Uncertainty in Deep Learning

Candidate Number 1091797

1 Introduction

Active Learning (AL) frameworks solve a key problem in machine learning applications: data labelling. By training a model on a small dataset and asking an expert to label samples based on uncertainty quantification, we reduce data acquisition costs. Nonetheless, AL's reliance on uncertainty modelling and small datasets limits its usability with high-dimensional data such as images. This prevents fields with high data acquisition costs such as healthcare to benefit from AL frameworks.

To tackle this challenge, this paper leverages Bayesian deep learning. While methods like hierarchical parametrised basis functions provide a satisfactory baseline, Bayesian convolutional neural networks (BCNNs) exceed 95% test accuracy on MNIST with 390 acquired images, a $2\times$ improvement.[1] Given potential label noise, we also design a novel label likelihood filter. In addition, we add clustering to maximise diversity and speed up convergence, yielding a 21% acceleration for 95% test accuracy.

2 Background

Previous literature approaches active learning of image data in various ways. Originally, kernel-based methods obtained model uncertainty using margin-based classification (Joshi et al., 2009) and Gaussian processes (Li and Guo, 2013). Other approaches like Gao et al. (2020) deviated slightly to explore semi-supervised methods, learning the unlabelled data distribution to enhance accuracy on the labelled set. Following this, the emergence of Bayesian CNNs (Gal and Ghahramani, 2015) yielded convincing results. By placing a prior distribution over model parameters and performing approximate inference with Dropout (Srivastava et al., 2014)—an equivalent to variational Bayesian approximation as proven in Gal and Ghahramani (2016)—one can extract uncertainty from small datasets. These models are then combined with acquisition functions that select informative samples using measurements such as standard deviation, variation ratios, entropy, etc (Gal et al., 2016).

More recently, efforts have focused on data acquisition via novel uncertainty quantification and increased sample diversity (Tharwat and Schenck, 2023). Based on information theory, modern methods maximise information gain by targeting epistemic uncertainty (information gaps) through spatial embeddings and acquisition functions. These approaches can also be combined with clustering techniques for improved representativeness in the initial training set (Yan et al., 2022). In the case of noisy datasets (i.e. experts that provide incorrect labels), Bayesian networks allow for the development of noise filters based on label likelihood (Nagarajan et al., 2024). Those efforts reveal useful in applications with less than 95% accuracy such as medical diagnoses of Alzheimer's disease.

3 Methodology

To reproduce Section 5.1 of Gal et al. (2017), we reimplement all of the acquisition functions (BALD, variation ratios, max entropy, mean standard deviation, and random acquisition). Then, we recreate the same Bayesian CNN (convolution-ReLU-convolution-ReLU-max pooling-Dropout-dense-Dropout-dense-softmax) and integrate it within the active learning pipeline, composed of four phases: training, scoring, dataset updating, and evaluation. For Dropout approximation, we use 100 MC samples—based on Gal and Ghahramani (2016) which suggests sampling rates above $T = 10$. Lastly, we reproduce Section 5.2 of Gal et al. (2017) to compare our results with an equivalent deterministic approach which removes Dropout at test time, preventing uncertainty quantification.

[1] Full code available at `www.github.com/anonymous-account-67/UncertaintyInDeepLearning`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

4 Minimum extension

In this section, we add a new baseline to the MNIST experiments using a hierarchical parametrised basis function regression model. This concept from Rumelhart et al. (1985) composes a feature vector $\varphi(\mathbf{x})$ with learnt weights as $f(\mathbf{x}) = \mathbf{W}^T\varphi(\mathbf{x})$ to obtain the final output. To achieve this, we change to a regression task with continuous multi-dimensional outputs by converting one-hot encodings into vectors in $\mathbb{R}^{10}$. We take the same architecture as in previous experiments but freeze layers up to the penultimate one to obtain a feature matrix $\Phi \in \mathbb{R}^{N\times 128}$ (with rows $\varphi(\mathbf{x}_n)^T$). We replace the last dense layer with a Bayesian linear regression layer, for which we derive the posterior $p(\mathbf{W}|\mathcal{D})$ and predictive distribution $p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D})$. Detailed derivations can be found in Appendix A.

For analytic inference, we assume a Matrix Normal posterior $p(\mathbf{W}|\mathcal{D}) = \mathcal{MN}_{128\times 10}(\mu', \Sigma', \Sigma)$. Given the predictive mean $\mu^* = \mu'^T\phi(\mathbf{x}^*)$ and scalar predictive variance component $var^* = \sigma^2 + \varphi(\mathbf{x}^*)^T\Sigma'\varphi(\mathbf{x}^*)$, the predictive distribution for a new input $\mathbf{x}^*$ is a multivariate Gaussian:

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D}) = \mathcal{N}(\mu^*, var^*\Sigma)$$

For analytic MFVI inference, we approximate the true posterior $p(\mathbf{W}|\mathcal{D})$ with a factorised distribution $q_{m,\sigma}(\mathbf{W})$ assuming weight independence to maximise the ELBO. Using the optimal variational parameters, the predictive variance for class d becomes:

$$\mathrm{Var}(y_d^*) = \sigma^2 + \sum_{k=1}^{128} \varphi_k(\mathbf{x}^*)^2\sigma_{k,d}^2$$

Here, we use scalar predictive variance in acquisitions. To recover a scalar from the predictive covariance matrix, we decide to take its trace. This captures epistemic uncertainty by aggregating marginal variances. Compared to other tools like the determinant, it is more robust to singularities and directly minimises our RMSE.

5 Novel extension

This section aims to enhance the capabilities of deep Bayesian active learning by diving deeper into two problems which are essential to integrating AL frameworks within real-world applications.

5.1 Data robustness and cleaning

To investigate acquisition functions' robustness to label noise, we start by adding label noise to the pool of examples and analyse the results. At each acquisition step, we add a probability of a mislabel and plot the results, comparing BALD—our most performant acquisition function—to the random acquisition, for varying levels of noise.

Then, to reject mislabels in a noisy pool, we design a filter that acquires points maximising BALD while avoiding noise. Formally, given y_{pool} the noisy label from the expert and a threshold τ, we rely on the posterior predictive $p(y|x, \mathcal{D}_{train})$ approximated by MC Dropout to define a trust set

$$\{x \in \mathcal{D}_{pool} | p(y = y_{pool}|x, \mathcal{D}_{train}) \geq \tau\}$$

5.2 Improving sample representation with clustering and diversity-aware acquisitions

To improve speed of convergence from the initial training set onwards, it is essential to move away from the random but balanced approach for a more representative method. Ensuring that the model sees both ways to write the digit "1" from the very start seems crucial. In our novel extension, we thus use k-means clustering to choose the two most representative samples within each class.

Speed of convergence can also be improved through diversity-aware acquisitions. At acquisition, a model has specific weaknesses. The issue is that similar images get similar scores, despite information overlap. Furthermore, reducing acquired samples per step slows down training without addressing the underlying problem. Here, we thus compare pool samples using pairwise cosine similarity $\frac{\langle X_i, X_j\rangle}{||X_i||\ ||X_j||}$ to reduce information overlap by adding a penalty that favours diversity within acquisition steps.

6 Experiments

6.1 Reproduction of Gal et al. (2017)

After redefining the acquisition functions, the Bayesian CNN, and the active learning pipeline using PyTorch (Paszke et al., 2019), we study the performance of various acquisition functions. Experiments use identical training parameters as in Gal et al. (2017) and are averaged over three repetitions. As seen in Figure 1, standard acquisition functions visibly outperform random acquisition. Variation Ratios is the fastest, exceeding 95% test accuracy on MNIST with 390 acquired images, a $2\times$ improvement. BALD is the most stable and performant with 3% better test accuracy over random acquisition. Table 1 also shows the number of acquisition steps needed to get 5% and 10% test errors.

To assess the importance of modelling uncertainty, we also compare our results with deterministic equivalents of our Bayesian CNNs. As shown in Figure 2, Bayesian models converge to a higher accuracy, showing that propagated uncertainty has a significant effect on measure of confidence.

6.2 Minimum extension experiments

We implement a hierarchical parametrised basis function regression model and perform analytic and analytic MFVI inference using the mathematical equations derived in Section 4. Those methods differ in assumption of independence. While analytic inference provides exact posterior for the weights with the full covariance matrix, the mean field assumption implies independent epistemic uncertainties of features. In practice, this gives MFVI lower predictive variance (indicating higher confidence) despite higher RMSE, as shown in Figure 3. Since KL-divergence minimisation approximation is mode-seeking, MFVI underestimates the variance of the posterior.

6.3 Novel extension experiments

To study the robustness of BALD, our most performant acquisition function, to label noise, we start by adding mislabelling probability at each acquisition step and comparing the impact for various noise levels. Results shown in Figure 4 reveal how BALD confuses noise (aleatoric uncertainty) for informative samples (epistemic uncertainty), performing worse than random acquisition over the first acquisition steps and struggling to outperform it when noise exceeds 10% noise—indicating that AL frameworks are highly sensitive to noise.

Adding a filter based on label likelihood to discard samples for which the model is confident enough that their label is wrong, we manage to mitigate noise slightly as shown in the Figure 5 comparison. Nonetheless, despite comparing various thresholds options using grid search and selecting the best one as shown in Figure 6, our results indicate that models tend to over-filter (c.f. Table 3), potentially losing out on information. In addition, witnessing a performance decrease when filtering a pool set without noise also reveals the severe impact of incorrectly rejected samples, which prevent the model from learning from genuine samples that are contradictory with its beliefs.

Using k-means to increase representativeness of numbers in the initial training set, we witness a 20% increase in test accuracy. Using cosine similarity, we can study various diversity penalty coefficients. As shown in Figure 7, giving a small priority to diverse samples increases performance by about 5% up until the fifth acquisition steps. With more samples, the advantage disappears quickly and can even become prohibitive for overly large penalties.

7 Analysis & conclusions

Our experiment reproductions show that Bayesian inference can be combined with deep learning methods to improve active learning frameworks. Bayesian CNNs can accurately model uncertainty in order to query the most informative images. Our results only deviate from the original paper for mean standard deviation, which performs better than random acquisition in our experiments.

In addition, we see that hierarchical parametrised basis function regression provides a satisfactory baseline with analytic inference but that adding mean-field variational inference is clearly suboptimal.

Lastly, our novel experiments show that clustering and diverse acquisitions enable faster convergence, a significant benefit given limited labelled data. Our results also suggest that, despite a filter mitigating the impact of noise, acquisition functions struggle to distinguish informative from erroneous samples.

A Minimum extension: mathematical derivations

A.1 Analytic inference

In this section, we detail the mathematical derivations used for analytic inference within the minimum extension. Our computations are based on slides 49-56 of lectures 3-4 extended to multivariate outputs as well as on several assumptions from the Q&A.

A.1.1 Likelihood and prior

Here, our goal is to derive the posterior mean and variance (c.f. slide 49) for the last layer weights $\mathbf{W} \in \mathbb{R}^{128\times 10}$

Let N the number of data samples. We start by assuming that the output classes are correlated through a covariance matrix $\Sigma \in \mathbb{R}^{10\times 10}$.

The likelihood of observing a matrix $\mathbf{Y}$ given weight $\mathbf{W}$ becomes

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \mathcal{MN}_{N\times 10}(\Phi\mathbf{W}, \sigma^2\mathbf{I}_N, \Sigma)$$

Here, this tells us that the samples are independent but that output classes are correlated by the covariance matrix.

Then, for the prior, we place a Matrix Normal prior on $\mathbf{W}$:

$$p(\mathbf{W}) = \mathcal{MN}_{128\times 10}(\mathbf{0}, s^2\mathbf{I}_{128}, \Sigma)$$

A.1.2 Posterior

We can now derive the posterior. Since $p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) \times p(\mathbf{W})$, the posterior also has a Matrix Normal distribution:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) = \mathcal{MN}_{128\times 10}(\mu', \Sigma', \Sigma)$$

where μ' and Σ' are the posterior (row) covariance and mean that we need to derive as in slide 49 of the lecture slides by extending to multivariate outputs (i.e. to matrices).

Therefore, the posterior covariance which models the uncertainty over features in a 128 × 128 matrix is $\Sigma' = (\sigma^{-2}\Phi^T\Phi + s^{-2}\mathbf{I}_{128})^{-1}$ and the posterior mean is $\mu' = \Sigma'(\sigma^{-2}\Phi^T\mathbf{Y})$

A.1.3 Predictive

As in slide 56 of the lecture slides, for a new input $\mathbf{x}^*$, the predictive distribution for $\mathbf{y}^*$ is a Multivariate Gaussian (which is equivalent to a 1×10 Matrix Normal) defined as

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D}) = \mathcal{N}(\mu^*, \mathbf{\Sigma}^*) \quad (= \mathcal{MN}_{1\times 10}(\mathbf{y}^*|\mu^*, var^*, \Sigma))$$

where the predictive mean is $\mu^* = \mu'^T\varphi(\mathbf{x}^*)$ and the predictive covariance effectively scales the class correlations by the predictive scalar variance $var^* = \sigma^2 + \varphi(\mathbf{x}^*)^T\Sigma'\varphi(\mathbf{x}^*)$, yielding

$$\mathbf{\Sigma}^* = (\sigma^2 + \varphi(\mathbf{x}^*)^T\Sigma'\varphi(\mathbf{x}^*)) \times \Sigma$$

A.2 Analytic MFVI inference

In order to perform inference for this new Mean Field Variational Inference baseline, this section also extends the derivation in slides 59-68 in lecture 5-6 to multivariate outputs.

Here, our goal is to approximate the true posterior $p(\mathbf{W}|\mathcal{D})$ with a different distribution $q_{m,\sigma}$ that assumes weight independence. As in the slides, it is defined as

$$q_{m,\sigma}(\mathbf{W}) = \prod_{k=1}^{128}\prod_{d=1}^{D}\mathcal{N}(w_{kd}|m_{kd}, \sigma_{kd}^2)$$

Now, the objective is to maximise the ELBO with respect to M and S (where the definition of M and S is taken from the slides), which has the following form

$$\mathcal{L}(\mathbf{M}, \mathbf{S}) = \mathbb{E}_q[\ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W})] - \mathrm{KL}(q(\mathbf{W})||p(\mathbf{W}))$$

Note that we assume that weights are fully factorized, implying that we must ignoring class correlations Σ to replace them with independent noise $\sigma^2\mathbf{I}$.

A.2.1 Computing the expected log-likelihood

As in slide 62, the expected log-likelihood is

$$\mathbb{E}_q[\ln p] = -\frac{1}{2\sigma^2}\left(\sum_{n=1}^{N} ||\mathbf{y}_n - \mathbf{M}^T\phi_n||^2 + (\sum_{d=1}^{D}\sum_{k=1}^{128}\phi_{nk}^2\sigma_{kd}^2)\right) - \frac{ND}{2}\log 2\pi\sigma^2$$

Here, the difference is that cross-terms of the diagonal weight covariance matrix cancel out under mean-field factorisation, simplifying the whole expectation to a sum of variances. Also, since out goal is to maximise expected log likelihood, our goal can be simplified to minimizing the term in between parenthesis.

A.2.2 Computing the Kullback-Leiber divergence

As in the slides, given prior variance s^2 and Gaussian prior $p(w_{kd}) = \mathcal{N}(0, s^2)$, we have

$$\mathrm{KL}(q||p) = \sum_{k,d}\frac{1}{2}\left(s^{-2}\sigma_{kd}^2 + s^{-2}m_{kd}^2 - 1 + \ln\frac{s^2}{\sigma_{kd}^2}\right)$$

A.2.3 Optimal variational posterior

To find the optimal parameters for the variational posterior q, we use the partial derivative of the ELBO $\frac{\partial\mathcal{L}}{\partial M}$ and $\frac{\partial\mathcal{L}}{\partial\sigma^2}$

Setting the partial derivative to zero and solving for the posterior mean yields an optimal solution identical to μ' in the Analytic Inference method:

$$\mathbf{M}^* = (\sigma^{-2}\Phi^T\Phi + s^{-2}\mathbf{I}_{128})^{-1}(\sigma^{-2}\Phi^T\mathbf{Y})$$

For the optimal posterior variance σ_{kd}^2, we set

$$\frac{\partial\mathcal{L}}{\partial\sigma_{k,d}^2} = -\frac{1}{2}\left(\sigma^{-2}\sum_n\phi_{nk}^2 + s^{-2} - \sigma_{k,d}^{-2}\right) = 0$$

We then solve the equation and get

$$\sigma_{kd}^2 = (\sigma^{-2}\sum_{n=1}^{N}\phi_{nk}^2 + s^{-2})^{-1}$$

A.2.4 Predictive distribution

As in slide 68, we get that for a new input $\mathbf{x}^*$, the predictive mean is $\mathbf{y}^* = \mathbf{M}^{*T}\varphi(x^*)$. However, since the weights are independent, the variance simplifies to a sum of squares. We can thus express the predictive variance for class d as:

$$\mathrm{Var}(y_d^*) = \sigma^2 + \sum_{k=1}^{128}\varphi_k(\mathbf{x}^*)^2\sigma_{k,d}^2$$

B Experiment results

B.1 Reproduction of Gal et al. (2017)

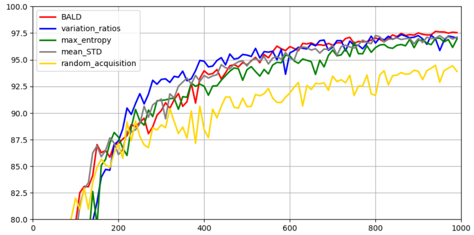

Figure 1: MNIST test accuracy as a function of number of acquired images from the pool set.

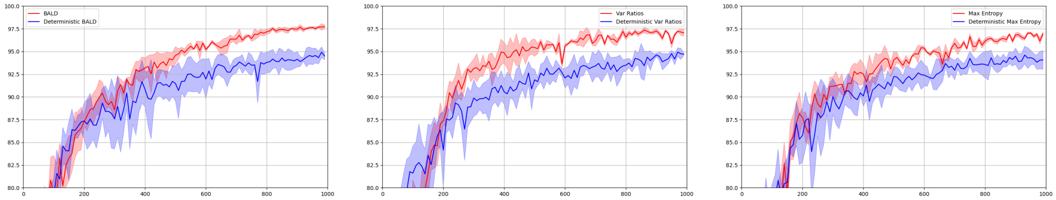

Figure 2: Test accuracy as a function of acquired images, using both a Bayesian CNN and a deterministic CNN.

% error	BALD	Var Ratios	Max Ent	Mean STD	Random
10%	105	115	125	230	385
5%	260	265	365	510	940

Table 1: Number of acquired images to get to model error of % on MNIST

B.2 Minimum Extension

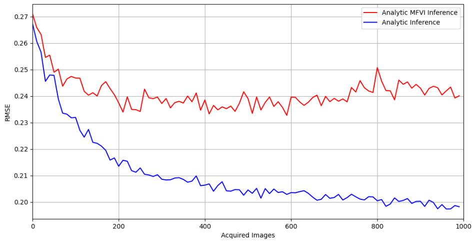

Figure 3: Root Mean Squared Error (RMSE) as a function of number of acquired images from the pool set

B.3 Novel Extension

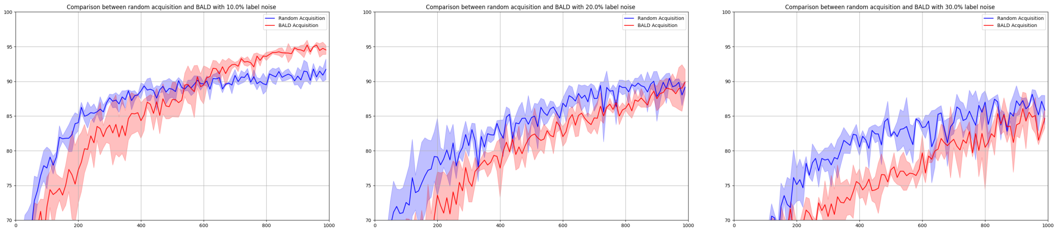

Figure 4: Test accuracy as a function of acquired images with label noise, using both BALD and random acquisitions

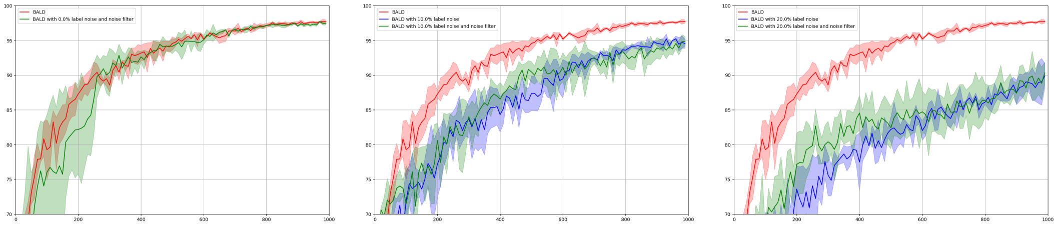

Figure 5: Test accuracy as a function of acquired images with label noise, using both BALD, filtered BALD, and random acquisitions

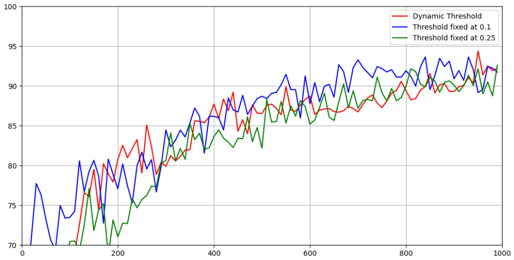

Figure 6: Test accuracy as a function of acquired images for various threshold designs and 20% label noise

Threshold	Rejects	Excess over optimal
Dynamic τ	2.98	+49%
$\tau = 0.1$	3.18	+59%
$\tau = 0.25$	5.10	+155%

Table 2: Average samples rejected per step for various threshold designs and 20% label noise

Label noise (%)	Avg. rejected samples	Excess over optimal
0%	0.39	0.39 samples
10%	1.47	+47%
20%	2.98	+49%
30%	4.96	+65%

Table 3: Average number of samples rejected per acquisition step at various noise levels

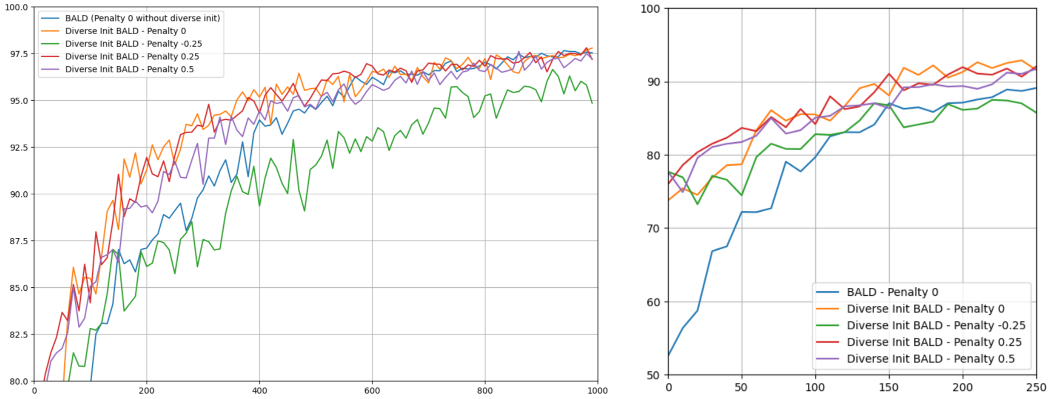

Figure 7: Test accuracy as a function of acquired images for various diversity penalty coefficients - full experiment (left) and early steps (right). Positive penalty coefficients indicate a preference for diversity, while negative penalty coefficients indicate a preference for uniformity.

References

Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, 2017. URL `https://arxiv.org/abs/1703.02910`.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009. doi: 10.1109/CVPR.2009.5206627.

Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 859–866, 2013.

Bhalaji Nagarajan, Ricardo Marques, Eduardo Aguilar, and Petia Radeva. Bayesian dividemix++ for enhanced learning with noisy labels. *Neural Networks*, 172:106122, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106122. URL `https://www.sciencedirect.com/science/article/pii/S0893608024000364`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL `https://arxiv.org/abs/1912.01703`.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, 1985.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

Xuyang Yan, Shabnam Nazmi, Biniam Gebru, Mohd Anwar, Abdollah Homaifar, Mrinmoy Sarkar, and Kishor Datta Gupta. A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, 52(11):13250–13267, 2022.