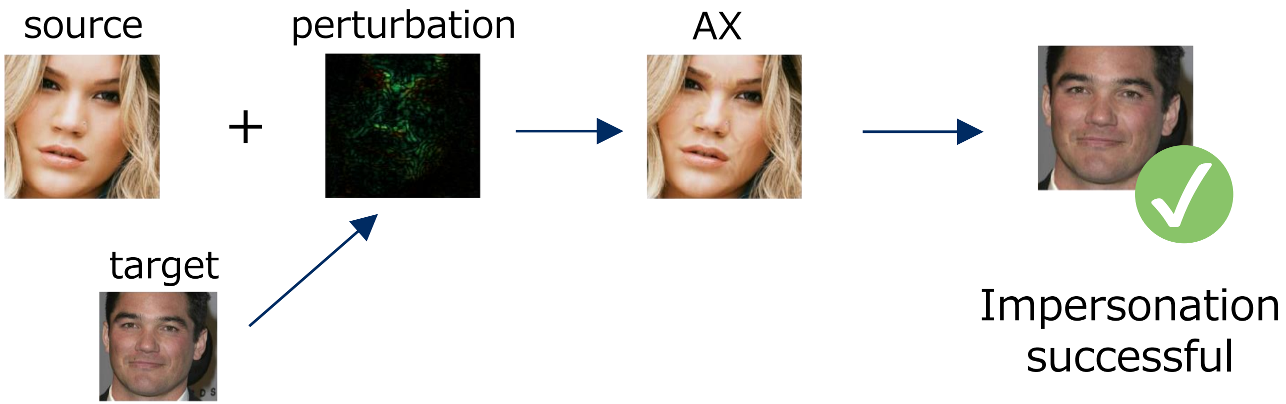
\Orchestrating a brighter world NEC

Adversarial Robustness for Face Recognition: How to Introduce Ensemble Diversity among Feature Extractors?

Takuma Amada[1], Kazuya Kakizaki[1],
Toshinori Araki[1], Seng Pei Liew[1*], Joseph Keshet[2],
Jun Furukawa[3]
[1]NEC Corporation, [2]Bar-Ilan University,
[3]NEC Israel Research Center

Adversarial Examples in Deep Learning

- Deep learning is useful for many applications including security critical services such as face recognition.
- An adversarial example (AX) is inconceivably perturbed input that can deceive deep learning.

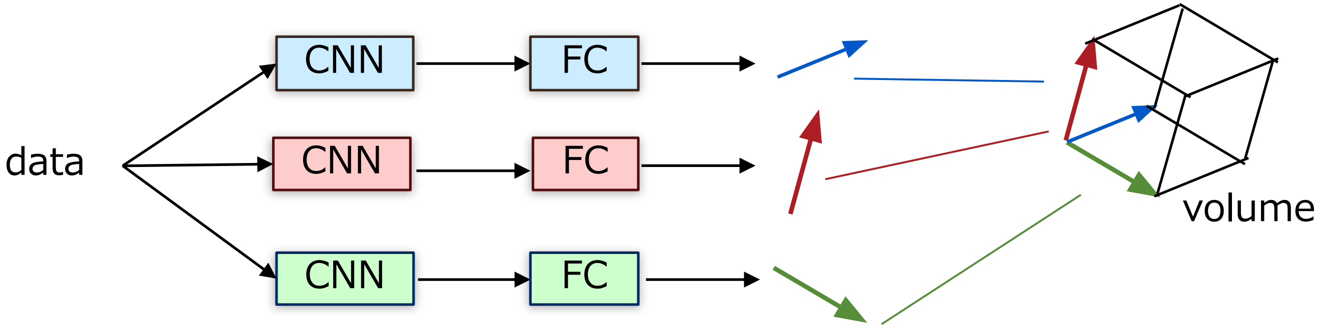


- A deep learning-based security critical service is no longer reliable under attacks by AXs.

Previous Methods for Preventing AXs

Adversarial Training and Ensemble Diversity Promotion are most successful methods for mitigating AXs.

- **Adversarial Training** generates AXs and use them in the training [Zhong and Deng, 2019].
- **Adaptive Diversity Promoting (ADP)** promotes non-maximal predictions of multiple models to be diverse [Pang et al., 2019].

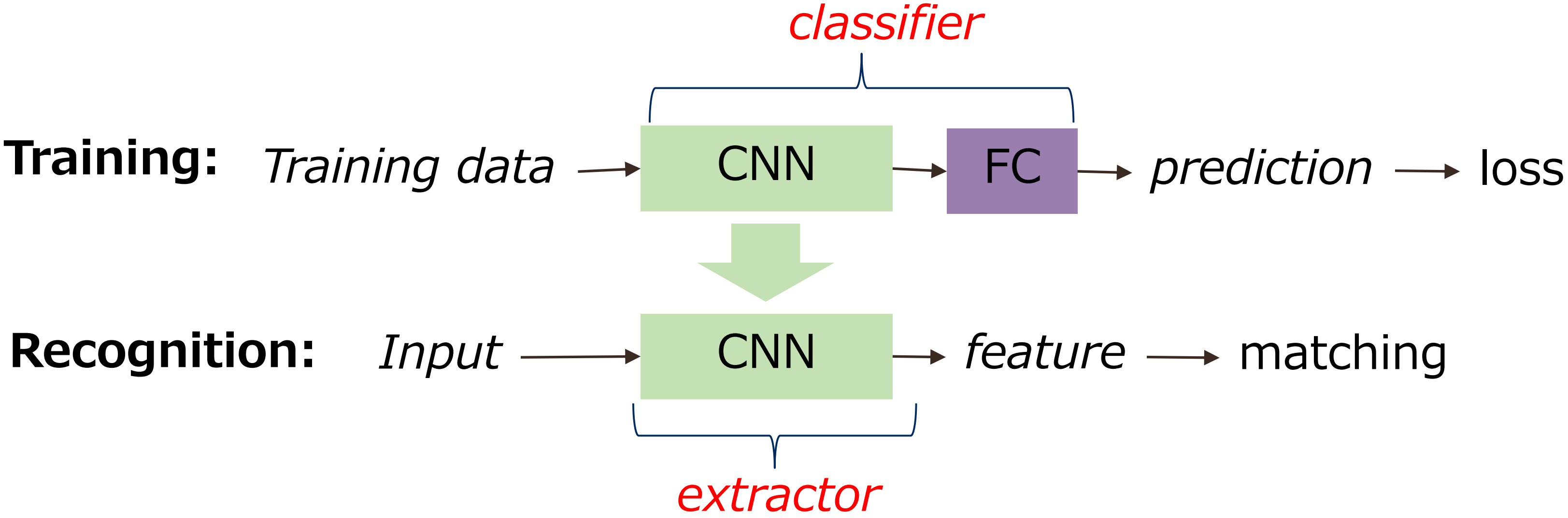


Vectors of non-maximal prediction

No methods are sufficient, and all methods need improvement.

Feature Extractor and Ensemble Diversity

- A feature extractor, once trained, works for new classes without training. Face recognition typically relies on a feature extractor.



- Ensemble diversity has not been applied to feature extractor for preventing AXs

Problem and Goal

Problem

- We applied Adaptive Diversity Promotion (ADP) to face recognition directly.
- Our experiment shows that it neither improved the robustness to AXs nor sacrificed accuracy at all.

Goal

- Introduce the ensemble diversity to feature extractor in the right manner.
- Obtain face recognition that is more robust to AXs.
- Help apply more deep learning to security critical infrastructures.

Our Diagnosis

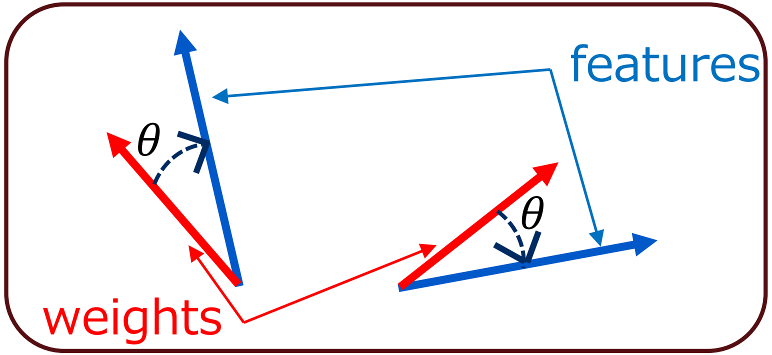
Features are comparable only with respect to weight vectors.

weights

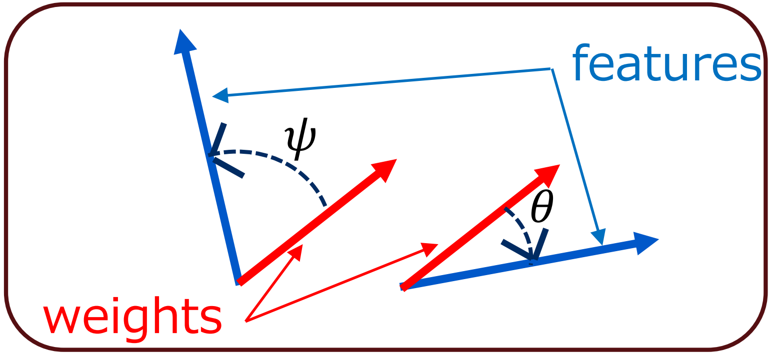
features

$$\mathcal{L}_{CE}(x,y) = \log \frac{e^{W_y \cdot f(x)}}{\sum_{\ell=1}^{n} e^{W_\ell \cdot f(x)}}$$

Direct Application



Our Method



Features are in different directions but not compared to their weights

Features are in different directions compared to their same weights

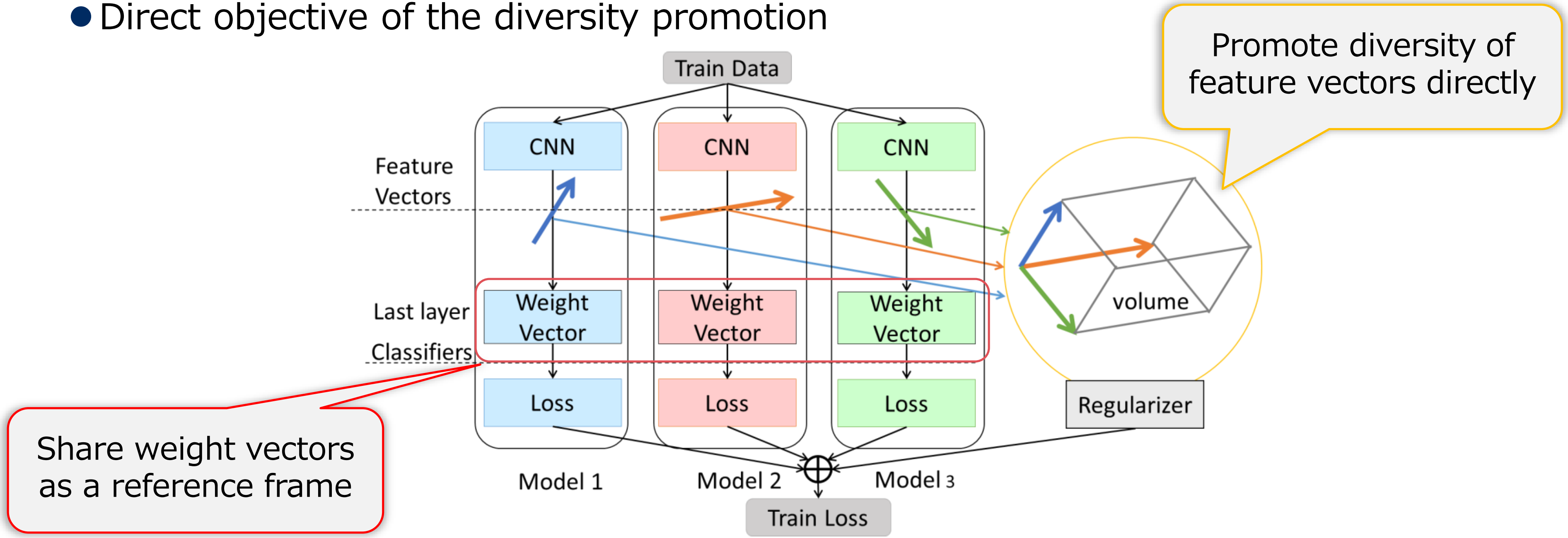
Our Method

Share weight vectors of the final layer by all models.

- The weight vectors provide a measure for the direction of the feature.

Promote the diversity of ensemble features

- Direct objective of the diversity promotion



Experiment - Setup

Data, Model, Attack,

- Training dataset: MS1MV2
- Test dataset: VGG2
- Number of models in ensemble: 3
- Architecture: MobileFaceNet
- Attack: LOTS via I-FGSM, BIM, CW until successful impersonation. No limit for perturbation size

Evaluation Metric: Attack success rate by AXs with different size of input perturbation/feature distance.

Experiment - Accuracy

Test in various data sets

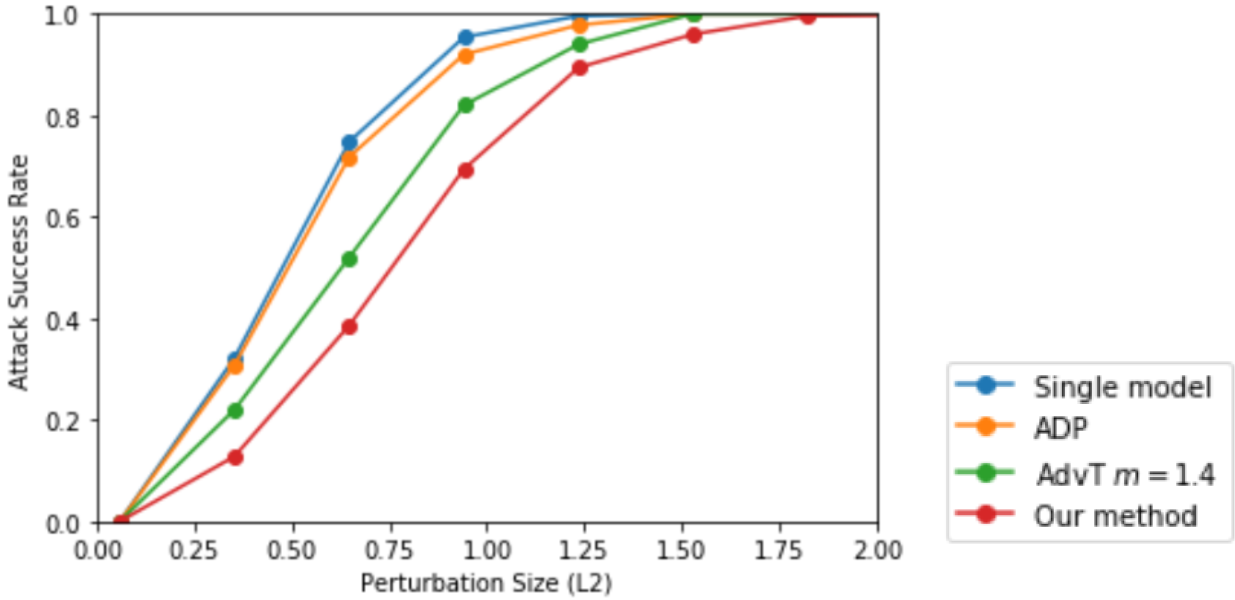
	LFW	CFP-FP	AgeDB-30
Single model	99.30	89.60	94.22
ADP	98.90	86.20	90.52
AdvT*	99.21	**90.80**	94.38
Our method	**99.40**	89.97	**95.15**

*Zhong and Deng, 2019

<u>Our method does not sacrifice accuracy.</u>

Experiment - Robustness against White-box Attacks

- Comparison of “attack success rate” for different “perturbation size”

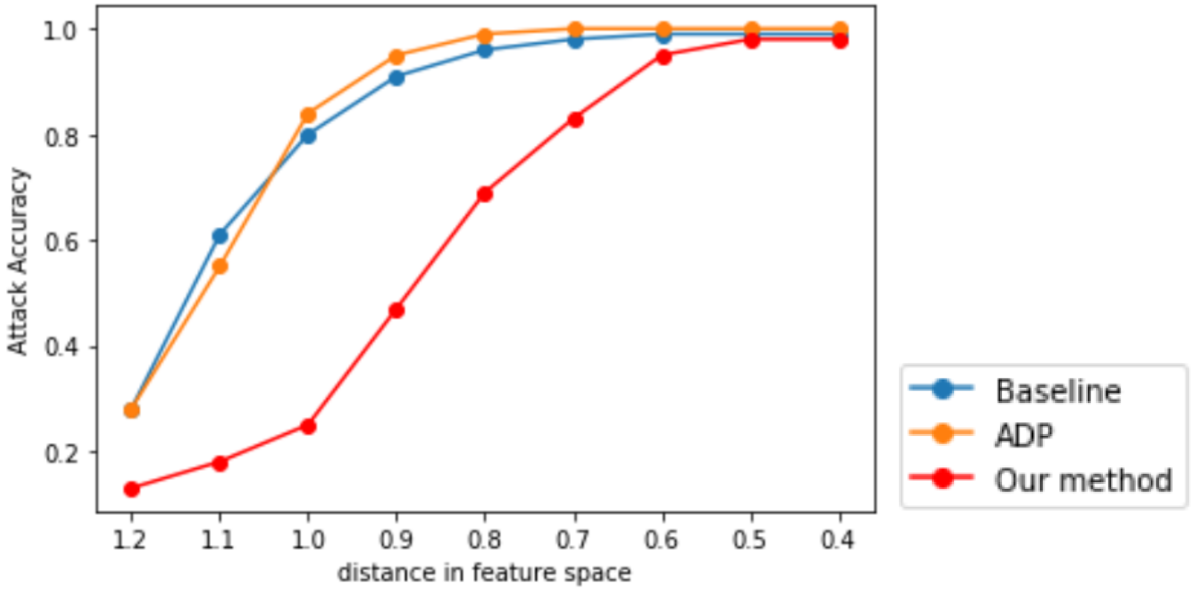


LOTS via I-FGSM

<u>Our method is most robust in the white-box attacks.</u>

Experiment - Robustness against Black-box Attacks

- Comparison of "attack success rate" for AX to the single model with different "distances in feature"



LOTS via I-FGSM

<u>Our method is most robust under black-box attacks</u>

Conclusion

- Feature extractor is essential for face recognitions.
- Promotion of ensemble diversity is one of promising method to prevent AXs. However, we could not apply it to feature extractor directly.
- We presented how to introduce ensemble diversity among feature extractors for robust face recognition without compromising the accuracy.
- Our method shows better robustness compared to adversarial training (although the evaluation is not versatile as others.)