 12.8M Images
- JFT-300: 300M Images

Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

Calibration of (Modern) Deep Neural Networks

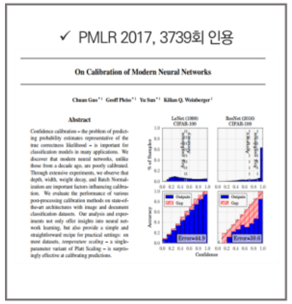


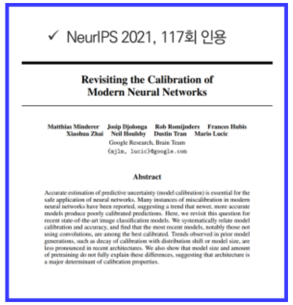


- ❖ Convolutional Model: 모델 사이즈가 커질수록 정확도 성능은 향상되지만, 신뢰도 성능은 떨어지는 추세
- ❖ Non-Convolutional Model: 신뢰도 성능이 전반적으로 좋은 편이며, 다량의 데이터 기반 사전학습 시 정확도까지 우수함

Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

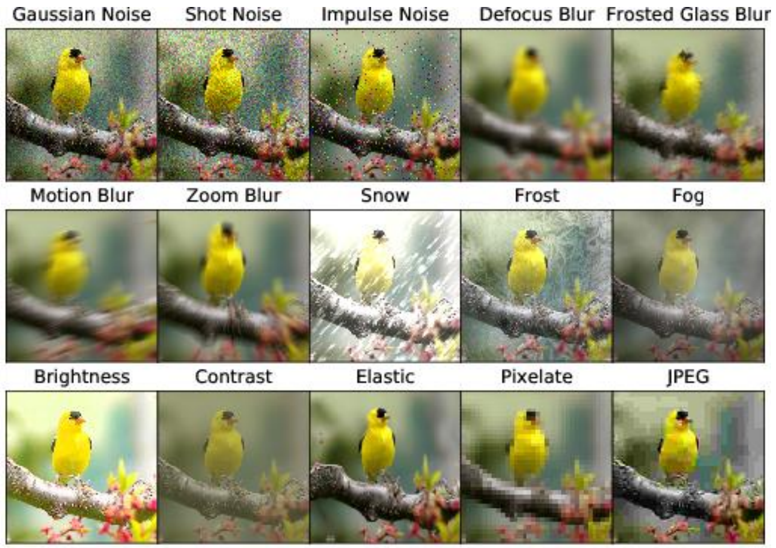
Calibration of (Modern) Deep Neural Networks





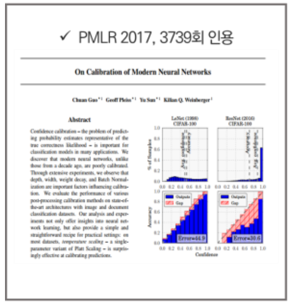
- ❖ Convolutional Model: 모델 사이즈가 커질수록 정확도 성능은 향상되지만, 신뢰도 성능은 떨어지는 추세
- ❖ Non-Convolutional Model: 신뢰도 성능이 전반적으로 좋은 편이며, 다량의 데이터 기반 사전학습 시 정확도까지 우수함
- ❖ Noisy(Corrupted) Data에 대해서는 어떠한 특성을 가지고 있을까?

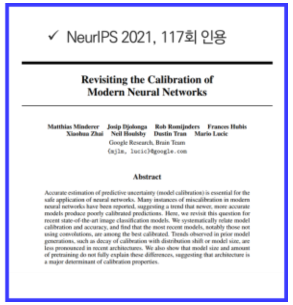
Noisy(Corrupted) Data



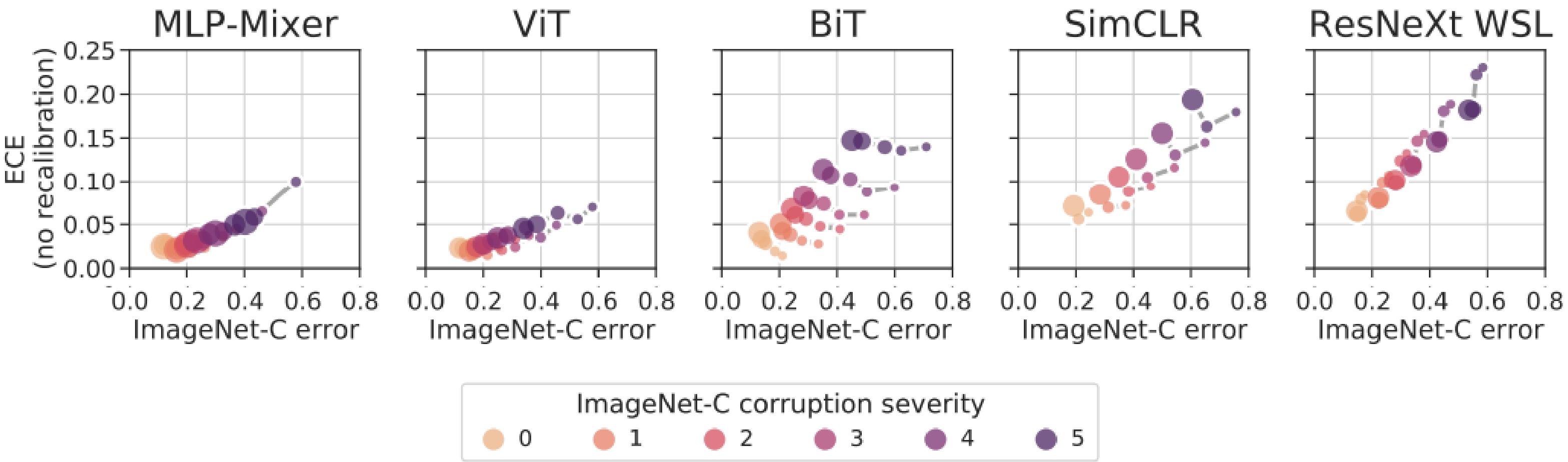
Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

Calibration of (Modern) Deep Neural Networks

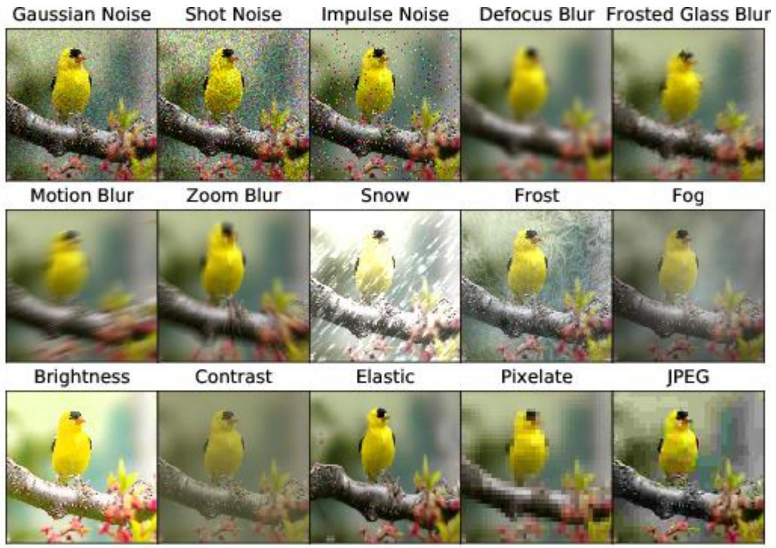




- ❖ Convolutional Model: 모델 사이즈가 커질수록 정확도 성능은 향상되지만, 신뢰도 성능은 떨어지는 추세
- ❖ Non-Convolutional Model: 신뢰도 성능이 전반적으로 좋은 편이며, 다량의 데이터 기반 사전학습 시 정확도까지 우수함
- ❖ Noisy(Corrupted) Data에 대해서는 어떠한 특성을 가지고 있을까?
  - 모든 모델에 대해서, 정확도와 신뢰도가 양의 상관관계를 가지고 있음
  - Non-Convolutional Model의 성능이 전반적으로 노이즈 강도에 대해 Robust함



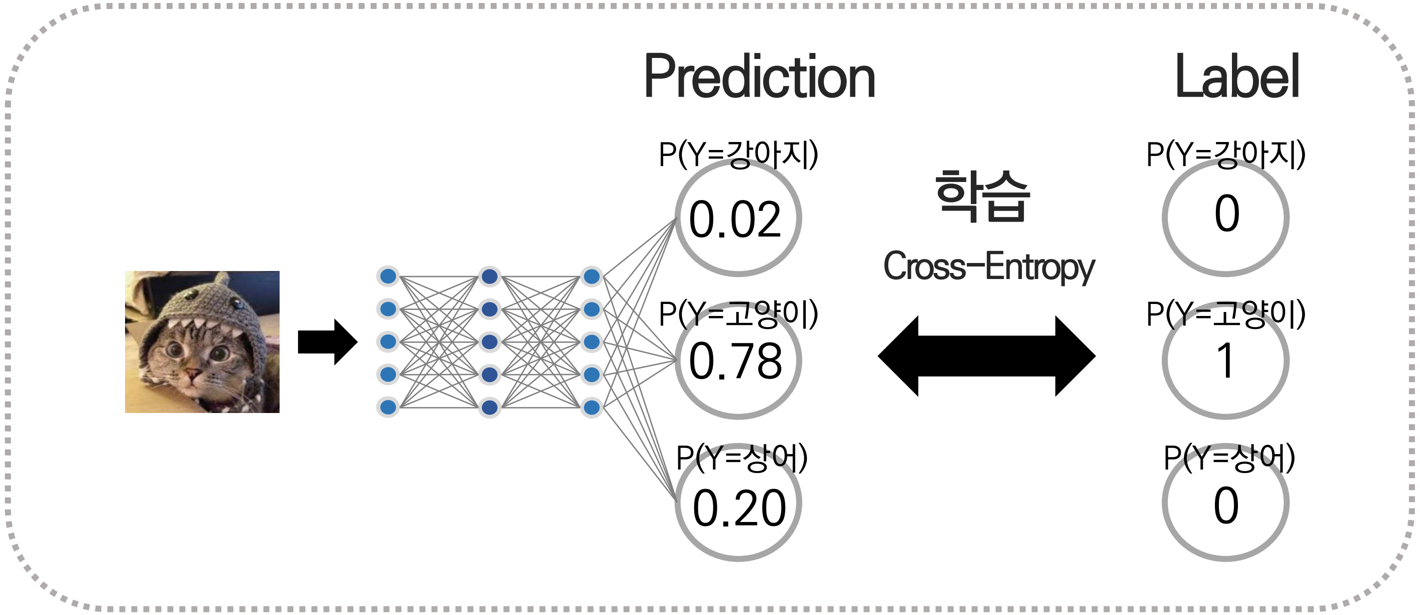
Noisy(Corrupted) Data



Minderer, Matthias, et al. "Revisiting the calibration of modern neural networks." Advances in Neural Information Processing Systems 34 (2021)

3. Improving Calibration of Deep Neural Networks

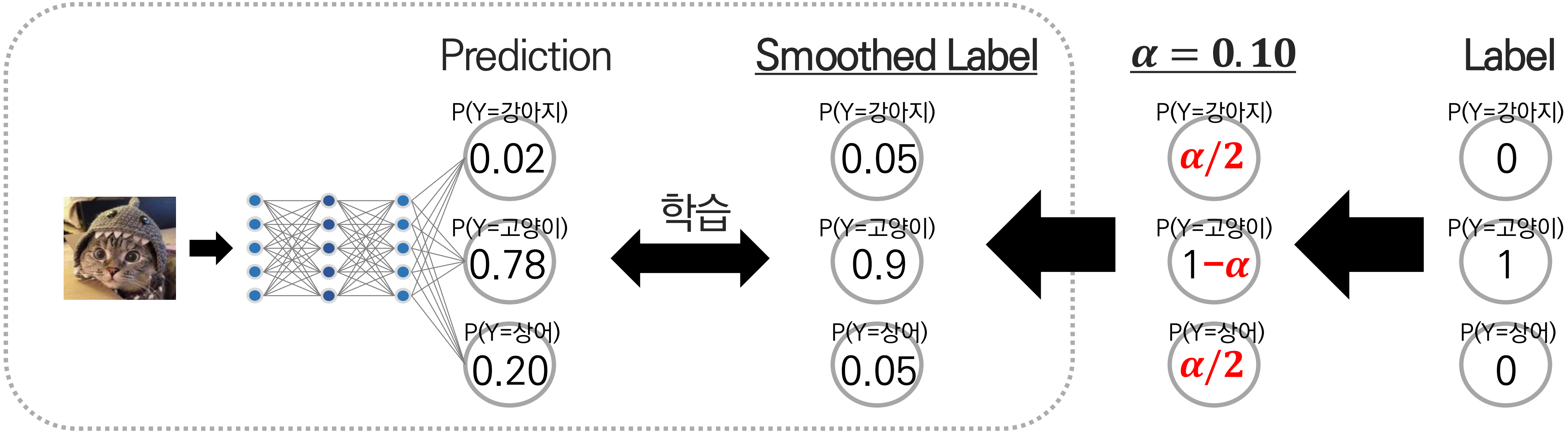
어떤 부분을 개선해야 과대 확신을 하지 않을까?



Improving Calibration of Deep Neural Networks

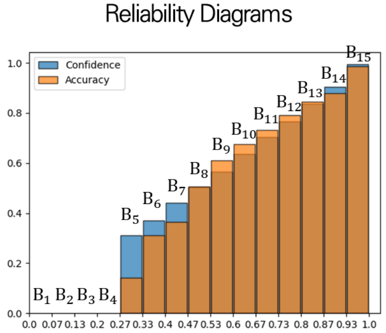
❖ Label에 대한 Smoothing 적용 → Smoothing된 Label을 학습 → 과한 확신 방지

- Smoothing 적용 전 Label: [0,0,0,⋯,1,⋯,0]
- Smoothing 적용 후 Label: [$\frac{\alpha}{K-1}$,$\frac{\alpha}{K-1}$,$\frac{\alpha}{K-1}$,⋯,1$-\alpha$,⋯,$\frac{\alpha}{K-1}$], K= Class 개수



Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

Improving Calibration of Deep Neural Networks

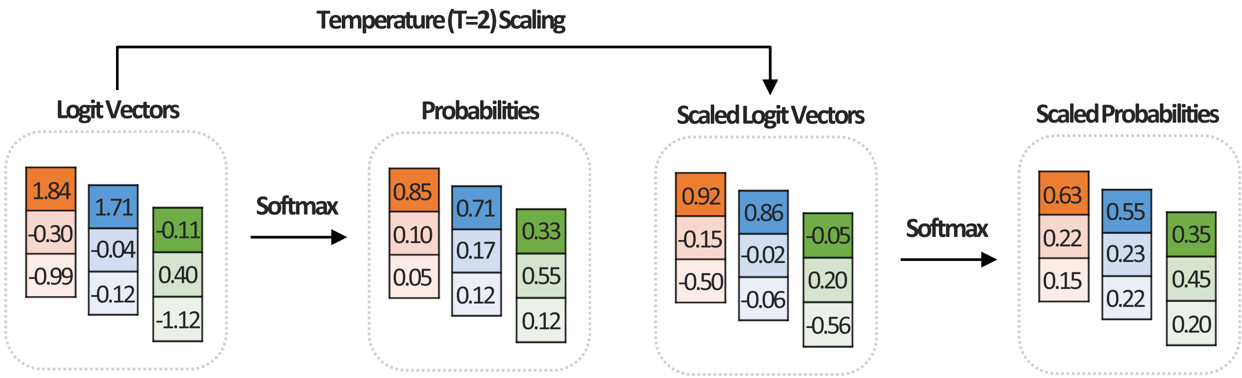


❖ Label에 대한 Smoothing 적용 → Smoothing된 Label을 학습 → 과한 확신 방지

- 간단하고 직관적인 아이디어로 ECE, Reliability Diagram를 효과적으로 개선시킴
- Temperature Scaling: Logit 값을 Temperature로 Scaling하여 Overconfidence를 해결한 비교방법론

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 / α (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



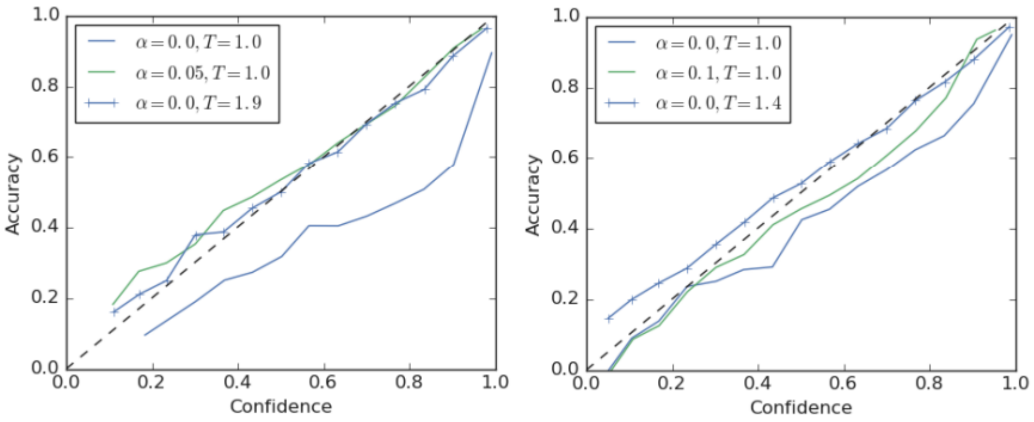


Figure 2: Reliability diagram of ResNet-56/CIFAR-100 (left) and Inception-v4/ImageNet (right).

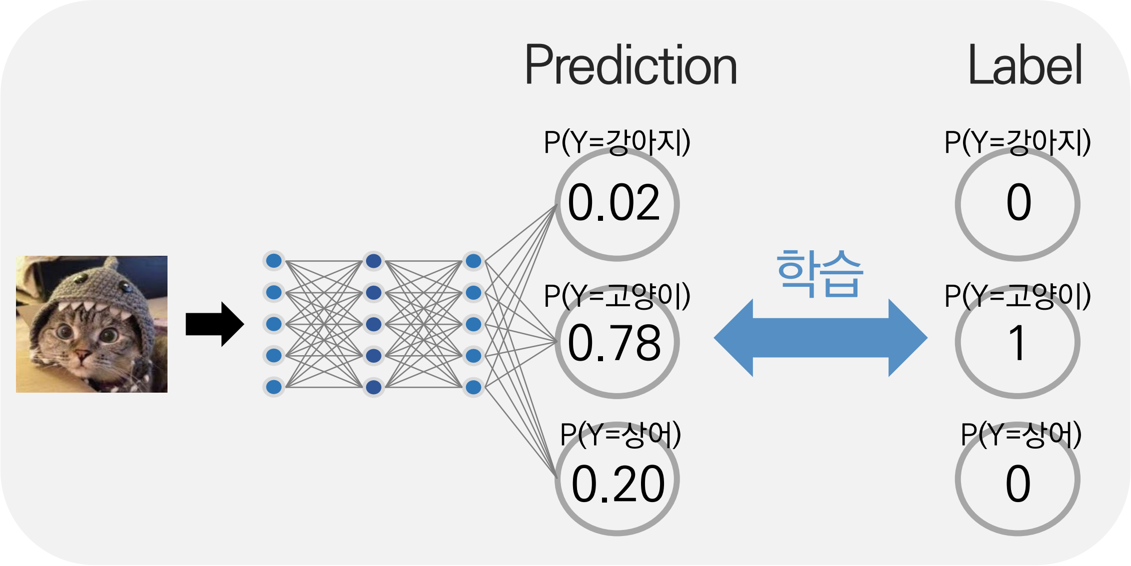
Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

Improving Calibration of Deep Neural Networks

$* \sum_{k=1}^{3} p_{i,k} \, log p_{i,k} = Negative\ Entropy$

❖ 과하게 확신하는 예측 사례에 대하여 패널티 부여 (=엔트로피 정규화)

- 과하게 확신하며 예측했던 Id(2)보단 Id(1)와 같은 예측이 되길
- Negative Entropy Loss Term을 기존 Cross Entropy 손실함수에 추가하여 제약을 가함



$$\boldsymbol{cross\_entropy\_loss(label_i, prediction_i) + \lambda \sum_{k=1}^{3} p_{i,k} \, log p_{i,k}}$$

Id	강아지 $p_{i,k=1}$	고양이 $p_{i,k=2}$	상어 $p_{i,k=3}$	$\sum_{k=1}^{3} p_{i,k} \, log p_{i,k}$
1	0.02	0.78	0.20	−0.5939
2	0.01	0.98	0.01	−0.1190

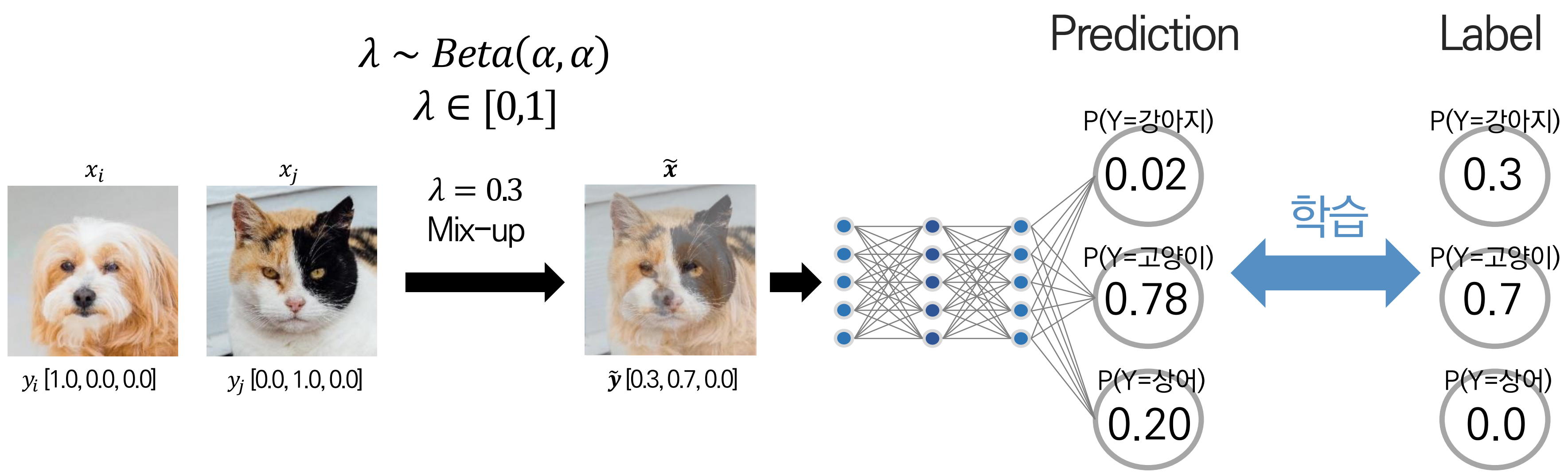
Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

Improving Calibration of Deep Neural Networks

❖ Mix-up 증강 기법 적용을 통한 과한 확신 방지

- 0 과 1로만 구성되어 있는 One-Hot Encoding Label을 학습하는 대신, 0과 1 사이로 Convex combination된 Label을 학습 → 개선
- One-Hot Encoding Label을 학습했던 것이 Overconfidence(Miscalibration)의 주된 문제점임을 의미함

$$\lambda \sim Beta(\alpha, \alpha)$$
$$\lambda \in [0,1]$$

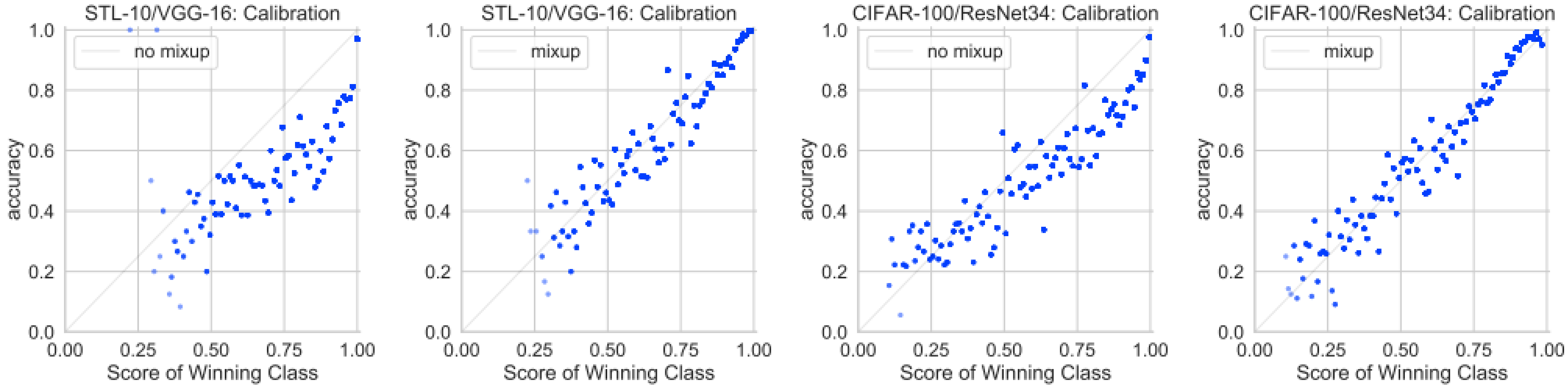


Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

Improving Calibration of Deep Neural Networks

❖ Mix-up 증강 기법 적용을 통한 과한 확신 방지

- No mixup: Y=X 직선 아래에 여러 점들이 찍혀 있음 → 확률값(Score of Winning Class)이 전반적으로 정확도보다 큼 → Overconfidence
- Y=X 직선 아래에 위치하고 있었던 점들이 mixup 적용 후 Y=X 직선 근방으로 옮겨지게 됨 → Improving Calibration



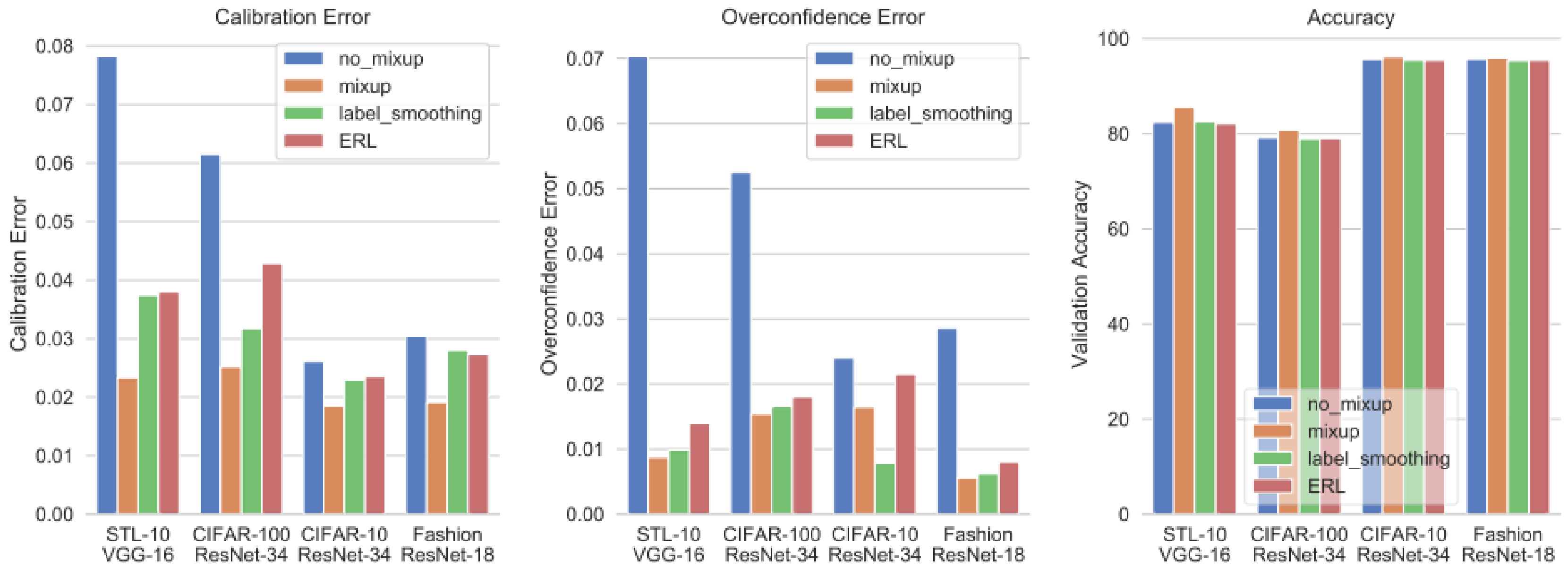
(a) (b) (c) (d)

Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

Improving Calibration of Deep Neural Networks

❖ Mix-up 증강 기법 적용을 통한 과한 확신 방지 → 정확도와 신뢰도가 모두 향상 되는 긍정적 효과

- Calibration Error = $\Sigma_{m=1}^{M} \frac{|\mathrm{B_m}|}{\mathrm{n}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})|$
- Overconfidence Error = $\Sigma_{m=1}^{M} \frac{|\mathrm{B_m}|}{\mathrm{n}} [\mathrm{conf}(\mathrm{B_m}) \times \max(\mathrm{conf}(\mathrm{B_m}) - acc(B_m), 0)]$ / * 정확도가 더 높은 경우는 상관 없는 지표



Müller, Rafael, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?." Advances in neural information processing systems 32 (2019).
Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. "Regularizing neural networks by penalizing confident output distributions". arXiv (2019)
Thulasidasan, Sunil, et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." Advances in Neural Information Processing Systems 32 (2019).

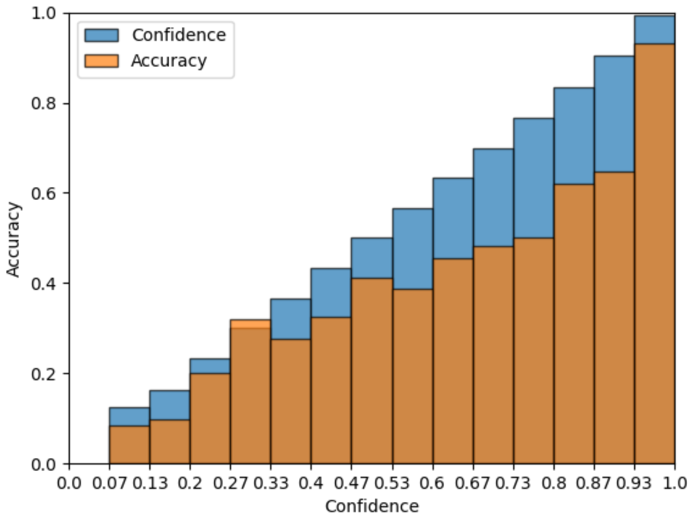
Conclusions

Conclusions

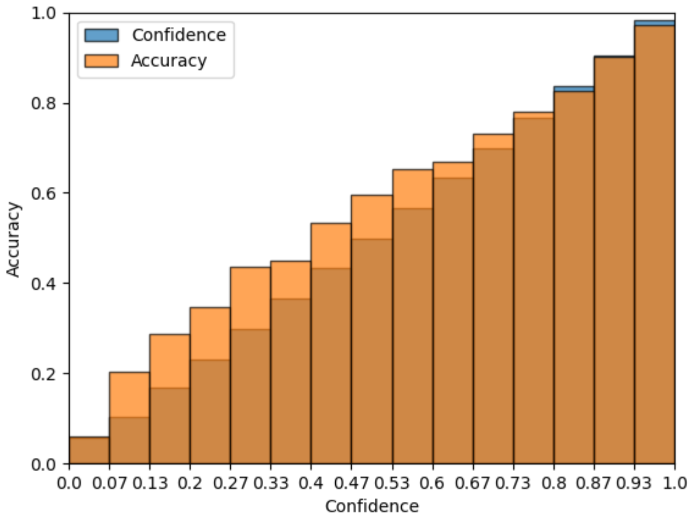
❖ Model Calibration: 모델이 갖고 있는 특성 중 하나로, 모델이 출력한 결과와 실제 정확도가 얼마나 비슷한 지

- Reliability Diagrams, Expected Calibration Error를 통해 Calibration 관점 성능을 평가
- <u>예측 정확도 Performance와 신뢰도 Performance가 모두 우수해야 실제 산업에서 믿고 사용될 수 있음</u>

모델 1: 정확도 80%



모델 2: 정확도 80%



$$\forall p \in [0,1],$$

$$\left|P\left(\hat{Y} = Y \middle| \hat{P} = p\right) - p\right|$$

Conclusions

❖ Convolutional Neural Networks: 정확도 성능이 향상되고 있지만 Calibration 성능은 떨어지고 있는 추세

❖ Non-Convolutional Neural Networks: 정확도와 Calibration 성능이 모두 우수하나, 대규모 사전학습을 필요로 함

- Self-Attention(ViT), Perceptron(MLP-Mixer) 연산을 사용하는 모델이 Well-Calibrated 되는 경향이 있음

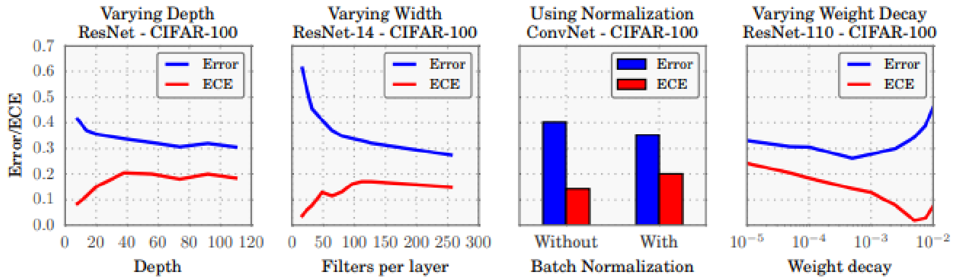
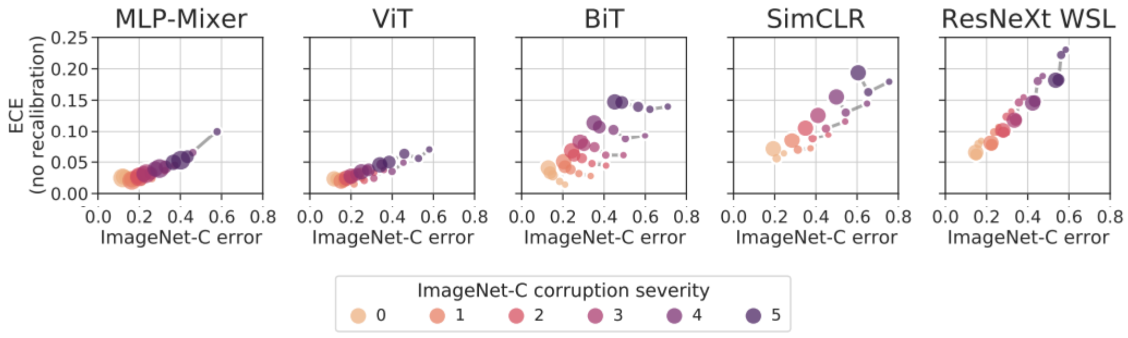


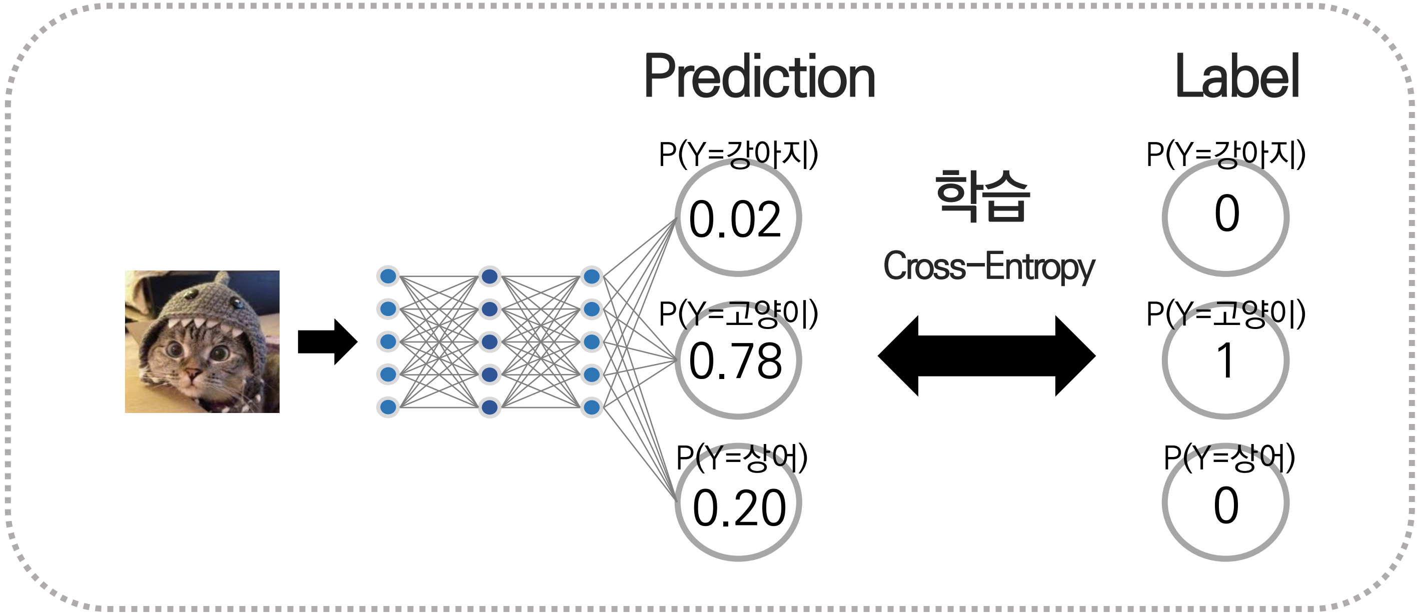
Figure 2. The effect of network depth (far left), width (middle left), Batch Normalization (middle right), and weight decay (far right) on miscalibration, as measured by ECE (lower is better).



Conclusions

- 과하게 확신하는 결과에 패널티를 준다.
  → Entropy Regularization Loss (ERL)
- 과하게 확신하지 않도록 보정된 정답값을 학습한다.
  → Label Smoothing, Mixup

어떤 부분을 개선해야 Calibration 성능을 올릴 수 있을까?



고맙습니다.