Human-in-the-loop *mixup*

Katherine M. Collins*, Umang Bhatt, Weiyang Liu, Vihari Piratla, Bradley Love, Adrian Weller
*kmc61@cam.ac.uk

Motivation

- **Synthetic data** is increasingly powering advances in machine learning (ML)[1,2,3]
- Not always clear whether **human perceptual judgments** of synthetically-generated exampes **match the generative process used to create them**
- First step: consider synthetic examples used in *mixup*[4]

Fig 1: Example mixed image

Why care about human percepts of synthetic data generation?

- Further ensure model reliability and trustworthiness
- Realigning may improve downstream model performance
- Help guard against gamification and manipulation

Why *mixup*?

- Simple data generation: access to "ground truth" parameterization
- Powerful technique and popular baseline[5,6,7]
- Cognitive neuroscience suggests likely misalignment[8,9,10]

Problem Setting

- *mixup*[4] is an effective regularizer, which trains on **linear combinations** of examples
- Examples constructed via **data** and **label mixing policies**

$$f(x_i, x_j, \lambda_f) = \lambda_f x_i + (1 - \lambda_f) x_j = \tilde{x}$$

Eq 1: Data mixing policy

$$g(y_i, y_j, \lambda_g) = \lambda_g y_i + (1 - \lambda_g) y_j = \tilde{y}$$

Eq 2: Label mixing policy

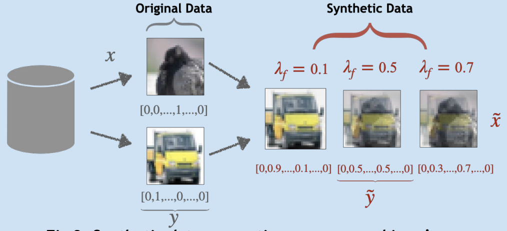

Fig 2: Synthetic data generative process used in *mixup*

- **Our Approach:** Study human perception of the generative process through **two human subject experiment (HSE) paradigms**

HSE 1: What $\tilde{x}$ do humans believe matches a given $\tilde{y}$?

Elicitation (N = 70)

- Elicit **perceived 50/50 point** over 249 pairs of mixed CIFAR-10[11] examples
- Employ different elicitation interfaces
  - **Construct:** press arrow keys to select mixed image
  - **Select-Shuffled:** choose from a shuffled set of mixed images
    - Controls for order effects

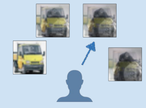

Fig 3: Generic elicitation paradigm

Findings

- In general, humans recover 50/50 mix
- But nuanced picture at individual-level suggests misalignment
- Decent agreement across interfaces

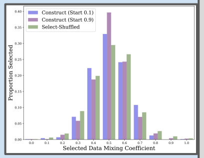

Fig 4: Individual endorsements of the perceived 50/50 point

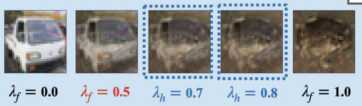

Fig 5: Example consensus misalignment of 50/50

HSE 2: Conditioned on $\tilde{x}$, what do humans perceive to be a good as $\tilde{y}$?

Elicitation (N = 81)

- 2070 mixed images
- Tell people the underlying labels
- Ask to **infer the mixing coefficient**, and **provide their confidence** in estimate

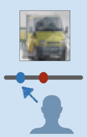

Fig 6: Generic elicitation paradigm

Findings

- Discrepancies elucidated between humans' internal models of synthetically-generated data vs. label mixing policy used in *mixup*
- In aggregate + individual-level misalignment

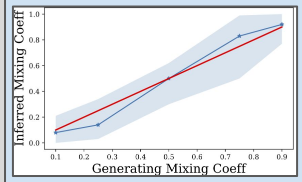

Fig 7: Aggregate human-inferred mixing coeff (blue) vs. *mixup* (red)

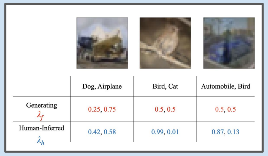

Fig 8: Example synthetic image-level relabelings

Learning with Human Relabelings

- Can we ***align*** **the labeling of mixed images to human perception** to **learn better category boundaries?**

Set-up

- Augment training set with mixed images and constructed labels
- Explore levering **averaged human relabelngs** without confidence
- And **leveraging elicited human confidence** (ω) **to smooth** between a uniform distribution and the averaged human relabeling

Label Type	CE	FGSM	Calibration
Regular (No Aug)	2.02±0.12	13.12±2.65	0.28±0.011
+ Random Labels	2.11±0.13	12.81±2.84	0.24±0.014
+ Uniform Labels	2.16±0.14	12.71±2.79	0.25±0.012
+ *mixup* Labels	1.65±0.11	10.62±2.44	0.23±0.005
+ Ours (Avg Relabelings)	1.78±0.12	11.69±2.90	0.24±0.009
+ Ours (Avg with ω)	**1.48±0.06**	**8.89±1.59**	**0.19±0.001**

Table 1: Evaluating on CIFAR-10H[11,12] holdout, with and without human feedback

Learning with Automatic Label Policy Grounded in Human Inferences

- How can we go beyond the constraint of finite human labelings for an infinite set of possible synthetic examples?
- Category boundaries have diverse structures, many non-linear — can we leverage capture this structure in *mixup* label policy?

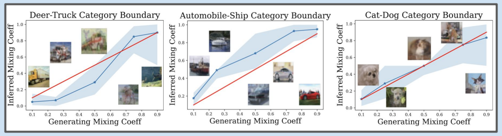

Fig 9: Example inferred mixing coefficient for category pairs

Set-up

- Fit logistic function to each category pair
- Compare learning with transformed mixing coefficient against classicial, full mixup[4]

Label Policy	CE	FGSM	Calibration
mixup	**1.15±0.08**	7.46±2.40	**0.10±0.01**
Human-Fits (Ours)	1.16±0.08	**7.32±2.27**	**0.10±0.01**

Table 2: Comparing automated full relabeling schemes

Takeaways

- Human percepts **not consistently aligned** with data generation used in *mixup*
  - When considering both data and label mixing policy
- Relabeling with human percepts, particularly when **leveraging human confidence**, has potential to improve downstream model performance

References

1. Celso M. de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. 2022. Next-generation deep learning based on simulators and synthetic data. Trends in Cognitive Sciences 26, 2 (2022), 174-187. https://doi.org/10.1016/j.tics.2021.11.008
2. James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data-what, why and how? arXiv preprint arXiv:2205.03257 (2022).
3. Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of big data* 6.1 (2019): 1-48.
4. Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. CoRR, abs/1710.09412.
5. Thulasidasan, S.; Chennupati, G.; Bilmes, J.; Bhattacharya, T.; and Michalak, S. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks.
6. Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. 2022. When and How Mixup Improves Calibration. In International Conference on Machine Learning (ICML).
7. Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. 2022. Pixmix: Dreamlike pictures comprehensively improve safety measures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16783-16792.
8. Jonathan R Folstein, Thomas J Palmeri, and Isabel Gauthier. 2013. Category learning increases discriminability of relevant object dimensions in visual cortex. Cerebral Cortex 23, 4 (2013), 814-823
9. Robert L Goldstone and Andrew T Hendrickson. 2010. Categorical perception. Wiley Interdisciplinary Reviews: Cognitive Science 1, 1 (2010), 69-78.
10. Stevan Harnad. 2003. Categorical perception. (2003).
11. Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical Report.
12. Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. CoRR, abs/1908.07086
13. Battleday, R. M.; Peterson, J. C.; and Griffiths, T. L. 2019. Capturing human categorization of natural images at scale by combining deep networks and cognitive models. CoRR, abs/1904.12690

Presented at The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI