ISITCAP?: Identification of Synthetic Images Through Convolutional Autoencoder Preprocessing

Saisriram Gurajala, Daniel Vogler, Isha Gokhale

Motivation

- Generative image model output can fool humans[1]
- Potential for fraud & misinformation – need to detect fakes

Intro & Related Work

- Some classifiers detect fakes with >90% accuracy when tested on images from the **same generative model as the one that produced their training set[2,3]** ("in-sample")
- But **most classifiers generalize poorly:** accuracy falls on datasets generated by different ("out-of-sample") models [4]
- These classifiers tend to learn specific visual glitches, not general patterns[5]
- Stanciu et al. mitigate poor generalizability of deepfake classifiers through **(1) standard training set augmentations** and **(2) autoencoders[6]** (AECs)
- We apply standard **augmentations[7] and autoencoder to CIFAKE dataset** to test if they (a) **disperse feature importance** & (b) **improve generalizability**

Datasets

- Primary (classifier training): **CIFAKE[8]**
  - 60K from CIFAR-10, **60K diffusion-generated images - same classes**
  - 100K train / 10K val / 10K test split
- Secondary (out-of-sample testing):
  - 2000 images from TinyImageNet[9] (real)
  - 1768 fake images from two different diffusion models[10]

Hypothesis

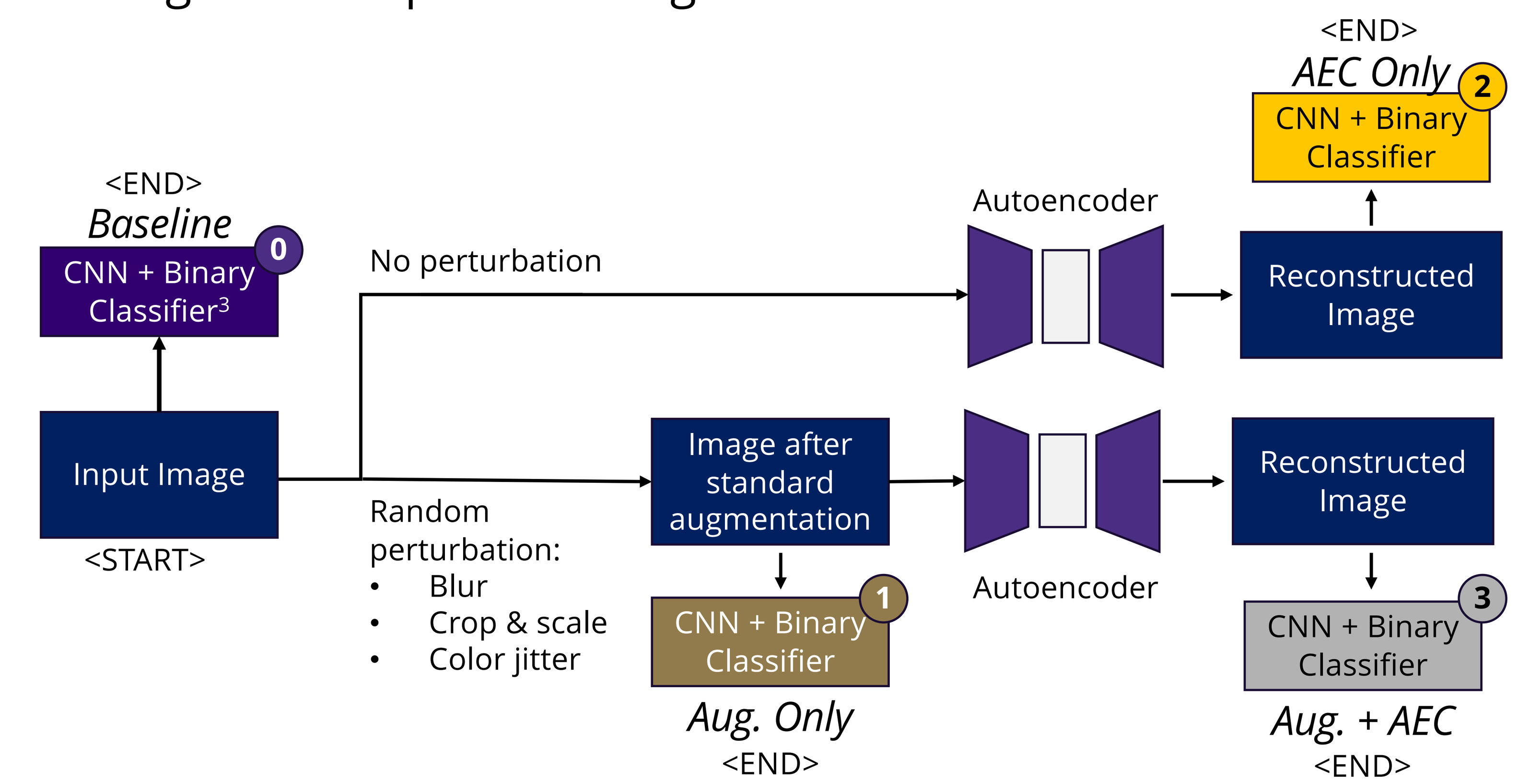
- Out of sample, autoencoder forces CNN to learn general features rather than specific glitches
- As a result, expect autoencoder accuracy > baseline accuracy

References and Notes

1. Dan-Cristian Stanciu and Bogdan Ionescu. Autoencoder- based data augmentation for deepfake detection. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, ICMR '23. ACM, June 2023.
2. Jordan J. Bird and Ahmad Lotfi. Cifake: Image classificationand explainable identification of ai-generated synthetic images, 2023.
3. Bird and Lotfi introduce a classifier architecture that achieves 92% accuracy on CIFAKE-10; our baseline model replicates their architecture
4. Stanciu and Ionescu. Autoencoder- based data augmentation for deepfake detection.
5. Bird and Lotfi. . Cifake.
6. Stanciu and Ionescu. Autoencoder- based data augmentation for deepfake detection.
7. We use 'standard augmentations' to refer to perturbations that directly affect image appearance , **not** resulting from passing the image through the autoencoder. In this study, we use random blur, crop and scale, and color jitter as standard augmentations.
8. Bird and Lotfi. Cifake.
9. Accessed via Kaggle (https://www.kaggle.com/c/tiny-imagenet) – originally introduced by Fei Fei Li, Andrej Karpathy, and Justin Johnson as part of cs231n course at Stanford University (http://cs231n.stanford.edu/)
10. Models are from the FakeImageDataSet in the open source SentryImage Project. Datasets accessed at : https://huggingface.co/datasets/InfImagine/FakeImageDataset. We randomly selected a set of synthetic images for testing. Diffusion Model 1 refers to the dataset produced by their IFv1-CC1M model; Diffusion Model 2 refers to the set from SDv15R-CC1M.
11. Real: CIFAKE, TinyImageNet; Fake: Fake images from CIFAKE plus fake images from diffusion models referenced in (10)

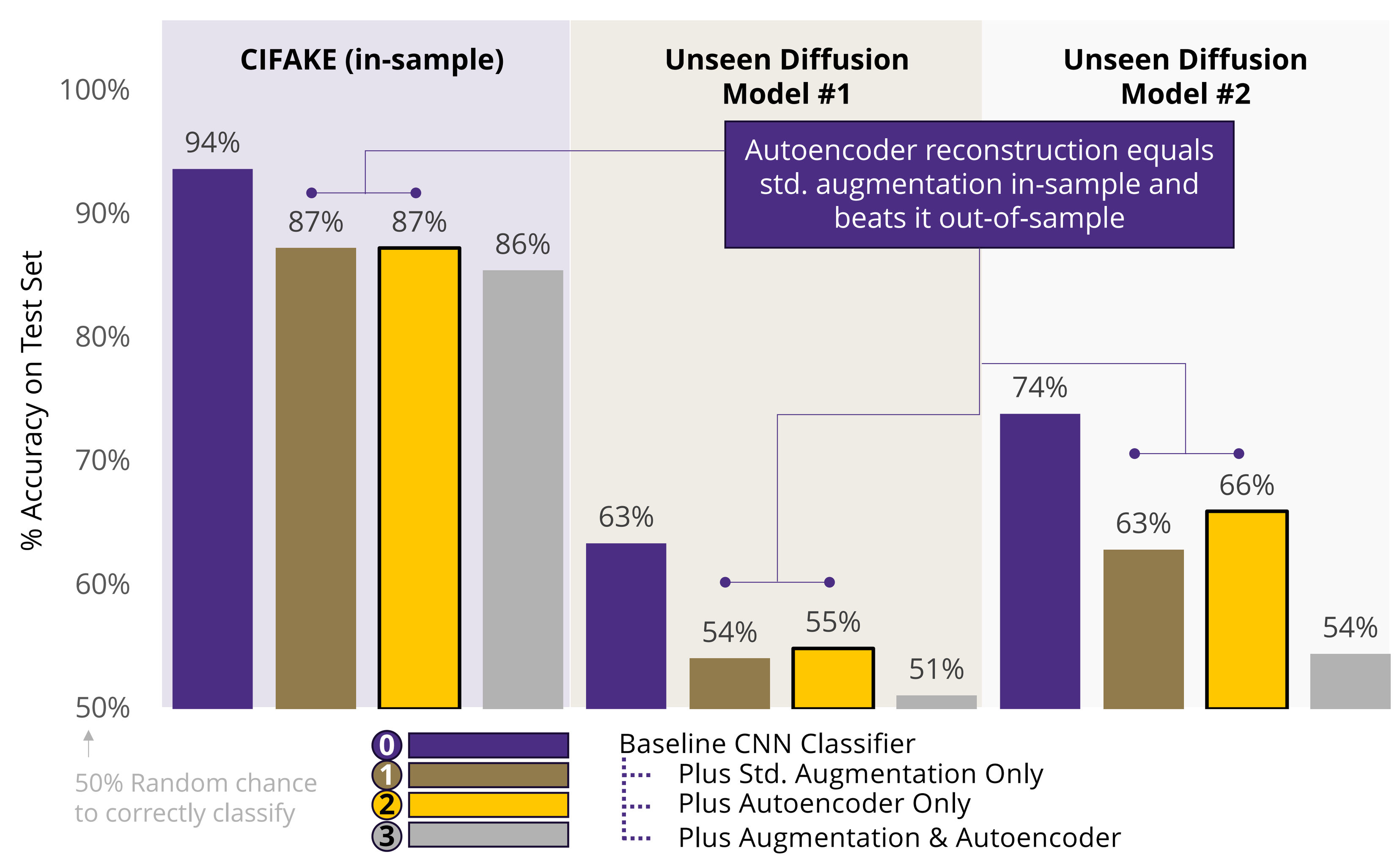
Experiments & Model Training

An image takes 4 paths through our architecture



Results

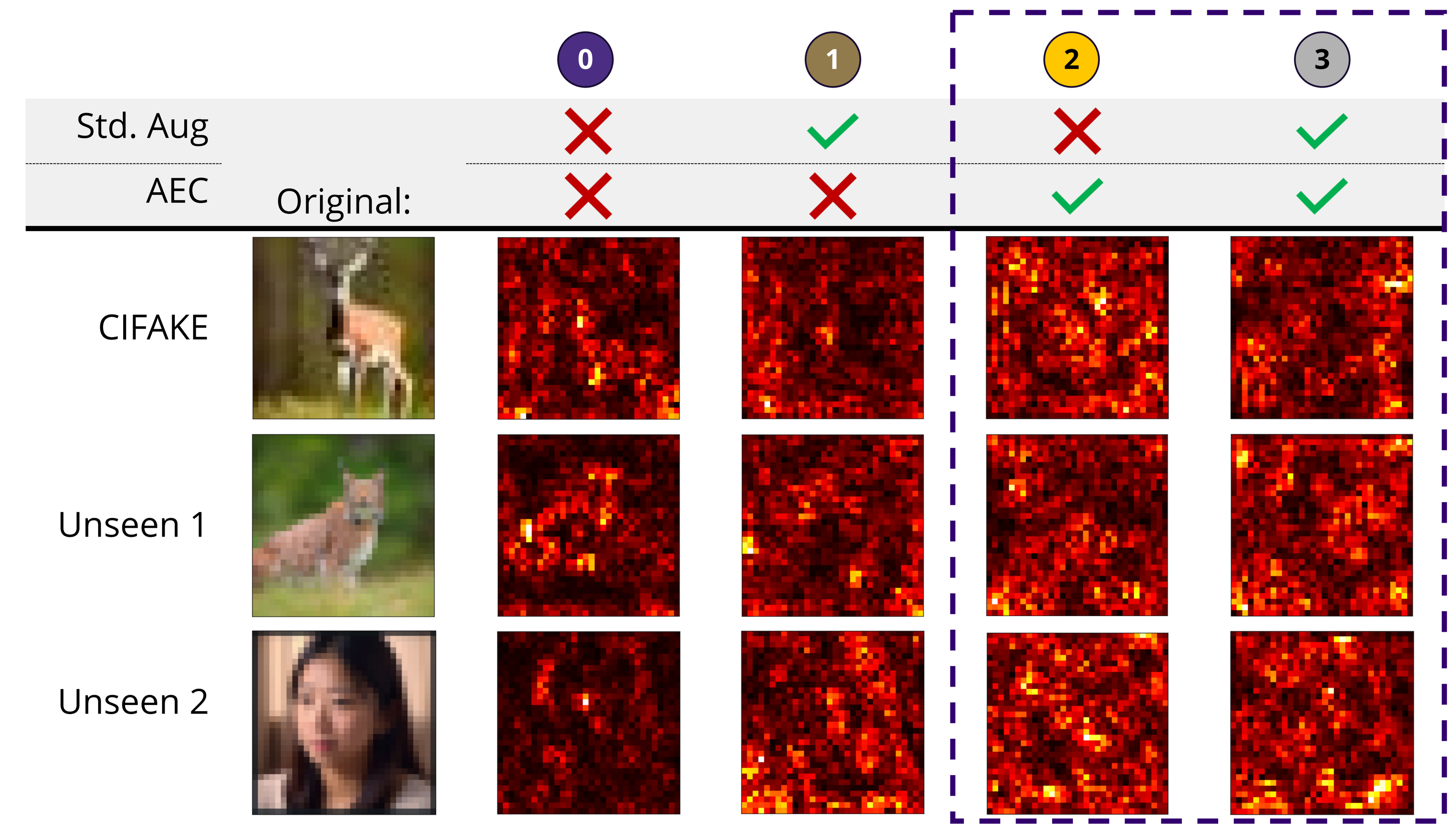
Augmentations & autoencoder preserve classifier performance on CIFAKE, but do not help classifier generalize



- Autoencoder preserves in-sample performance, but **does not help CNN classifier generalize** out of sample
- Conclude that **augmentation cannot offset low dataset diversity**
- To confirm this, next step is to **re-train with more diverse dataset** – expect performance uplift

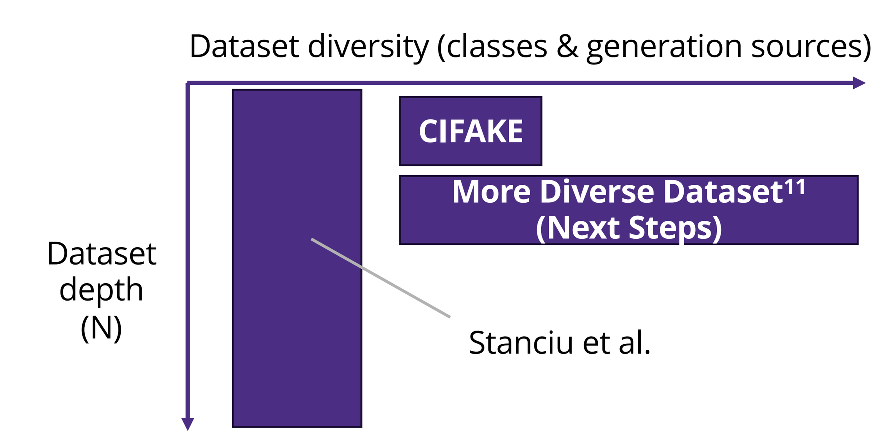
Interpretation

Example Saliency Maps under Test Conditions:



- Across generators, classifiers on unaugmented image focus on **specific visual features** to determine they are synthetic
- **Autoencoder disperses feature importance** more than augmentation alone
- Systematic analysis of image samples confirms effect illustrated by examples

If autoencoder does help CNN learn general features, why isn't out-of-sample accuracy better? Our interpretation:



Conclusion

- **Takeaway:** Training on diverse datasets is key to generalizable classifier; augmentation, including via autoencoder, does not offset low dataset diversity
- **Contribution:** systematic analysis of autoencoder effect on feature learning; thorough characterization of CIFAKE
- **Next Steps:** re-train classifier with on multi-sample dataset and re-test