Counterfactual Gradient Alignment: Optimizing Directional Expert Supervision for Data-Efficient Learning

Jonathan Erskine
University of Bristol
Bristol, United Kingdom

Matt Clifford
University of Bristol
Bristol, United Kingdom

Alexander Hepburn
University of Bristol
Bristol, United Kingdom

Raúl Santos-Rodríguez
University of Bristol
Bristol, United Kingdom

ABSTRACT

Typically, machine learning algorithms must find patterns in samples drawn from an unknown data distribution, with no prior contextual information about the input space or intuition for the problem they are solving. Humans, by contrast, can apply both intuition and prior knowledge to quickly solve new problems, requiring less training data. In this paper we investigate how to address this disparity by enriching traditional labels for supervised classification tasks; embedding additional information through gradient-based *expert annotations*. We present a framework for learning from an oracle by (1) annotating existing observations with the direction in the input space which indicates a change of class and (2) utilising a loss function which rewards negative model gradients along these directions. We demonstrate this approach for a synthetic 2D problem and two natural language sentiment classification tasks, learning the task in fewer epochs than the standard method.

KEYWORDS

Counterfactuals, Supervised Learning

1 INTRODUCTION

Highly tailored machine learning models typically require large amounts of data and may fail to generalise well to out-of-distribution examples, leading to unexpected or inaccurate predictions [5, 9, 26]. Humans are, by contrast, relatively fast learners. We are able to utilise information outside of the usual data-label pair that is provided to machine learning systems, requiring fewer data points in order to draw conclusions and generalise to unseen data. It would be trivial, for example, for a human to adapt from film reviews to political tweets when performing sentiment classification, where a machine learning (ML) model may require refinement [4]. When operating in real environments the underlying distribution of observable data is often dynamic, subject to change and has limited availability [19]. Particularly when errors carry a high cost, be it monetary or a risk to human safety, strong generalisation is a necessary component of any model which aims to assist or replace a highly adaptable human supervisor.

Whether we must maximise the utility of a few available samples or adapt to new observations quickly, the objective is often to learn the underlying distribution from a minimal number of examples while avoiding over-fitting. In online settings such as Active Learning, lack of data is addressed by the machine learning algorithm communicating with a (typically human) oracle who returns a new label for each query example, usually selected to reduce a form of uncertainty [16]. Formulations like Machine Teaching approach the problem by allowing humans to select minimal groups of samples that effectively describe concepts and patterns [30]. Both approaches can be said to improve data efficiency and the ability of models to generalise to new data through an informed selection of training data.

In this work we describe an alternative approach; rather than querying an oracle, we enrich the data in the form of more *complex annotations*. We propose that embedding more information in the annotation and learning process can increase data efficiency, at the expense of annotation complexity. We test our hypothesis on two binary classification tasks; (1) a synthetic 2D dataset, and (2) a sentiment classification task for the IMDB Large Movie Review Dataset [14]. We present an example of one possible complex annotation by enriching our annotations with a direction vector which indicates the direction a sample must move in feature space in order to change the samples' classification; a *counterfactual direction*. We also propose a novel loss function which enforces negative model gradients in the direction of the counterfactual, with the intuition being that a change in class should happen along this direction. Where gradients have been used previously as a natural method for providing human-interpretable explanations [24], this paper aims to reverse this pipeline and provide machine-interpretable supervision.

Finally, we investigate entropy- and diversity-based importance sampling measures to maximise data utilization in scenarios of highly-limited annotation budgets, strategically selecting the most informative samples for counterfactual annotation. This approach combines uncertainty-based selection (targeting samples where the model is most confused) with diversity constraints (ensuring broad coverage across different regions of the feature space), maximizing the value of each human-provided counterfactual direction while avoiding redundant annotations in similar data regions.

2 RELATED WORK

In this section we review state-of-the-art methods for including additional information in machine learning systems that go beyond data-label pairs, namely; human priors, counterfactual observations and gradient supervision.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

2.1 Embedding human priors in model training

One significant motivation of this work is to enable human annotators to provide domain knowledge to models to improve data efficiency. Rieger et al. [17] demonstrated the ability to embed human priors in their work developing contextual decomposition explanation penalisation (CDEP). CDEP enables embedding domain knowledge during model training to ignore spurious correlations, correct errors, and generalise to different types of dataset shifts. This method assumes that the explanation method is truthful and penalises model explanations where these do not align with explanations provided by an oracle. Domain knowledge can be encoded by a human or in an automated fashion. In a similar fashion, Ross et al. embed human priors as explanations and develop a loss function which balances cross entropy loss with an additional function to penalise misaligned model explanations [18]. In their applications, they reinforce the magnitude of gradients to align to pre-supposed explanation; penalising the model where features are inappropriately neglected or considered to be too important.

2.2 Training with counterfactuals

Kaushik et al. [11] present a system for collecting counterfactually augmented data by tasking human annotators to minimally edit individual observations to produce a counterfactual on a subset of the IMDB dataset [14]. They demonstrate that incorporating counterfactuals within the training data leads to improved generalisation and show that training purely with unaltered data or counterfactual pairs leads to over-fitting on spurious signals, while a combined dataset produces a model which is less sensitive to these signals.

In a follow up paper, Kaushik et al. investigate why counterfactually augmented data reduces spurious correlations [12]. They propose that adding noise over causal features should worsen in-domain and out-of-domain performance, but that adding noise to non-causal features should improve relative performance on out-of-domain tasks. To test this hypothesis, they apply noise to the complement of the edited sections of text within their counterfactually augmented dataset. They repeat this process for sections of text identified by an attention-based classifier (BiLSTM with Self-Attention [8]) and by gradient-based feature attribution [13], and find that human annotations served as the best augmentation for performance on out-of-distribution data.

In our work we evaluate models trained on the counterfactually augmented data produced both by humans and algorithmically. However, we take a different approach to learning from these counterfactual augmentations through gradient supervision.

2.3 Learning from counterfactuals through gradient supervision

Teney et al. [27] introduce a novel training objective called "gradient supervision", which involves penalising the angle between the gradients of a neural network classifier and the vector difference between counterfactual examples in the input space. Using a counterfactual, Teney et al. define a loss function which aims to minimise the angle between the model gradient and vector pointing to the counterfactual as

$$\mathcal{L}_{GS}\left(\mathbf{g}_i, \hat{\mathbf{g}}_i\right) = 1 - \langle \mathbf{g}_i, \hat{\mathbf{g}}_i \rangle / (\|\mathbf{g}_i\| \|\hat{\mathbf{g}}_i\|) \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between two vectors, $\mathbf{g}_i$ denotes the gradient of the model with respect to an observation $\mathbf{x}_i$ and $\hat{\mathbf{g}}_i$ is the vector between counterfactual examples.

Eq. 1 is used as a regulariser alongside the conventional cross-entropy loss in order to align the model's decision surface between pairs of counterfactual examples. The authors demonstrate the effectiveness of their approach on tasks known for their susceptibility to poor generalisation due to biases in the training data; visual question answering, multi-label image classification, sentiment analysis, and natural language inference. Where Teney et al. assume alignment between the feature space and the output space and attempt to minimise the angle between the gradient of the model and the counterfactual direction, our proposed loss function is less restrictive, enforcing only the sign of the gradients of the model.

2.4 Active Learning and Importance Sampling

The selection of highly informative samples is a cornerstone of efficient model training under data constraints. Traditionally, active learning (AL) research has been dominated by *uncertainty sampling* strategies, such as least confidence, margin sampling, and entropy-based selection [21]. While effective at identifying samples near the decision boundary, these methods are prone to the "batch redundancy" problem—selecting multiple samples that are informative but highly similar to one another—which leads to suboptimal gain per annotation dollar [23].

To mitigate this, recent literature has shifted toward *diversity-aware* or *hybrid* sampling. Sener and Savarese [20] framed AL as a core-set problem, using geometric distance in the embedding space to ensure the selected subset covers the entire data distribution. This representativeness ensures that the model learns the global structure of the manifold rather than over-focusing on specific noisy boundaries.

Contemporary approaches often integrate these two paradigms. Hybrid strategies, as explored by Citovsky [6] and Ash et al. [1], utilize a weighted combination of predictive uncertainty and geometric novelty. By maximizing the distance between a candidate and the currently selected set, these methods enforce a diverse batch composition. In our work, we adapt this hybrid approach to the specific task of counterfactual enrichment, ensuring that the limited human supervision we collect is distributed across distinct regions of the feature space. This ensures high utility for gradient-based alignment without the redundant overhead typical of pure uncertainty-driven selection.

3 GRADIENT-BASED LEARNING

3.1 Directional Expert Annotation

In traditional supervised machine learning, we assume a dataset of N random independent identically distributed (i.i.d.) observations $\mathbf{x_i}, y_i \sim P(\mathbf{x}, y)$, where $\mathbf{x_i} \in \mathcal{R}^d$ and y_i is the target class label [10]. In the CGA framework, we consider triplets $\mathbf{x_i}, y_i, \mathbf{d_i}$, where $\mathbf{d_i} \in \mathcal{R}^d$ defines a unit vector from $\mathbf{x_i}$ pointing towards a proximal

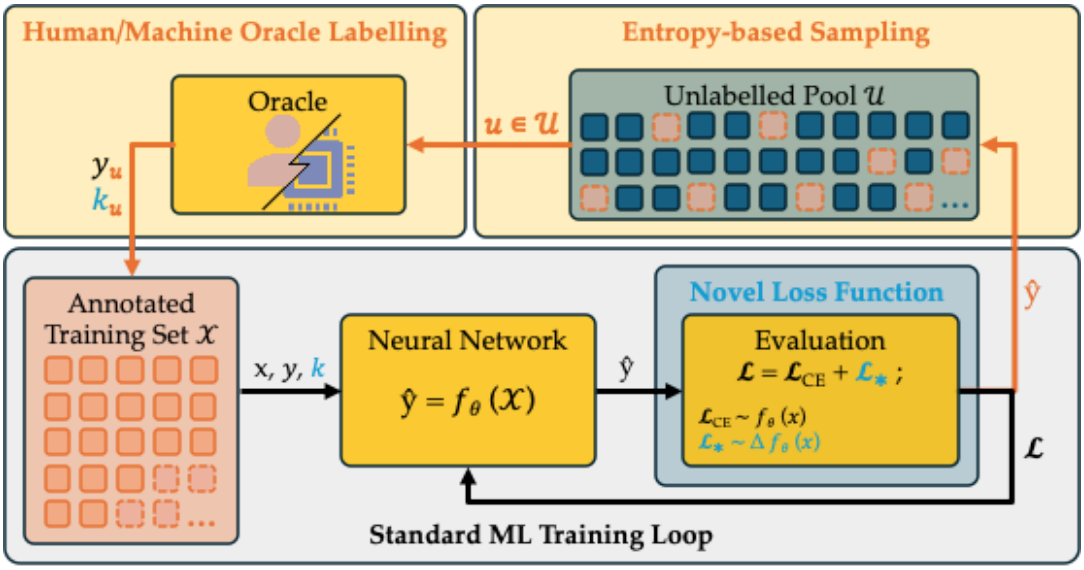

Figure 1: Overview of the proposed counterfactually-augmented active learning framework. The model leverages both factual and counterfactual samples, selected via importance sampling, and trained using a hybrid loss function.

counterfactual $\hat{\mathbf{x}}_i$:

$$\mathbf{d_i} = \frac{\hat{\mathbf{x}}_i - \mathbf{x_i}}{||\hat{\mathbf{x}}_i - \mathbf{x_i}||}. \tag{2}$$

We propose that moving from $\mathbf{x_i}$ towards $\hat{\mathbf{x}}_i$ should result in a decrease in the model's prediction probability for the true class y. Let $f_y(\mathbf{x})$ be the model's logit score for class y. We define the **Directional Derivative** d as the projection of the gradient onto the counterfactual direction:

$$d = \nabla_x f_y(x) \cdot \mathbf{d}_i \tag{3}$$

This derivative quantifies how much the class confidence changes as we perturb the input along the expert-provided direction.

3.2 Optimized Loss Function Variants

We evaluate three loss variants $\mathcal{L}_d$ designed to minimize d, aligning the model's confidence drop with the counterfactual path. The intuition is that d should be negative at the counterfactual direction, with each variant modifying the profile of this negative loss in terms of magnitude and steepness.

Sign (Tanh) Directional Loss. The purest form of our proposed loss function penalises only the sign of the model gradient; disregarding negative model gradients and generating a fixed loss penalty for positive model gradients, regardless of the magnitude of the gradient. The Sign variant is formulated as:

$$\mathcal{L}_{\text{Sign}} = \tanh(200 * d) + 1 \tag{4}$$

where the high multiplier creates a very steep tanh function, squashing the derivative into the range $\mathcal{L}_{\text{Sign}} \in 0, 2$. We hypothesise that focusing purely on the sign of the alignment ensures a bounded, stable regularizing signal across different manifolds.

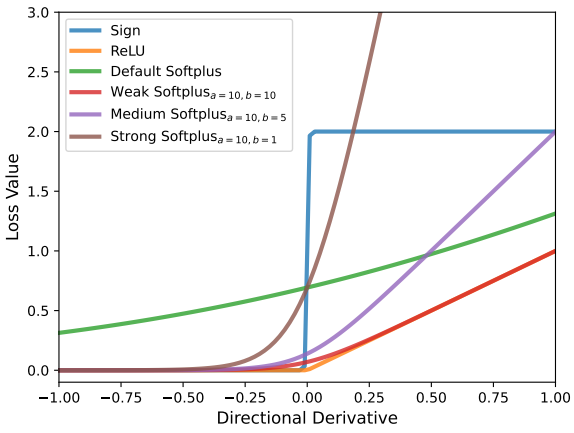

Figure 2: Comparison of the three directional loss variants $\mathcal{L}_{\text{Sign}}$, $\mathcal{L}_{\text{ReLU}}$, and $\mathcal{L}_{\text{Softplus}}$ as functions of the directional derivative d. The Sign loss provides a steep, bounded penalty based only on the gradient sign, the ReLU loss applies a linear penalty for positive (misaligned) gradients, and the Softplus loss offers a smooth, tunable alternative that can still impose a non-zero penalty even when d is negative, depending on its magnitude.

ReLU Directional Loss.

$$\mathcal{L}_{\text{ReLU}} = \max(0, d) \tag{5}$$

This applies a linear penalty only if the model's confidence is incorrectly *increasing* toward the counterfactual boundary. It is a "hinge-like" objective that stops regularizing once the gradient sign

is correct. By attributing higher penalties to steeper positive gradients, this function is more communicative, but may prove to be over-restrictive, particularly for noisy datasets.

Softplus Directional Loss. Softplus is a smooth version of ReLU that provides a continuous alignment signal. Unlike ReLU, it can provide a non-zero penalty even when d is negative dependent on the magnitude of the derivative. While this may encourage the model to maintain a "steep" confidence drop-off even when it is already locally "correct", it is unclear how informative this signal will be, particularly in sparse regimes. We therefore utilise a modified Softplus function defined as:

$$\mathcal{L}_{\text{Softplus}} = \log(1 + e^{ad})/b \tag{6}$$

where we can control the strength of this supervision by modifying a and b. Figure 2 compares the behaviour of each loss function variant.

3.3 Total Training Objective

The total training loss is a weighted mixture of classification and alignment objectives:

$$\mathcal{L}_{\text{Total}} = (1 - \alpha)\mathcal{L}_{\text{CE}} + \alpha \left(\frac{1}{N_d} \sum_{i=1}^{N_d} \mathcal{L}_{d,i} \right) \tag{7}$$

where N_d is the number of examples in the training set for which we have labelled directions, and $\alpha \in [0, 1]$ controls the trade-off between standard classification and geometric supervision. In our JAX-based implementation, we leverage efficient automatic differentiation to compute the directional derivative d within the optimization loop.

3.4 Informative and Diverse Sample Selection

To maximize the utility of human-in-the-loop annotations under data scarcity, we employ a hybrid importance sampling strategy that balances *informativeness* (via entropy) and *diversity* (via geometric distance). This approach addresses a known failure mode of pure uncertainty sampling: the redundancy problem, where a model selects high-entropy samples that are spatially clustered, leading to overlapping and inefficient gradient updates [23].

Our selection mechanism follows a two-stage process. First, we identify a candidate pool of the most informative samples based on Shannon entropy, $H(p) = -\sum p_i \log p_i$. To ensure this subset provides broad coverage of the feature manifold, we apply a greedy selection process inspired by the K-Center-Greedy approach [20]. For each candidate x in the high-entropy pool, we compute a selection score:

$$\text{score}(x) = H(x) + \alpha \cdot \min_{s \in \mathcal{S}} \|\phi(x) - \phi(s)\|_2 \tag{8}$$

where $\mathcal{S}$ is the set of already selected samples, $\phi(\cdot)$ denotes the embedding function, and α is a diversity weight.

This weighted hybrid formulation, as explored in batch active learning [6], ensures that each selected sample is both locally uncertain and globally distant from existing annotations. By employing this strategy, we ensure that the counterfactual direction annotations enrich the model across a diverse range of the data distribution, which is critical for effective supervision when annotation budgets are strictly limited.

4 EXPERIMENTAL SETUP

Our method is applicable to datasets where counterfactuals exist, or can be generated. We first validate our method on several two dimensional classification tasks for which counterfactuals can be easily generated. Subsequently, we evaluate on two language classification tasks; (1) binary sentiment classification on IMDB movie reviews, with *human-generated* counterfactuals collected by Kaushik et al. [11], and (2) a topic classification task on news articles spanning four topics, first introduced by Zhang et al. [29].

4.1 Datasets

The objective of our experiments is to measure how incorporating counterfactuals using Eq. 7 may improve data efficiency. To this end we utilise small subsets of the available training data to simulate data scarcity, a common challenge in real-world applications, and report results for training sets of various sizes. In all experiments we compare our method against the gradient supervision model proposed by Teney et al. [27] (replacing $\mathcal{L}_d$ in Eq 7) and a baseline model trained on cross-entropy (Eq. 7, $\alpha = 0$).

Table 1: Dataset Split Sizes.

Dataset	Train	Validation	Test
Synthetic 2D			
Circles	6	1,000	–
XOR	6	1,000	–
Two Moons	6	1,000	–
9-Blobs	60	1,000	–
NLP Classification			
IMDB	1,707	245	488
AGNEWS	500	3,550	3,550

4.1.1 2D Topologies. We evaluate our method across four synthetic 2-dimensional topologies generated using the `scikit-learn` library [15]. These consist of three binary classification tasks—XOR, Concentric Circles, and Two Moons—and a more complex multi-class objective featuring nine classes of distinct Gaussian clusters. This selection allows us to assess the framework's performance on both interleaved non-linear boundaries and high-entropy multi-class partitions. Counterfactuals are generated using a brute force grid search, implemented with the FAT Forensics toolkit [25]. We first tune the model hyperparameters and dataset size on the cross-entropy baseline, finding the minimal amount of data to fully learn the problem and then marginally reducing the dataset further to introduce epistemic uncertainty. We train with relatively small subsets of data and omit a test set in lieu of a suitably large validation set. Training and validation set sizes are given in Table 1. To compare data efficiency between counterfactuals and simply learning from additional labels, we also introduce an **upper bound baseline which is trained on twice as many examples**.

Model Architecture. For these 2D datasets, we utilize an over-parameterized Multi-Layer Perceptron (MLP) consisting of a single hidden layer with 512 units and `ReLU` activation. While this width is high relative to the input dimensionality, we observe that overparameterization consistently boosts validation accuracy on the control model trained with cross-entropy and provides a high-dimensional feature space conducive to geometric alignment. The model is optimized using the **AdaBelief** optimizer [31] with β_1=0.9, β_2=0.999, and ϵ=1^{-16}, with the learning rate l = 0.001.

4.1.2 Natural Language. We utilise two openly available text datasets which provide counterfactual examples alongside training data for binary sentiment classification on movie reviews (IMDB) and topic classification for news articles (AGNEWS). Table 1 gives the sizes of available data for training, validation and testing for each dataset. For IMDB, we utilised the human-generated counterfactually-augmented dataset curated by Kaushik et al. [11], maintaining separation of the test set into the original and augmented reviews as in the original work[11]. For AGNEWS, we utilised machine-generated counterfactuals generated by Wang et al. [28]. In their study, Wang et al. [28] evaluate several generative models with both machine and human evaluators. To investigate potential variance from counterfactual quality, we evaluate two sets of counterfactually-augmented data:

(1) Counterfactuals generated by (1) Llama3-70B using the FIZLE generation methodology[3] for which label flip score (LFS) is well aligned between human and machine evaluators; 63.6% and 64.4% respectively.
(2) Counterfactuals generated by Llama3-8B using the more polarising FLARE methodology [2] which achieves only 42.2% LFS against a machine judge and a much higher 86.7% for humans.

Data Preprocessing. All text datasets were stripped of HTML, bracketed content, URLs, email addresses, and numbers, followed by tokenization into lowercase words, stopword removal, and rule-based stemming to normalize lexical forms.

Feature Representation. A `CountVectorizer` was fitted to the training corpus, limited to the 20,000 most frequent tokens. This vocabulary was then reused for both development and test sets. Each document was transformed into a fixed-length sequence of token indices mapped to the vocabulary, truncated or padded to a length of 32.

Model Architecture. For text classification, we employed a **Bag-of-Words (BoW)** neural network implemented in `JAX/Flax`. Two model variants were used depending on the task: a binary classifier (`BagOfWordsClassifierSimple`) and a multi-class classifier (`BagOfWordsClassifierMultiClass`). Both models share the same general structure:

(1) **Embedding layer:** A trainable embedding matrix maps token indices to dense vectors of dimension 50.
(2) **Average pooling:** Embeddings for all tokens in a document are averaged, ignoring padding tokens.
(3) **Dropout:** A dropout layer (rate = 0.5) is applied to mitigate overfitting.
(4) **Linear projection:** A single dense layer maps the pooled representation to either one logit (binary classification) or N logits (multi-class classification).

The model outputs both the predicted logits and the intermediate text representation, which are used for downstream tasks such as similarity computation and importance sampling.

4.2 Ablation Study and Importance Sampling

For our 2D experiments, we provide a counterfactual on a per-example basis. We then test all loss function variants across all problems, varying α to modify the influence of our loss function. After optimising on the 2D datasets, we fix α and select the most effective variant of our proposed loss function to perform experiments within the NLP setting. To understand the relationship between counterfactual annotations and information gain we randomly ablate generated counterfactuals to some ratio $r \in [0, 1]$ of total training examples **in the NLP setting only**. Our interest in low-data regimes motivates further investigation whereby we perform importance sampling; we repeat each NLP experiment starting from 10% of available data, iteratively selecting 5% unlabelled observations and adding these to the training data every 2 training epochs before re-initialising the model and re-training. We run for the same amount of epochs as the 'AL Off' scenario as we are mainly interested in data efficiency in earlier epochs. Data selection is performed by balancing label uncertainty and sample diversity. A full description of our sampling strategy is provided in Appendix A.

5 RESULTS

5.1 2D Experiments

Table 2 shows results for these experiments. We prove that enabling learning through counterfactuals (either through gradient supervision or through some variant of our proposed CGA method) leads to faster learning or better accuracy, if not both. This is unsurprising given the amount of additional information which the model has access to. What is noteworthy is that, for **Two Moons** and **9-Blobs**, our CGA method seems to be able to improve on data efficiency compared to the upper bound. In the case of **Two Moons**, *Softplus-Strong* is able to approximate the upper bound while training for less than 10% of the number of epochs. For **9-Blobs**, all of our proposed methods surpass the upper bound. The best performing variant is inconsistent across experiments.

5.2 NLP Classification Tasks

5.2.1 IMDB. Table 3 and Table 4 show results for sentiment classification on the IMDB dataset for the case of full data availability and the active learning setting respectively. At N = 20 (and extending to N = 50 under the active learning paradigm), all variant collapse to a baseline accuracy of 52.9% at Epoch 0. This indicates that for this specific architecture and task, 20 to 50 samples are simply insufficient to overcome initial random initializations, regardless of the loss function variant applied.

As the dataset scales to N = 100 and N = 500, the Gradient Supervision (GS) variant with r = 1.0 and the standard Cross-Entropy (C.E.) baseline consistently outpace the Softplus-Strong

Table 2: Validation Accuracy (%) and Epochs for 2D Classification Tasks trained with and without counterfactuals, using different loss function variants while varying the relative strength of the supervisory signal ($\alpha \in [0, 1]$). Bold values show the best accuracy for that specific loss function variant, while Underlined Bold shows the best accuracy across the entire experiment. Epochs in bold indicate that the run is the most efficient in its group, but only if it is also the most accurate. An upper bound is trained on twice as much data, to allow comparison of data efficiency between learning from counterfactuals or simply adding more labelled examples.

		Circles		XOR		Two Moons		9-Blobs	
Variant	α	% Acc	Ep	% Acc	Ep	% Acc	Ep	% Acc	Ep
Baseline ($\mathcal{L}_{\text{CE}}$)	0.0	87.8 (±0.054)	35	99.1 (±0.001)	86	77.3 (±0.003)	18	84.8 (±0.002)	194
Upper Bound ($\mathcal{L}_{\text{CE}}$)	0.0	100.0% (±0.029)	37	100.0% (±0.002)	35	86.6% (±0.001)	58	85.8% (±0.004)	162
Gradient Supervision	0.3	**74.7 (±0.027)**	**74**	99.7 (±0.001)	78	**85.7 (±0.031)**	95	**85.8 (±0.004)**	162
	0.5	74.2 (±0.020)	76	99.7 (±0.001)	65	81.3 (±0.008)	46	85.4 (±0.003)	123
	0.7	74.4 (±0.019)	85	**99.8 (±0.001)**	88	80.4 (±0.035)	37	85.1 (±0.001)	187
Sign	0.3	**80.9 (±0.078)**	**36**	95.9 (±0.028)	99	77.0 (±0.014)	17	86.4 (±0.016)	197
	0.5	80.4 (±0.068)	47	**97.2 (±0.036)**	**99**	77.0 (±0.016)	8	**87.1 (±0.016)**	196
	0.7	80.6 (±0.065)	47	95.7 (±0.054)	99	**77.5 (±0.011)**	13	86.3 (±0.024)	191
ReLU	0.3	**90.2 (±0.062)**	44	**99.5 (±0.002)**	**72**	78.8 (±0.006)	13	87.0 (±0.006)	196
	0.5	88.5 (±0.046)	40	**99.5 (±0.002)**	78	79.0 (±0.010)	8	87.6 (±0.002)	195
	0.7	85.4 (±0.062)	39	99.5 (±0.004)	95	**79.8 (±0.013)**	**5**	**88.6 (±0.007)**	**190**
Softplus-Weak	0.3	87.2 (±0.041)	43	99.5 (±0.001)	66	80.2 (±0.005)	21	87.5 (±0.005)	199
	0.5	86.7 (±0.035)	50	99.5 (±0.001)	70	81.2 (±0.005)	12	88.7 (±0.006)	190
	0.7	**88.2 (±0.075)**	99	**99.6 (±0.001)**	88	**82.8 (±0.018)**	**4**	**89.2 (±0.004)**	**159**
Softplus-Medium	0.3	87.0 (±0.035)	49	99.5 (±0.001)	69	81.1 (±0.004)	13	88.4 (±0.006)	194
	0.5	**88.6 (±0.081)**	99	**99.6 (±0.001)**	92	82.3 (±0.018)	4	89.0 (±0.006)	173
	0.7	88.4 (±0.028)	99	**99.6 (±0.001)**	91	**84.6 (±0.030)**	**2**	**89.1 (±0.008)**	**130**
Softplus-Strong	0.3	**90.2 (±0.044)**	168	99.6 (±0.002)	89	84.6 (±0.024)	3	88.9 (±0.005)	139
	0.5	82.5 (±0.041)	27	**99.7 (±0.003)**	96	**86.3 (±0.026)**	**3**	**89.3 (±0.004)**	192
	0.7	82.1 (±0.080)	26	99.5 (±0.004)	99	85.5 (±0.011)	5	**89.3 (±0.004)**	183

alternative. In the standard AL Off setting, GS ($r = 1.0$) edges out C.E. to achieve the global best accuracy at both $N = 100$ (63.1%) and $N = 500$ (75.8%). Softplus-Strong, conversely, lags behind the baseline by nearly 8% at $N = 500$.

We observe a strict sensitivity to the mixing parameter r within the GS framework. While $r = 1.0$ yields peak performance, reducing the ratio to $r = 0.5$ triggers catastrophic forgetting or learning failure; at $N = 500$ (AL Off), accuracy drops from 75.8% down to a near-baseline 53.9%.

In the active learning setting, neither method is able to surpass the baseline model trained with cross-entropy. However, for the **N = 100** dataset, we are able to approximate the baseline using 50% less training epochs, with our loss function learning slightly faster than the gradient supervision method.

5.2.2 AGNEWS. Table 5 and Table 6 present the validation accuracies for the AGNEWS text classification dataset, evaluated across the FIZLE and FLARE experimental subsets for both standard (AL Off) and active learning (AL On) sampling strategies.

Consistent with our previous NLP observations, an extreme low-data bottleneck exists at $N = 20$, where all models fail to exceed the 24.2% baseline (random chance across the four AGNEWS classes). However, a significant divergence occurs at the $N = 50$ threshold under standard uniform sampling (AL Off). Here, the Softplus-Strong ($a = 0.5$) variant demonstrates exceptional sample efficiency, leaping to a global best of 44.9% (FIZLE) and 48.7% (FLARE) with $r = 0.5$, while the Cross-Entropy (C.E.) baseline stagnates around 30.2%. As the dataset scales to $N = 100$ and $N = 500$, the Softplus-Strong variant with $r = 1.0$ produces the best results, although the discrepancy is smaller. Importantly, we note only a marginal reduction in performance when utilising half as many counterfactuals for our loss variant (at $N = 50$ we actually improve by reducing counterfactual annotations). The gradient supervision method is markedly more sensitive to ablation.

Across almost all viable training scales ($N \geq 50$), the FLARE subset consistently yields higher peak accuracies than FIZLE. At $N = 500$ (AL Off), the best FLARE configuration outperforms its FIZLE counterpart by 2.5% (77.8% vs. 75.3%). Counterfactuals generated by the FIZLE method reported a much higher LFS across all model variants compared to FLARE, but FLARE scored a much higher LFS with humans[28]; this suggests that the features or preprocessing transformations inherent to the FLARE split provide a more separable representation space for the classifier, particularly when leveraged by the Standard loss variant.

Table 3: Test Accuracy (%) and Epochs on IMDB (AL Off). GS = Gradient Supervision, C.E. = Cross-Entropy. For loss functions which utilise counterfactuals we fix the parameter, $\alpha = 0.5$. Bold values show the best accuracy for that specific loss function variant, while Underlined Bold shows the best accuracy across the entire experiment. Epochs in bold indicate that the run is the most efficient in its group, but only if it is also the most accurate.

		Dataset Size							
		20		50		100		500	
Variant	r	% Acc	Ep	% Acc	Ep	% Acc	Ep	% Acc	Ep
C.E.	-	52.9 (±0.033)	0	50.4 (±0.002)	55	62.9 (±0.005)	51	75.4 (±0.003)	14
GS	1.0	52.9 (±0.033)	0	50.4 (±0.002)	56	**63.1 (±0.004)**	45	**75.8 (±0.004)**	14
	0.5	52.9 (±0.033)	0	**54.3 (±0.016)**	**18**	54.3 (±0.024)	5	53.9 (±0.027)	0
Softplus-Strong	1.0	52.9 (±0.033)	0	**49.8 (±0.006)**	**1**	**57.6 (±0.008)**	53	**67.6 (±0.009)**	**22**
	0.5	52.9 (±0.033)	0	**49.8 (±0.005)**	**1**	56.8 (±0.006)	52	66.8 (±0.012)	33

Table 4: Test Accuracy (%) and Epochs on IMDB (AL On). GS = Gradient Supervision, C.E. = Cross-Entropy. For loss functions which utilise counterfactuals we fix the parameter, $\alpha = 0.5$. Bold values show the best accuracy for that specific loss function variant, while Underlined Bold shows the best accuracy across the entire experiment. Epochs in bold indicate that the run is the most efficient in its group, but only if it is also the most accurate.

		Dataset Size							
		20		50		100		500	
Variant	r	% Acc	Ep	% Acc	Ep	% Acc	Ep	% Acc	Ep
C.E.	-	52.9 (±0.033)	0	52.9 (±0.033)	0	**58.8 (±0.004)**	94	**72.3 (±0.002)**	81
GS	1.0	52.9 (±0.033)	0	52.9 (±0.033)	0	**58.6 (±0.006)**	67	**71.5 (±0.003)**	83
	0.5	52.9 (±0.033)	0	52.9 (±0.033)	0	53.7 (±0.023)	63	53.5 (±0.018)	9
Softplus-Strong	1.0	52.9 (±0.033)	0	52.9 (±0.033)	0	**57.8 (±0.007)**	**61**	**67.4 (±0.005)**	94
	0.5	52.9 (±0.033)	0	52.9 (±0.033)	0	57.0 (±0.006)	98	66.6 (±0.003)	73

For AL On, learning with counterfactuals does not significantly improve on the baseline, with the exception of $N = 100$, where results are heavily dependent on the counterfactual augmentation and loss variant applied). For $N = 500$, both methods which utilise counterfactuals learn faster, with our method training in 33% the amount of epochs required for the baseline.

6 DISCUSSION

Our experiments on the IMDB and AGNEWS datasets reveal several findings for low-data regimes. Across both datasets, the extreme low-data bottleneck ($N = 20$) proved insurmountable for all loss variants, resulting in stationary learning. However, as data scales up to $N = 50$, our proposed Softplus-Strong variant demonstrated a remarkable capacity for sample-efficient learning, particularly on the AGNEWS dataset. Under uniform random sampling (AL Off), this variant significantly outperformed the standard cross-entropy baseline in both accuracy and learning speed, suggesting that our continuous, magnitude-sensitive directional loss effectively regularizes the model by enforcing a smooth confidence drop-off along the counterfactual trajectory.

The performance hierarchy shifts for larger datasets ($N = 100, 500$) depending on the task. For binary sentiment classification (IMDB), the Gradient Supervision (GS) variant with a full mixing ratio ($r = 1.0$) achieved the global best accuracy, slightly edging out the standard cross-entropy baseline. Conversely, in the multi-class topic classification task (AGNEWS), the Softplus-Strong variant maintained its dominance across all viable training scales. This discrepancy implies that the optimal geometry for counterfactual alignment may be highly task-dependent. The strict angular penalty of GS may be well-suited for the relatively simple binary manifold of sentiment, whereas the more flexible Softplus penalty better accommodates the complex, multi-dimensional decision boundaries of topic classification.

Furthermore, we observed a strict sensitivity to the mixing parameter r. For the GS variant, reducing the proportion of counterfactually augmented data to $r = 0.5$ resulted in catastrophic learning failure on both datasets. The Softplus-Strong variant, however, exhibited remarkable robustness to this parameter shift on the AGNEWS dataset, maintaining high accuracies even with partial counterfactual supervision.

7 CONCLUSIONS

We proposed enriching supervised learning with gradient-based counterfactual annotations to embed human intuition. Experiments

Table 5: Test Accuracy (%) and Epochs for FIZLE and FLARE datasets (AL Off) across different learning approaches. GS = Gradient Supervision, C.E. = Cross-Entropy. For loss functions which utilise counterfactuals we fix the parameter, $\alpha = 0.5$. Bold values show the best accuracy for that specific loss function variant, while Underlined Bold shows the best accuracy across the entire experiment. Epochs in bold indicate that the run is the most efficient in its group, but only if it is also the most accurate.

		Dataset Size							
		20		50		100		500	
Variant	r	% Acc	Ep	% Acc	Ep	% Acc	Ep	% Acc	Ep
FIZLE Dataset									
C.E.	-	24.2 (±0.005)	0	30.2 (±0.035)	59	41.1 (±0.009)	26	69.1 (±0.019)	9
GS	1.0	24.2 (±0.005)	0	**29.6 (±0.035)**	59	**43.3 (±0.019)**	**29**	**71.0 (±0.024)**	11
	0.5	24.2 (±0.005)	0	25.7 (±0.003)	59	26.7 (±0.004)	41	27.7 (±0.004)	3
Softplus-Strong	1.0	24.2 (±0.005)	0	40.4 (±0.050)	19	**59.8 (±0.014)**	59	**75.3 (±0.009)**	**46**
	0.5	24.2 (±0.005)	0	**44.9 (±0.044)**	**18**	59.6 (±0.012)	59	75.2 (±0.010)	47
FLARE Dataset									
C.E.	-	24.2 (±0.005)	0	30.2 (±0.035)	59	41.1 (±0.009)	26	69.1 (±0.019)	9
GS	1.0	24.2 (±0.005)	0	**31.0 (±0.031)**	59	**43.9 (±0.015)**	36	**70.3 (±0.018)**	12
	0.5	24.2 (±0.005)	0	25.1 (±0.004)	47	26.3 (±0.003)	27	29.8 (±0.007)	3
Softplus-Strong	1.0	24.2 (±0.005)	0	45.0 (±0.063)	21	**62.8 (±0.019)**	59	**77.8 (±0.006)**	43
	0.5	24.2 (±0.005)	0	**48.7 (±0.062)**	22	62.5 (±0.008)	59	77.3 (±0.007)	41

Table 6: Test Accuracy (%) and Epochs for FIZLE and FLARE datasets (AL On). GS = Gradient Supervision, C.E. = Cross-Entropy. For loss functions which utilise counterfactuals we fix the parameter, $\alpha = 0.5$. Bold values show the best accuracy for that specific loss function variant, while Underlined Bold shows the best accuracy across the entire experiment. Epochs in bold indicate that the run is the most efficient in its group, but only if it is also the most accurate.

		Dataset Size							
		20		50		100		500	
Variant	r	% Acc	Ep	% Acc	Ep	% Acc	Ep	% Acc	Ep
FIZLE Dataset									
C.E.	-	24.2 (±0.005)	0	**27.1 (±0.008)**	31	43.4 (±0.048)	38	60.9 (±0.006)	94
GS	1.0	24.2 (±0.005)	0	**27.1 (±0.023)**	59	**38.0 (±0.043)**	**23**	**63.5 (±0.006)**	56
	0.5	24.2 (±0.005)	0	24.3 (±0.012)	29	32.3 (±0.003)	93	27.3 (±0.023)	9
Softplus-Strong	1.0	24.2 (±0.005)	0	25.6 (±0.012)	42	**45.7 (±0.039)**	69	57.6 (±0.050)	24
	0.5	24.2 (±0.005)	0	**25.7 (±0.026)**	57	45.2 (±0.030)	66	**59.6 (±0.026)**	87
FLARE Dataset									
C.E.	-	24.2 (±0.005)	0	27.1 (±0.008)	31	40.1 (±0.045)	31	63.4 (±0.006)	67
GS	1.0	24.2 (±0.005)	0	**27.2 (±0.006)**	31	**48.0 (±0.030)**	**46**	**63.8 (±0.017)**	34
	0.5	24.2 (±0.005)	0	24.2 (±0.013)	0	27.0 (±0.005)	57	26.7 (±0.017)	11
Softplus-Strong	1.0	24.2 (±0.005)	0	**28.2 (±0.027)**	59	46.4 (±0.045)	54	**63.9 (±0.024)**	32
	0.5	24.2 (±0.005)	0	27.0 (±0.024)	58	**46.9 (±0.025)**	71	60.1 (±0.037)	22

on synthetic topologies and NLP tasks demonstrated that penalizing gradients along counterfactual paths accelerates training and outperforms standard baselines in data efficiency, particularly using our **Softplus-Strong** variant, with the exception of experiments on the IMDB dataset. Comparing FIZLE and FLARE suggests that counterfactual quality is the major influencing factor, and future work

should further investigate counterfactual quality. Improvements in the entropy-based active learning setting were less significant than with full datasets, although this discrepancy *may* grow with further training epochs.

While our findings demonstrate the potential of gradient-based counterfactual alignment, several limitations remain. **Annotation cost** is high, as generating counterfactuals is labor-intensive; thus, future work must determine the minimal number required. Furthermore, there is inherent **subjectivity**, as models absorb annotator bias, requiring methods to estimate expert skill. **Generalization** also poses a challenge, as translating directional vectors to complex domains (e.g., high-resolution imaging) or non-linear paths remains difficult. Finally, there is a strict **quality dependence**, as the efficacy of our method is bound by counterfactual quality, necessitating better filtering metrics.

ACKNOWLEDGMENTS

Jonathan Erskine's PhD research is jointly funded by UK Research and Innovation (UKRI) and Thales Training & Simulation Ltd. 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

[1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. arXiv:1906.03671 [cs.LG] https://arxiv.org/abs/1906.03671

[2] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Towards LLM-guided Causal Explainability for Black-box Text Classifiers. arXiv:2309.13340 [cs.CL] https://arxiv.org/abs/2309.13340

[3] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation. In *2024 IEEE International Conference on Big Data (BigData)*. 1243–1248. https://doi.org/10.1109/BigData62323.2024.10825537

[4] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Annie Zaenen and Antal van den Bosch (Eds.). Association for Computational Linguistics, Prague, Czech Republic, 440–447. https://aclanthology.org/P07-1056

[5] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237. https://doi.org/10.1136/bmjqs-2018-008370 arXiv:https://qualitysafety.bmj.com/content/28/3/231.full.pdf

[6] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch Active Learning at Scale. arXiv:2107.14263 [cs.LG] https://arxiv.org/abs/2107.14263

[7] Paul Doucet, Benjamin Estermann, Till Aczel, and Roger Wattenhofer. 2025. Bridging Diversity and Uncertainty in Active learning with Self-Supervised Pre-Training. arXiv:2403.03728 [cs.LG] https://arxiv.org/abs/2403.03728

[8] Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[9] David J. Hand. 2006. Classifier Technology and the Illusion of Progress. *Statist. Sci.* 21, 1 (2006), 1 – 14. https://doi.org/10.1214/088342306000000060

[10] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[11] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sklgs0NFvr

[12] Divyansh Kaushik, Amrith Setlur, Eduard H Hovy, and Zachary Chase Lipton. 2021. Explaining the Efficacy of Counterfactually Augmented Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=O9ngzn09Bmc Poster.

[13] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and Understanding Neural Models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://www.aclweb.org/anthology/N16-1082

[14] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[16] Michael Prince. 2004. Does active learning work? A review of the research. *Journal of engineering education* 93, 3 (2004), 223–231.

[17] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*. PMLR, 8116–8126.

[18] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2662–2670.

[19] Berkman Sahiner, Wesley Chen, Ravi K. Samala, and Nicholas Petrick. 2023. Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology* 20220878 (2023). https://doi.org/10.1259/bjr.20220878

[20] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. arXiv:1708.00489 [stat.ML] https://arxiv.org/abs/1708.00489

[21] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison. http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[22] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27 (1948), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

[23] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep Active Learning for Named Entity Recognition. arXiv:1707.05928 [cs.CL] https://arxiv.org/abs/1707.05928

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*.

[25] Kacper Sokol, Alexander Hepburn, Rafael Poyiadzi, Matthew Clifford, Raul Santos-Rodriguez, and Peter Flach. 2020. FAT Forensics: A Python Toolbox for Implementing and Deploying Fairness, Accountability and Transparency Algorithms in Predictive Systems. *Journal of Open Source Software* 5, 49 (2020), 1904. https://doi.org/10.21105/joss.01904

[26] Vinicius M. A. Souza, Denis M. dos Reis, André G. Maletzke, and Gustavo E. A. P. A. Batista. 2020. Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery* 34 (2020), 1805–1858.

[27] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*. Springer, 580–599.

[28] Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. 2025. Truth or Twist? Optimal Model Selection for Reliable Label Flipping Evaluation in LLM-based Counterfactuals. arXiv:2505.13972 [cs.CL] https://arxiv.org/abs/2505.13972

[29] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

[30] Xiaojin Zhu. 2015. 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). https://doi.org/10.1609/aaai.v29i1.9761

[31] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James S. Duncan. 2020. AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients. arXiv:2010.07468 [cs.LG] https://arxiv.org/abs/2010.07468

A ENTROPY-BASED IMPORTANCE SAMPLING WITH DIVERSITY

Where gradients have been used previously as a natural method for providing human-interpretable explanations [24], this paper aims to reverse this pipeline and provide machine-interpretable supervision.

To ensure that our framework selects the most informative samples for counterfactual annotation while maintaining diversity

across the feature space, we employ an entropy-based importance sampling strategy with an integrated diversity parameter. This approach addresses the common limitations of purely uncertainty-based or diversity-based sampling methods by combining their respective strengths [7].

Implementation Details. Our sampling procedure operates iteratively during training. At each selection round, we first compute the softmax probabilities from the current model's logits:

$$\mathbf{p}_i = \text{softmax}(\mathbf{f}_\theta(\mathbf{x}_i)) \tag{9}$$

where $\mathbf{f}_\theta$ represents our neural network model with parameters θ. The uncertainty of each sample is then quantified using the entropy of these probability distributions:

$$U(\mathbf{x}_i) = -\sum_{y \in \mathcal{Y}} p_{i,y} \log p_{i,y} \tag{10}$$

where $p_{i,y}$ denotes the predicted probability for class y given input $\mathbf{x}_i$ [22]. Samples with high entropy values indicate regions where the model exhibits significant classification uncertainty, making them prime candidates for informative annotations.

However, relying solely on uncertainty sampling can lead to the selection of similar samples clustered near decision boundaries, potentially overlooking diverse regions of the feature space that could provide broader representational coverage [21]. To address this limitation, we incorporate a diversity parameter β that encourages the selection of samples spanning different regions of the feature space.

Our `select_informative_samples` function implements a hybrid selection strategy that operates in two phases. First, we identify candidate samples with high uncertainty scores based on the entropy measure above. Second, we apply diversity-aware selection using the learned embeddings $\mathbf{e}_i$ from our model's intermediate representations:

$$\text{Selected} = \text{select\_informative\_samples}(\mathbf{P}, \mathbf{E}, k, \beta) \tag{11}$$

where $\mathbf{P}$ contains the softmax probabilities, $\mathbf{E}$ represents the embedding matrix, k specifies the number of samples to select, and β controls the diversity weighting parameter.

We combine uncertainty and representativeness measures through a weighted formulation. For each sample $\mathbf{x}_i$, we compute its representativeness $Rep(\mathbf{x}_i)$ based on its similarity to other unlabeled instances in the embedding space:

$$Rep(\mathbf{x}_i) = \frac{1}{|U|} \sum_{\mathbf{x}_j \in U} \exp\left(-\frac{||\mathbf{e}_i - \mathbf{e}_j||^2}{2\sigma^2}\right) \tag{12}$$

where U represents the unlabeled dataset, $\mathbf{e}_i$ and $\mathbf{e}_j$ are the learned embeddings, and σ controls the bandwidth of the similarity function. The combined information content is then:

$$Info(\mathbf{x}_i) = U(\mathbf{x}_i)^\alpha \cdot Rep(\mathbf{x}_i)^\beta \tag{13}$$

where α and β are weighting parameters that balance uncertainty and representativeness considerations, with β corresponding to our `diversity_weight` parameter.

This iterative process ensures that our model continuously improves its understanding of challenging regions while maintaining diversity across the feature space. The entropy-based uncertainty measure ensures we focus annotation efforts on the most ambiguous classification boundaries, while the diversity parameter guarantees that our selected samples represent the global structure of the data distribution.

For our gradient supervision framework, this balanced selection is particularly important, as diverse counterfactual directions provide richer supervision signals that can improve model generalization across the entire feature space rather than just local decision boundaries [11]. The use of learned embeddings for diversity computation is especially valuable, as it ensures that similarity is measured in the model's internal representation space rather than raw input space, leading to more semantically meaningful diversity constraints.