An Interactive Human-Machine Learning Interface for Collecting and Learning from Complex Annotations

Jonathan Erskine [1] **Matt Clifford** [1] **Alex Hepburn** [2] **Raúl Santos-Rodríguez** [2]

Abstract

Human-Computer Interaction has been shown to lead to improvements in machine learning systems by boosting model performance, accelerating learning and building user confidence. In this work, we propose a human-machine learning interface for binary classification tasks with the goal of allowing humans to provide richer forms of supervision and feedback that go beyond standard binary labels as annotations for a dataset. We aim to reverse the expectation that human annotators adapt to the constraints imposed by labels, by allowing for extra flexibility in the form that supervision information is collected. For this, we introduce the concept of task-oriented meta-evaluations and propose a prototype tool to efficiently capture the human insights or knowledge about a task. Finally we discuss the challenges which face future extensions of this work.

1. Introduction

Artificial intelligence (AI) has undoubtedly made remarkable progress in recent years, with the development of various sophisticated tools and models. However, it is important to recognize that this expansion does not necessarily imply a proportionate step towards an advanced form of artificial intelligence that has ability to understand, learn, and apply knowledge across various domains. Instead, what we are witnessing is the proliferation of specialized tools, each designed to tackle specific tasks with remarkable proficiency. For example, Large langauge models (LLMs) excel in their designated domains (Wei et al., 2022), but their exceptional performance is often reliant on the availability of large datasets (Devlin et al., 2019) and their capabilities are often limited when confronted with unfamiliar or unforeseen challenges (Collins et al., 2022). In this paper we ask if the twin issues of low data efficiency and poor model generalisation can be alleviated through human-machine interaction by enabling human annotators to evaluate model behaviour and increase the complexity of individual annotations where required; can we reduce the number of annotations required while ensuring we do not over-fit to the training data and/or learn spurious correlations?

Where simple statistical evaluations such as model accuracy or mean-squared error cannot always capture whether spurious correlations have been learned, higher-level human investigations can often reveal insights about model behaviour which are not immediately apparent from the results, such as bias or failure to perform in critical regions. We call this a *meta-evaluation* and aim to demonstrate how this type of evaluation can be used by an annotator to introduce knowledge into the machine learning pipeline.

In online settings such as Active Learning, this is addressed by the machine learning algorithm communicating with a (human) oracle who returns a new label for each query example, usually selected to reduce a form of uncertainty (Prince, 2004). However, the flexibility for encoding supervision information provided by labels might be too limited for some tasks and might not be adapted to the skills of the annotators. Formulations like Machine Teaching move in this direction by allowing humans to select groups of samples that effectively describe concepts and patterns (Zhu, 2015).

In this work, take an alternative view: rather than incurring additional costs during training by querying a human oracle, we offer the oracle the flexibility to enrich the supervision of the data in the form of more *complex annotations*. We present a prototype interface which enables humans to provide relevant contextual information in between training epochs, where they deem necessary, adapting the loss function to learn from these annotations during training.

1.1. Meta-Evaluation

We propose the term *meta-evaluation* to encapsulate any human evaluations that consider the machine learning pipeline from a broader perspective (beyond conventional metrics)

[1]Department of Computer Science, University of Bristol, Bristol, United Kingdom [2]Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom. Correspondence to: Jonathan Erskine <jonathan.erskine@bristol.ac.uk>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

to understand the underlying mechanisms driving model decisions. These meta-evaluations might be task-dependent, and could be as simple as visualising a dataset to determine any redundant or noisy features, or more complex, such as defining a set of tests to determine bias and fairness in model outputs. Put plainly, we are referring to human evaluations where we look at the bigger picture.

Figure 1 illustrates a simple meta-evaluation for a two-dimensional binary classification task. We call this a *meta-evaluation* as it requires insight from a human observer, instead of being objectively quantified in the model results.

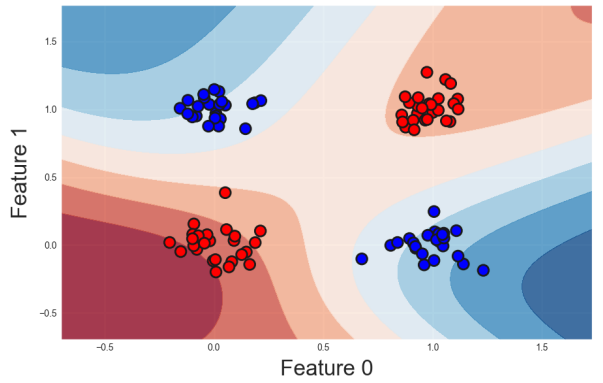

Figure 1. Meta-evaluation in the context of low data availability. We decrease the number of available samples from (a) $n = 100$ to (b) $n = 10$ and visualise the model predicted probabilities across the feature space after a fixed number of training epochs. For the reduced dataset (b) it is visually apparent that the under-sampled region in the top right of the feature space has led to under-fitting of the red class.

1.2. Complex Annotations

Having evaluated a model and determined that improvements can be made, we explore how to communicate our improvement strategy by providing contextual information in the form of *complex annotations*, adding an additional dimension to our observations which captures some useful information. Here we emphasise that *complex annotation* can refer to any enriched embedding. For the example presented in Figure 1, assuming that we only have access to the minimal data in Figure 1 (b), a human annotator could assign a weighting to each observation, amplifying the significance of specific regions of the model in the loss function and reinforcing our desired model boundaries. This is not a novel proposal and can be likened to over/under-sampling strategies (Mohammed et al., 2020). For illustrative purposes, in this work we propose a complex annotation in the form of direction vector which points to a counterfactual observation for a given point. Section 3 describes how this annotation, in combination with a novel loss function, can enable model improvements.

2. Related Work

Recent approaches in *human-in-the-loop* (HITL) machine learning have focused on empowering humans to effectively interact with and understand complex machine learning models (Wang et al., 2021; Kurzendorfer et al., 2017). These interactions are typically visual, such as attention maps in image captioning tasks (Xu et al., 2015) or activation patterns in deep neural networks (Singh et al., 2019), and are typically utilised to generate post-hoc explanations for human observers. Explanations can be used to validate models by enabling humans to judge whether spurious correlations are being learnt. Depending on the assessment, the model can then be deployed, or re-trained using better data or with a different learning architecture. By integrating these post-hoc explanations into the pipeline, researchers aim to empower humans to actively participate in the training process, further improving our understanding of black-box models.

Active learning mitigates the need for large datasets when training deep learning models. Unlike traditional machine learning methods that rely on randomly or statically labeled data, active learning typically engages humans in the annotation process to strategically choose samples that will most effectively improve the model's performance (Ren et al., 2022). This iterative process allows the model to focus on areas of the data distribution that are initially difficult to learn, leading to better generalization and more efficient training. In the context of deep learning, which requires large labelled datasets to effectively model complex problems, active learning presents an opportunity to optimise for the smallest amount of data required to train these complex networks.

Machine teaching is a field which aims to tackle the problem of finding an optimal training set given a machine learning algorithm and a target model (Zhu, 2015). Recent work com-

bines active learning and machine teaching to automatically generate the minimum viable dataset to learn a concept, and then apply an active learning approach to refine this initial prototype (Mosqueira-Rey et al., 2021).

Human involvement in the learning pipeline is not restricted solely to the examples above. Kurzendorfer et al. (2017) utilise a human expert to provide adjusted boundaries for a 3D segmentation task. Human-interactive segmentation shares a key characteristic with active learning; involvement of human expertise in the annotation process, where the human selects and corrects the most informative regions or boundaries for the model to learn from. The human plays a crucial role in guiding the model's learning process and aligning model behaviour to human preferences. This demonstrates that integrating alternative forms of human feedback and interaction can enhance the accuracy and efficiency of machine learning models, and serves as motivation for further investigation of these techniques.

Our complex annotations rely on the provision of counterfactual observations, which have been proven as a valuable learning signal (Kaushik et al., 2020). In their work, Kaushik et al. aim to combat reliance on spurious correlations for a sentiment classification task by creating a new dataset of counterfactual pairs by employing human annotators to generate minimally different counterfactuals; changing film reviews from positive to negative and vice versa while avoiding any unnecessary changes. They demonstrate that models trained with a mixture of original data and human-generated counterfactuals produce robust models that improve generalisation for a sentiment classification task on the IMDB dataset. Instead of using counterfactuals as additional data points, we build on this work by having human annotators suggest the direction to the nearest counterfactual during model training and establishing a separate learning component which is focused on aligning model gradients along these direction vectors.

3. Methodology

In this work, we want to explore (and produce tools to show) how humans can provide complex forms of supervision during learning to improve model performance, accelerate learning, or both. To this end we (1) propose a simple synthetic learning task for which a human expert can provide valuable insights, (2) build an interface between human and machine learning system to enable meta-evaluations and complex annotations that are machine-intelligible and (3) adapt the machine learning training process to learn from these annotations. We demonstrate that, by combining our meta-evaluation with the standard evaluation of model accuracy, it is possible to enable a human to assess model behaviour and provide insights beyond traditional supervised learning approaches.

3.1. Learning Task

Our problem is 2-dimensional representation of the XOR logical condition with an arbitrary level of noise applied, illustrated in Figure 3. The XOR problem is not inherently challenging to neural networks with more than one layer. However, to a human familiar with the concept of a binary classification, the decision boundary should be immediately obvious. We therefore make the assumption that a human oracle should be able to provide insights to a machine learning algorithm regarding the solution to this classification task.

3.2. Human-Machine Learning Interface

Our hypothesis for this problem is that a human annotator can guide the early learning phase of model training to reduce the number of epochs required to achieve perfect performance. This requires a graphical user interface which enables our proposed meta-evaluation and also the translation of human knowledge into learning signals.

Here, we propose a tool for meta-evaluation which visualises our dataset and the predicted probabilities of the model across the region of interest of the feature space. This presents a live view of the model decision boundaries as they shift during training relative to the labelled examples. By observing how the parameter adaptations are affecting the predictive capabilities of the model across multiple iterations, it is possible to draw conclusions about whether the model is effectively and efficiently learning the correct solution. This evaluation, combined with the model accuracy, provides a deeper insight into the learning process. An example is demonstrated in Figure 1.

As we have discussed, the decision boundary for the XOR problem is readily solvable for any human observer familiar with the problem. Given any observation, a change in classification can be achieved through either a translation in the x- or y-axes exclusively. Formally, these axis refer to Feature 0, F_0, and Feature 1, F_1, in our dataset.

As the classes are clearly separable, it can be assumed that a human expert could take any observation and indicate whether a positive or negative shift in F_0 or F_1 would lead to a change in class. Along this direction, we expect to encounter an observation within the opposing class distribution. We refer to this new observation as a *counterfactual*. We propose a knowledge workbench which enables a human to define the direction of one or more counterfactuals. In traditional supervised learning for classification, we assume a dataset of pairs $\{\mathbf{x_i}, y_i\}$, where $\mathbf{x_i} \in \mathcal{R}^n$ and y_i is the target class label (Hastie et al., 2009). Here, let us consider instead that we have access to doublets of the form of an example $\mathbf{x_i}$ and any number of user-defined direction vectors $\mathbf{d_i} \in \mathcal{R}^n$, along each of which we expect to

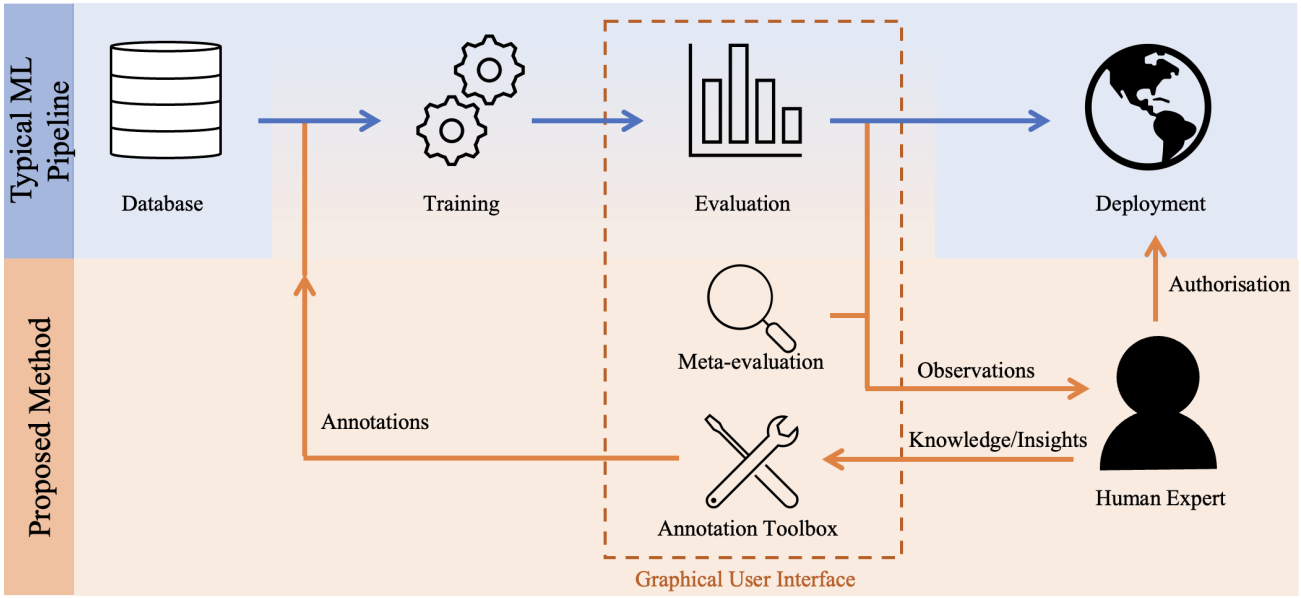

Figure 2. A typical machine learning pipeline contrasted with our proposed method.

intersect the decision boundary. This can be summarised as $\{\mathbf{x_i}, \mathbf{K_i} = \{\mathbf{d_j}, \mathbf{d_{j+1}}, ..., \mathbf{d_k}\}\}$ where k is the cardinality of the set of counterfactuals defined for any given example in the dataset, $0 \leq k < \infty$.

Put differently, we can define multiple counterfactual directions from one observation, and can repeat this process for any number of observations. Subsequently, model training is resumed, and this process can be repeated indefinitely. The annotator has access to our meta-evaluation and the model accuracy, and can add or remove experiments for comparison. A view of the entire interface is provided in Figure 4.

3.3. Leveraging Annotations in Learning

The knowledge vector, $\mathbf{K}$, must be intelligible to the machine learning pipeline to be used for learning. After defining the counterfactual directions in $\mathbf{K}$, we make the simplifying assumption that the gradient of our predictive model for our given class should be negative along these vectors, to indicate a change in probability. For example, as we move from an observation within the distribution of Class 0 towards the region of Class 1, we expect our probability of being in Class 0 to be decreasing.

3.3.1. Loss Function

We assume an Empirical Risk Minimisation approach with the binary cross entropy loss function as baseline (for which a derivation and additional theory is provided by (Goodfellow et al., 2016)). Eq. 1 describes our proposed gradient-based loss function,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j=1}^{k}|\text{sign}(\nabla_{\mathbf{d}_{i,j}} f(\mathbf{x}_i)) + 1| \qquad (1)$$

where $\nabla_{\mathbf{d}_i}$ is the gradient of the model in the direction of $\mathbf{d}_i$ and $f(\mathbf{x}_i)$ is the prediction for input $\mathbf{x}_i$ and N is the number of examples in the training set. The $sign$ function is approximated with a steep $tanh$.

For any observation for which we have attributed a counterfactual direction vector, we calculate the directional derivative of the model gradient in this direction, and penalise any instance where the gradient is not negative. The behaviour of this loss function is that the loss increases when the gradient of the model does not reflect the counterfactual directions provided by the annotator.

4. Future Work

In this paper we present our interface applied to a toy XOR problem. As next steps, we aim to enable this GUI to function on real-world (typically higher dimension) datasets so that we may test our hypothesis; that human expert annotations will either improve model performance or reduce learning time. This requires us to overcome several challenges.

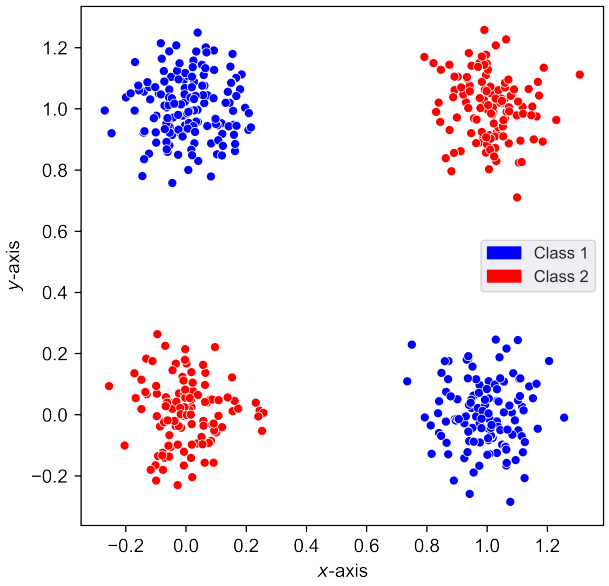

Figure 3. A two-dimensional representation of the logical XOR condition with noisy observations.

4.1. Challenges and Limitations

4.1.1. Moving Beyond Two Dimensions

An obvious shortcoming of the current proposed solution is the reliance on two-dimensional datasets. It may be possible to adapt the current meta-evaluation to three dimensions but for the majority of real world applications with three or more dimensions, a new meta-evaluation must be defined. Looking to active learning, we could provide the annotator with observations which have either been incorrectly classified or query areas of high uncertainty, instead of attempting to illustrate the entire feature space. By engaging an annotator who has expertise in the problem, the same annotation scheme can still be used.

We hope to explore dimensionality reduction techniques to determine whether this meta-evaluation could be effective on higher dimensional problems where the feature space can be realistically compressed into two dimensions. Alternatively, a subset of dimensions of the feature space could be selected by the annotator for evaluation, at the risk of biasing the process. For example, in a high-dimensional space where a relationship is known to exist between only two of those dimensions, the annotator could reinforce the decision boundary along the two dimensions using our meta-analysis.

Under the current annotation scheme, the annotator is required to assess an observation and determine a viable path to a counterfactual by assigning a positive, zero or negative weighting to the annotation vector. There is, however, a limit to the number of features which a human can feasibly interpret. It therefore stands that, within the confines of the meta-evaluation and annotation scheme defined in this paper, we are restricted to n-dimensional datasets where n is the intrinsic dimensionality implied by the skill of the human expert and the task.

4.1.2. Solution Ambiguity

The XOR classification task is easily solvable, but real-world data can be harder to handle, and can be interpreted subjectively by different audiences. Our approach requires an oracle, which begs the question of "Why do we need this approach if we already have a solution". However, the motivation behind this approach is to use humans to reduce the reliance on large data and to bring a human expert into the learning process so they can evaluate and validate models before deployment. A human with no experience on the task can still develop an understanding of both the problem and the model by providing unskilled annotations and observing how model evaluation is affected.

Choosing the right human expert for a task may prove challenging, and to some degree it must be accepted that this form of learning will inherently inject annotator bias into a model. There are, however, two sides to this coin; a trusted human expert could evaluate model behaviour and "train bias out" of models using our approach. Additionally, we could go down the path of estimating the level of skill of the annotators as in the learning from crowds literature (Raykar et al., 2010).

4.2. Potential Applications

As discussed in Section 4.1.1, our meta-evaluation and annotation scheme are highly domain specific, with the toy XOR problem enabling us to walk before we run into harder problems. our intention is that additional meta-evaluations and annotation schemes can be defined and added to a collective toolbox. An immediately obvious annotation scheme for our given meta-evaluation is to give the annotator the ability to add or remove (synthetic) observations. In the case of missing data or underrepresented classes, this could be used to reinforce decision boundaries in regions of interest.

4.2.1. Healthcare

Considering the meta-evaluation and compatible annotation scheme, defined in this paper, health data is of particular interest as the ability to process the numerous features used in predicting patient health is expected to be within the realm of human expertise. For example, the UCI heart disease dataset (Janosi et al., 1988) contains 76 attributes, but all published experiments rely on just 14 attributes. It is plausible that a human expert could examine the attributes of a patient who does not have heart disease, and identify

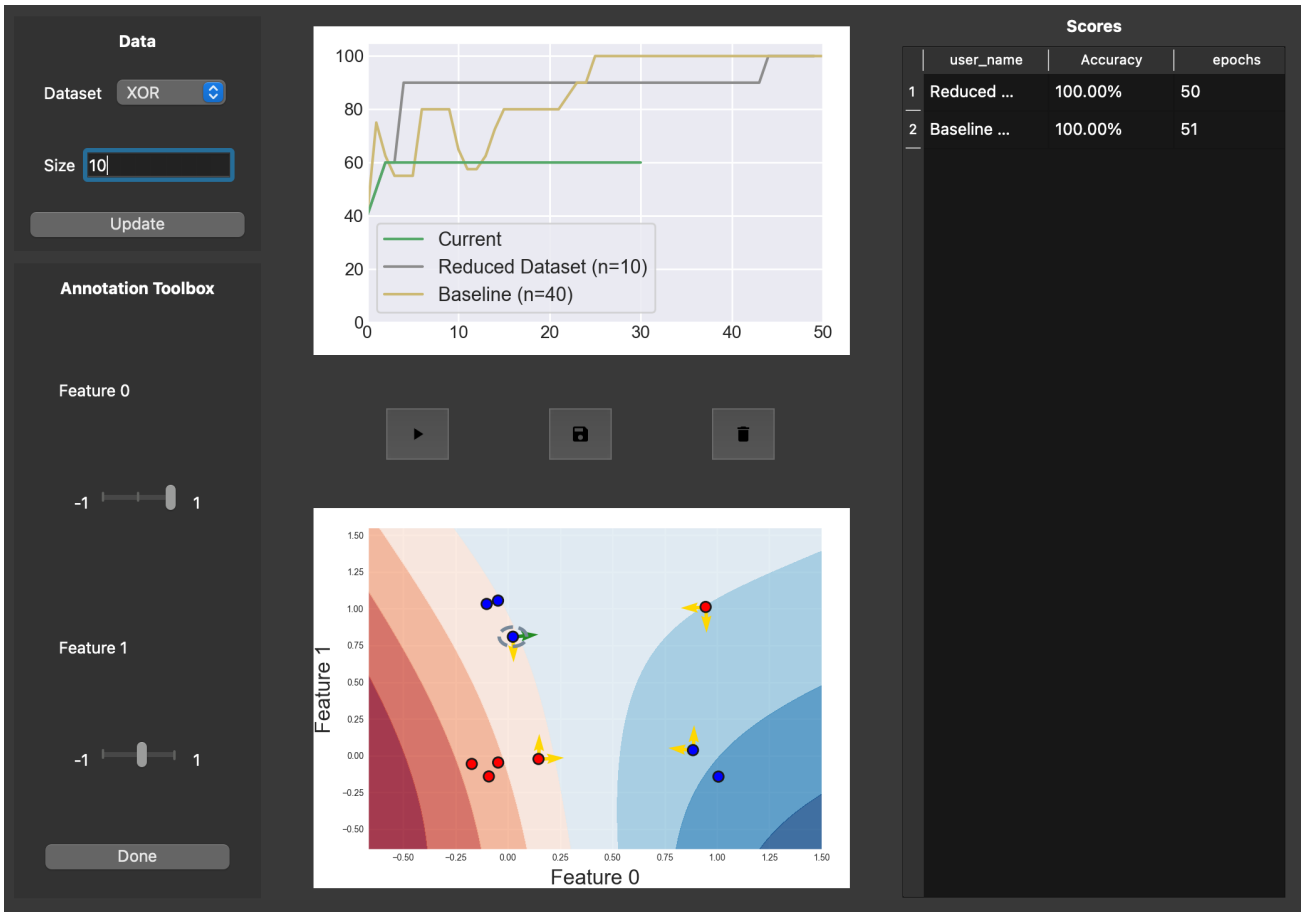

Figure 4. Our Graphical User Interface (GUI) which enables learning from human expert annotations. The current annotation is represented by a green arrow, with prior annotations represented in yellow. The annotation toolbox (bottom left) enables the annotator to define the vector of F_0 and F_1 which will lead to a counterfactual observation. The meta-evaluation (bottom-centre) consists of the dataset and the model predicted probabilities across the feature space. The evaluation console (top-centre) enables comparisons across different experiments. This example shows both the effect of reducing the dataset from $n = 40$ to $n = 10$ and of providing expert human annotations.

some possible directions that would lead to a heart disease diagnoses. These directions could be used in model training to avoid spurious correlations, improve the ability of the model to generalise to new data, or guide the model to learning the whole dataset more efficiently.

4.2.2. Image Processing

Teney et al. (2020) present a similar loss function to that defined in Equation 1 for use in image captioning and visual question answering. They provide pairs of similar images with critical regions masked and use these counterfactual pairs to reinforce model gradients and avoid spurious correlations. We could provide the annotator with counterfactual images which are perceptually close and ask them to choose the closest counterfactual, or ask them to generate a counterfactual image by masking a relevant object and removing this object from the caption.

4.2.3. Natural Language Processing

The research by Kaushik et al. described in Section 2 demonstrates the value of a counterfactually-augmented dataset derived from human annotators. In the language domain, our annotation tool could be similar or identical to theirs, with the exception of the sampling strategy; where they select a random subset and request a dataset containing modified examples prior to learning, we could actively adapt the subset during learning to target regions of high uncertainty during learning. In this context, our method could be thought of as an extension of their (Kaushik et al., 2020) pipeline with an active learning component and an adapted loss function to amplify the signal from these counterfactuals. Our GUI would enable future work to optimise the number of annota-

tions required by both methods to improve both the speed of learning and the effect on model performance on out-of-distribution data. This would help us understand whether our method reduces annotator burden.

5. Conclusions

In this work, we introduce a novel interface that allows users to provide complex forms of supervision to shape the way learning occurs in machine learning systems. As an example, we show how our proposed graphical user interface promotes human-machine learning interaction by enabling the meta-evaluation of model performance and subsequent annotation by a human oracle, reinforcing model gradients in critical regions. We demonstrate the learning pipeline on a toy XOR problem, discuss the potential benefits and existing limitations of this work, and suggest the exploration of other meta-evaluations and annotation schemes moving forward. We describe how this interface might be adapted to several domains, but further work is required within each domain to define the appropriate meta-evaluations and annotation tools, test viability, and determine the equilibrium between annotator burden and the utility of counterfactually-augmented data.

Acknowledgements

Thanks to UK Research and Innovation (UKRI) and Thales Training & Simulation Ltd. who are jointly funding Jonathan Erskine's PhD research through the UKRI Centre for Doctoral Training in Interactive Artificial Intelligence under grant EP/S022937/1.

This work was partially funded by the UKRI Turing AI Fellowship EP/V024817/1.

References

Collins, K. M., Wong, C., Feng, J., Wei, M., and Tenenbaum, J. B. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. Heart Disease. UCI Machine Learning Repository, 1988. DOI: 10.24432/C52P4X.

Kaushik, D., Hovy, E., and Lipton, Z. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Kurzendorfer, T., Fischer, P., Mirshahzadeh, N., Pohl, T., Brost, A., Steidl, S., and Maier, A. Rapid interactive and intuitive segmentation of 3d medical images using radial basis function interpolation. *Journal of Imaging*, 3(4), 2017. ISSN 2313-433X. doi: 10.3390/jimaging3040056. URL https://www.mdpi.com/2313-433X/3/4/56.

Mohammed, R., Rawashdeh, J., and Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–248, 2020. doi: 10.1109/ICICS49469.2020.239556.

Mosqueira-Rey, E., Alonso-Ríos, D., and Baamonde-Lozano, A. Integrating iterative machine teaching and active learning into the machine learning loop. *Procedia Computer Science*, 192:553–562, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.08.057. URL https://www.sciencedirect.com/science/article/pii/S1877050921015441. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021.

Prince, M. Does active learning work? a review of the research. *Journal of engineering education*, 93(3):223–231, 2004.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. URL http://jmlr.org/papers/v11/raykar10a.html.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):1–40, December 2022.

Singh, C., Murdoch, W. J., and Yu, B. Hierarchical interpretations for neural network predictions, 2019.

Teney, D., Abbasnedjad, E., and Hengel, A. v. d. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, pp. 580–599. Springer, 2020.

Wang, Z. J., Choi, D., Xu, S., and Yang, D. Putting humans in the natural language processing loop: A survey, 2021.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models, 2022.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention, 2015. URL https://arxiv.org/abs/1502.03044.

Zhu, X. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), March 2015. doi: 10.1609/aaai.v29i1.9761. URL https://doi.org/10.1609/aaai.v29i1.9761.