Human-in-the-loop *mixup*

Katherine M. Collins[1*], **Umang Bhatt**[1,2], **Weiyang Liu**[1,3], **Bradley Love**[2,4], **Adrian Weller**[1,2]

[1]University of Cambridge, United Kingdom
[2]The Alan Turing Institute, United Kingdom
[3]Max Planck Institute for Intelligent System, Germany
[4]University College London, UK
[*]kmc61@cam.ac.uk

Abstract

Synthetic data is powering advances in machine learning; however, it is not always clear if synthetic labels are perceptually sensible to humans. We take a step towards understanding human perceptual alignment of the synthetic labels in *mixup*, a powerful regularizer shown to improve model robustness, generalization, and calibration. We find that human perception does not consistently align with the labels traditionally used for synthetic points, and that aligning mixing coefficients with human perception has several advantages.

Training on synthetic data has unlocked tremendous advances in machine learning (Silver et al. 2016; de Melo et al. 2022; Emam et al. 2021; Jordon et al. 2022). However, it is not always clear whether the labels used for synthetic examples align with human perception. Aligning networks could help ensure model reliability and trustworthiness (Nanda et al. 2021; Chen et al. 2022). Therefore, it is worth *verifying* whether synthetic data aligns with human perception, and if not, whether training with *human-relabeled* examples improves model performance.

In this work, we take a step in this direction by focusing on *mixup* (Zhang et al. 2017): a method whereby a model is trained only on synthetic, linear combinations of conventional training examples. We focus on *mixup* for three key reasons. First, the generative process for synthetic examples is very simple. Second, despite this simplicity, *mixup* is a powerful and popular training-time method which has been leveraged to address model fairness (Chuang and Mroueh 2020), improve model calibration (Thulasidasan et al. 2019), and increase model robustness via implicitly regularizing the form of category boundaries learned (Zhang et al. 2020), and is frequently used as a strong benchmark for many new data augmentation and regularization techniques (Hendrycks et al. 2019, 2022). Third, prior work in human categorical perception – revealing that humans show nonlinear "warping" effects along category boundaries (Harnad 2003; Folstein, Palmeri, and Gauthier 2013; Goldstone and Hendrickson 2010) – leads us to believe that humans *will* differ in their percepts from the linear category boundaries encouraged by *mixup*.

To that end, we consider whether *mixup* labels match human perception, and if not, how the labeling scheme can be improved to better align with human intuition and potentially enhance model performance. We see our work as taking a small step in the exciting direction of a human-centric perspective on synthetic data used to train ML systems.

Background

We first review *mixup* (Zhang et al. 2017) and explicate the recipe by which synthetic examples are created. We assume access to a finite set of N samples $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$. *mixup* training (Zhang et al. 2017) consists of constructing synthetic training examples $(\tilde{x}, \tilde{y})$ via linear combinations of pairs of the training observations $(x_i, y_i), (x_j, y_j)$ for $i, j \in [1, N]$, corresponding to the following data and label mixing functions:

$$\text{Data Mixing: } f(x_i, x_j, \lambda_f) = \lambda_f x_i + (1 - \lambda_f) x_j = \tilde{x},$$
$$\text{Label Mixing: } g(y_i, y_j, \lambda_g) = \lambda_g y_i + (1 - \lambda_g) y_j = \tilde{y},$$

where λ_f and λ_g are defined as the ***data mixing coefficient*** and ***label mixing coefficient***, respectively. We refer to the combined images x_i, x_j and their labels y_i, y_j as the ***endpoints***. For a specified mixing coefficient λ, we denote the resultant image as $\tilde{x}$. *mixup* typically assumes $\lambda_f = \lambda_g$. instead decouple the data and label mixing functions.

Mixing Coefficient from Human Perception

We consider a generalized *mixup* where the data and label mixing functions can have different coefficients. We are particularly interested to find the intrinsic label mixing coefficient λ_g that best matches human perception. We employ crowdsourcing where human annotators estimate λ_g under various settings (which we refer to as λ_h for human).

Task We recruit $N = 33$ participants from Prolific (Palan and Schitter 2018) to infer λ_g for a mixed image. Participants are told the endpoint labels of the images mixed and asked to indicate their inference (λ_h) and confidence (ω) in this inference via sliders. We follow Prelec (2004); O'Hagan et al. (2006); Chung et al. (2019) and ask participants to respond how they think others would respond. Mixed images are constructed by combining two `CIFAR-10` (Krizhevsky 2009) images, with λ_f drawn from $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. A total of $N = 810$ mixed images were each seen by at least two different participants.

	Dog, Airplane	Bird, Cat	Automobile, Bird
Generating λ_f	0.25, 0.75	0.5, 0.5	0.5, 0.5
Human-Inferred λ_h	0.42, 0.58	0.99, 0.01	0.87, 0.13

Figure 1: Example of averaged human inferences (λ_h) over mixed examples, compared against the generating mixing coefficient λ_f.

Results On aggregate, participants recover close to the generating coefficient when considering the median over all images for a given mixing coefficient. However, error bars are wide and a closer look at how individual images are relabeled (see Fig. 1) reveals significant deviations which suggest human perception is *not* consistently aligned with the mixing coefficient. We therefore think calibration at the aggregate-level is likely due to averaging effects which may cancel out differences in participants' percepts. Inspecting the inferred mixing coefficient between particularly classes – as in Fig. 2 – reveals marked differences between human percepts and the labeling policy traditionally used in *mixup*. We also find that participants' *confidence* in their inferred mixing coefficients somewhat tracks the degree of ambiguity of the original images that are combined.

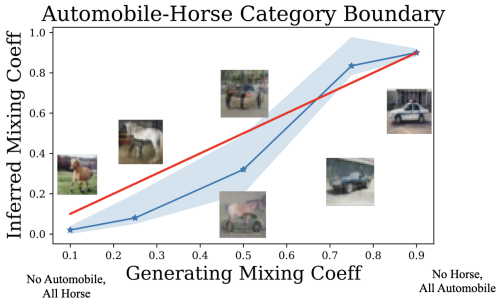

Figure 2: Hand-picked extracted "category boundary" from elicited inferences. People diverge from linearity in the generating coefficient. λ_f is depicted against λ_h (blue) and classical λ_g (red).

Learning with Human Relabelings

Given we find that human perception does not uniformly align with the traditional, linear target mixing target policy used in *mixup* at an individual level, we hypothesize that incorporating the human-elicited relabelings instead of the default *mixup* labels may improve model performance.

Setup We train a PreAct ResNet-18 (He et al. 2015) over 7,000 regular `CIFAR-10` combined with the 810 synthetically mixed images where we vary the labels. While we would ideally study human relabelings for every synthetic image that could generated with f, we only have labels for a small subset and therefore instead compare using our labels versus the traditional *mixup* labels over a *finite, augmenting set* of the combined images. 5 seeds are run per variant.

Labeling Scheme	CE	Calib	FGSM
Regular (No Aug)	2.25	0.29	14.53
+ *mixup* Labels	2.19	0.28	14.35
+ Ours (Agg, λ_h Only)	2.27	0.28	15.07
+ Ours (Sep, λ_h Only)	2.01	0.27	13.33
+ Ours (Agg, λ_h with ω)	2.09	0.27	14.01
+ Ours (Sep, λ_h with ω)	**1.83**	**0.24**	**11.81**

Table 1: Average performance of models trained on different labels for the N=810 synthetic augmenting images. Regular (No Aug) does not train on any synthetically combined *mixup* images.

Our Label Varieties We consider four variants of our labels: {average over participants (Agg), separate labels per participant (Sep)} x {just λ_h , λ_h with ω}. For varieties where human confidence is used, we smooth the label based on an exponentially-decaying transformation (i.e., smoothing = γ^{ω}) of the predicted confidence; here, $\gamma = 0.005$.

Evaluation We evaluate a suite of metrics over 3,000 examples from `CIFAR-10H`, a dataset containing labels from many humans over the `CIFAR-10` test set (Peterson et al. 2019). We compare: cross entropy between the model-predicted and the human-derived label distributions (CE), model calibration following (Hendrycks et al. 2022) and robustness to the Fast Gradient Sign Method (FGSM) adversarial attack (Goodfellow, Shlens, and Szegedy 2014).

Results Table 1 reveals that labeling mixed examples with human relabelings has the potential to improve model generalization, calibration, and robustness over the traditional synthetic labels used in *mixup*. In particular, we find that: 1) learning on separated, non-aggregated labels is beneficial, highlight the importance of capturing and maintaining inter-annotator differences in machine learning datasets (Prabhakaran, Davani, and Diaz 2021; Uma, Almanea, and Poesio 2022; Díaz et al. 2022), and 2) human confidence can be leveraged to construct more potent supervisory signals, indicating that studies aimed at aligning synthetic data to human percepts could benefit from also capturing and representing human uncertainty (Collins, Bhatt, and Weller 2022).

Conclusions and Future Work

While we acknowledge that our results are early and ought to be scaled – we find that the synthetic examples classically used in *mixup* may differ in fundamental ways from human perception, which if altered to align with individual human percepts (adjusted by human confidence), have potential to improve model performance. Our work also motivates the design of automated relabeling procedures for synthetic examples which leverage elicited human data (e.g., training a model to predict a likely human's mixing coefficient) to sidestep inherent issues with scaling human annotation over the (infinite) space of possible synthetic examples.

References

Chen, V.; Bhatt, U.; Heidari, H.; Weller, A.; and Talwalkar, A. 2022. Perspectives on Incorporating Expert Feedback into Model Updates. *arXiv preprint arXiv:2205.06905*.

Chuang, C.-Y.; and Mroueh, Y. 2020. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*.

Chung, J. J. Y.; Song, J. Y.; Kutty, S.; Hong, S. R.; Kim, J.; and Lasecki, W. S. 2019. Efficient Elicitation Approaches to Estimate Collective Crowd Answers. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Collins, K. M.; Bhatt, U.; and Weller, A. 2022. Eliciting and Learning with Soft Labels from Every Annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, volume 10.

de Melo, C. M.; Torralba, A.; Guibas, L.; DiCarlo, J.; Chellappa, R.; and Hodgins, J. 2022. Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, 26(2): 174–187.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Emam, Z.; Kondrich, A.; Harrison, S.; Lau, F.; Wang, Y.; Kim, A.; and Branson, E. 2021. On The State of Data In Computer Vision: Human Annotations Remain Indispensable for Developing Deep Learning Models. *CoRR*, abs/2108.00114.

Folstein, J. R.; Palmeri, T. J.; and Gauthier, I. 2013. Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4): 814–823.

Goldstone, R. L.; and Hendrickson, A. T. 2010. Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1): 69–78.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Harnad, S. 2003. Categorical perception.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

Hendrycks, D.; Zou, A.; Mazeika, M.; Tang, L.; Li, B.; Song, D.; and Steinhardt, J. 2022. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16783–16792.

Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S. N.; and Weller, A. 2022. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

Nanda, V.; Majumdar, A.; Kolling, C.; Dickerson, J. P.; Gummadi, K. P.; Love, B. C.; and Weller, A. 2021. Exploring Alignment of Representations with Human Perception. *CoRR*, abs/2111.14726.

O'Hagan, A.; Buck, C. E.; Daneshkhah, A.; Eiser, J. R.; Garthwaite, P. H.; Jenkinson, D. J.; Oakley, J. E.; and Rakow, T. 2006. *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: John Wiley.

Palan, S.; and Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9617–9626.

Prabhakaran, V.; Davani, A. M.; and Diaz, M. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, 133–138.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *science*, 306(5695): 462–466.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587): 484–489.

Thulasidasan, S.; Chennupati, G.; Bilmes, J.; Bhattacharya, T.; and Michalak, S. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks.

Uma, A.; Almanea, D.; and Poesio, M. 2022. Scaling and Disagreements: Bias, Noise, and Ambiguity. *Frontiers in Artificial Intelligence*, 5.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond Empirical Risk Minimization. *CoRR*, abs/1710.09412.

Zhang, L.; Deng, Z.; Kawaguchi, K.; Ghorbani, A.; and Zou, J. Y. 2020. How Does Mixup Help With Robustness and Generalization? *CoRR*, abs/2010.04819.

Acknowledgments

We thank (alphabetically) Alan Clark, Benedict King, Tuan Anh Le, Vihari Piratla, Joshua Tenenbaum, Richard E. Turner, Vishaal Udandarao, and Catherine Wong for helpful discussions. We also thank our reviewers for very helpful feedback on our manuscript.

KMC is supported by a Marshall Scholarship. UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI), and from the Mozilla Foundation. B.L. acknowledges support under the ESRC grant ES/W007347,

and AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI.