Deep Active Learning In The Presence Of Label Noise: A Review

Moseli Mots'oehli[0000−0002−9191−0565]

Department of Information and Computer Science,
University of Hawai'i at Manoa
moselim@hawaii.edu

1 Introduction

Machine learning algorithms are a sub-class of artificial intelligence that learn from data to perform a pre-defined task such as classification, regression, or clustering. Of the numerous algorithms for machine learning, artificial neural networks, deep neural networks in particular have done exceptionally well in tasks involving complex data representations such as images, text, and sound. The main reason for this was the discovery that if you have a large enough dataset, you can build more extensive and more complex models with little to no risk of over-fitting. While this works in theory, the practical applications have major drawbacks such as the need for labeled training examples that come at a high cost due to the time needed to label the data, the high cost of labor in very specialized fields, or the cost of running simulations that would produce the ground truth dataset. The solution comes in the form of Deep active learning (DAL) algorithms, which strive to let the learning algorithm iteratively pick data examples to be labeled from a larger unlabelled dataset, in such a manner that results in (1) a smaller labeled training set, (2) that is representative of the underlying data distribution leading to a near-optimal learner, (3) while at the same time not exceeding the labeling budget.

While this works well for most use cases, Real industry dataset labelling has inherent label noise due to a variety of factors such as redundant observations being labelled differently, the best human expert classification performance being low or use of auto-labelling software such as Mechanical Turk. This has adverse effects on these DAL algorithm's performance, and most existing DAL literature focus on noise free settings. We explore existing literature around the problem of using DAL algorithms in the existence of label noise. We are particularly interested in the image detection, segmentation and classification domain using different deep representation learning frameworks such as Deep Neural networks, convolutional neural networks (CNNs), and vision transformer networks. We conclude by exploring possible directions for future research in DAL in vision tasks under label noise.

2 Preliminary

In this section, we briefly describe Deep Learning (DL) focusing on the most important architectural choices for computer vision. The active learning (AL) framework for machine learning follows including key approaches for training deep learning algorithms on a labeling budget in the case of clean labels, and finally the scene is set for label noise and the literature addressing DL on noisy labels.

2.1 Deep Learning

Deep Learning (DL) refers to the use of Artificial Neural Networks with multiple hidden layers [43] to approximate known or unknown functions $f\colon X \mapsto Y$. Over the years, different domain specific DL architectures have been developed to enhance the quality of the learned representations from the different data Modalities. Early research focused on improving optimization, custom layers and connections, activation functions, loss functions and hyper-parameter tuning techniques for the multi-layer perceptron as a way to improve performance on different data modalities. For tabular data, tree based ensemble learning algorithms such as Random forest [5], XGBoost [6], and CatBoost [41] are preferred over DL for their superior performance and resource efficiency. A non-exhaustive selection of interesting neural network adaptations to tabular data includes [44,42,40,2,3]. In the Natural language processing domain, earlier work involved learning word and sentence representation using shallow neural networks in an unsupervised setting [37,35,19]. Until the wide adoption of attention based transformer language models [49,45], word and sentence level embeddings are fed to a deep neural network with recurrent connections such as a Long-Short-Term-Memory(LSTM) network [24] to achieve state-of-the-art results on down-stream text classification, sentence completion, named entity recognition or summarization tasks. For static visual tasks such as image classification, regression and segmentation, CNN based architectures with specialized layers and prepossessing transformations were eventually superseded by vision transformer models [25].

Since this work focuses on DL algorithms for vision tasks, we explore models for supervised classification, regression, and segmentation. We first give a brief overview of CNNs that are responsible for a large share of progress in vision based tasks. We then highlight the use of hybrid LSTM and CNN networks for tasks with visual and temporal properties, such as is the case for video based classification. Finally we highlight literature on state of the art (S.O.T.A) spatial attention based models in the context of visual tasks (Vision Transformers) [25].

Convolutional Neural Networks : Convolutional neural networks [27] were introduced by Yann Lecun and Yashua Bengio as a replacement for human feature extraction in training multi-layer neural networks on spacial data

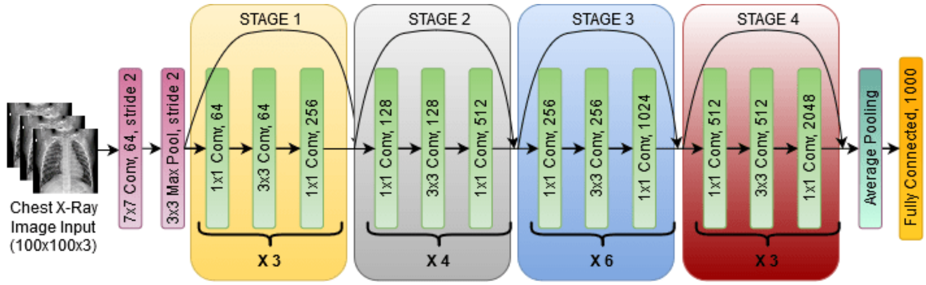

Fig. 1: [Source: Resnet50] architecture with convolutional blocks of different filter sizes and max pooling. Residual connections acting as memory cells arch above the blocks passing initial information all the way to the final layer.

[28]. The key deficiencies with training fully connected feed-forward networks (FFNN) using back-propagation for computer vision are efficiency and transformation (rotation, translation) invariance. Handling high-dimensional image input data given that very low playground resolution images are 28×28, leading to an input layer with 784 neurons. If the second layer had as little as 100 neurons, the 2-layer network has more than 78400 weights and bias terms already. The number of input neurons and hidden layer depth to approximate input-output mappings on high-resolution real-world image classification problems using FFNNs is exponentially higher than in the 28×28 case. This can lead to over-fitting on smaller datasets, and would mean the application cannot run on most low-resource hardware.

Pre-processing and normalization operations applied to images move pixels around so that feature extraction based on fixed input neurons fails for non-conforming images. CNNs solve this by learning lower-dimensional filters through convolution and pooling operations, allowing the network to learn features based on local pixel proximity. More advanced CNNs have multiple convolution and pooling blocks (earlier layers capture low level features, deeper layers capture higher level features), residual connections to provide input level context [23] to higher level blocks as depicted in fig:ResNet50. The different layers are connected by non-linear activation functions such as the popular ReLU and Elu [34,10]. CNNs have been the dominant approach on computer vision benchmarks for a large part of the last decade mainly for addressing issues with FFNNs, the introduction of large training datasets, advances in computing hardware, and reduction in the cost computational cost of training such models. These networks have been used in the feature extraction pre-training step of most fine-tuned S.O.T.A approaches in different vision tasks.

Vision Transformers : Before full transformer models in the language domain, the best LSTM models used a low dimensional vector representation to passed information from an encoder network to a decoder network, while using an attention mechanism. Attention in this setting is used to learn what parts of an input sequence are most important IN predicting different parts of the output. In the original paper "Attention is all you need" [45], the authors demonstrate long temporal dependencies can be learned without the need for recurrence. The three fundamental components in a transformer network are positional encoding of tokens, the attention, and self attention mechanisms. Positional encoding of both input and output tokens is achieved by assigning integer values to tokens/words based on their relative position in the input and output sequences. Unlike LSTMs, the work of learning word progression and relationships between input and output words is learned implicitly by the network instead of designing networks with explicit recurrent cells and sequential processing. Self-attention makes it possible to learn good representations for any language given a sufficiently large collection of text in a semi-supervised manner by masking tokens and letting the network learn what the missing word is in any given input sequence. The learned representations are then used on a down-stream task with fewer labelled data. Because transformers do not process input tokens in sequence, they are perfect for parallel GPU training.

Like most great innovations, the fundamental ideas of the transformer have been incorporated into CNNs [52,54,13,48], and in some cases completely replacing CNNs to produced S.O.T.A results in various computer vision benchmarks [25,53,8]. These models are designed in a modular fashion to easily be able to learn both language and image representations for image captioning, classification, scene-text understanding, and visual question answering. [53] Is particularly interesting since the authors present a joint constrastive loss (image to text and text to image), image classification loss and image to language captioning loss, allowing for efficient training of a single network for multiple tasks, and the ability to transfer the learn representations to a different downstream task and dataset.

2.2 Active learning

In most supervised machine learning use cases, there is an initial data collection and labeling cost, in both money and time. In some domains and tasks, datasets are inherently difficult to label for a variety of reasons, meaning more time is needed even by an expert human annotator to assign a label to each sample. In other cases the cost of hiring expert annotators is high, such as is the case in medical imaging [18,26] , or the cost of producing the samples is high, such as is the case in experimental physics where observations come from very expensive telescopes or particle accelerators. This presents a challenge to the real world use of machine learning systems, especially as unlabeled dataset sizes increase. Active learning is a machine learning paradigm that seeks to address this problem by letting learning algorithms iteratively select a subset L^m of size m, from a larger unlabelled dataset U^n of size $n : m \leq n$, to be labelled by an oracle O for training. The active learning mantra can be stated as follows: *Train a machine*

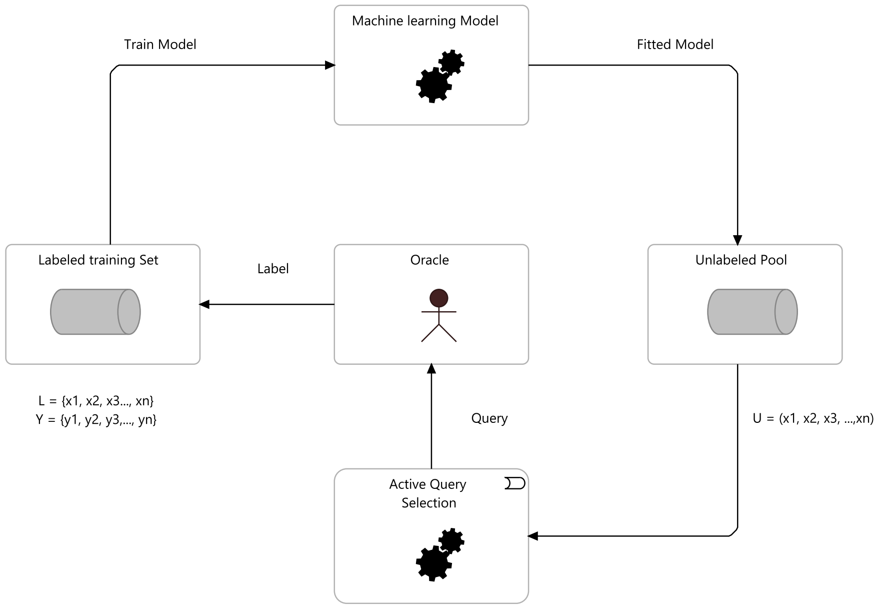

Fig. 2: The five main components to the standard Active Learning Framework. Each of these components may vary depending on the complexity of data to be learned, and available resources. The oracle could be either a human annotator of a software or simulation process.The machine learning Model can be a deep learning neural network, kernel method or standard tree based learning algorithm.

learning model on a significantly smaller labelled dataset, with little to no drop in test performance, all the while staying within a pre-determined labelling budget B.

Active Learning algorithms, while overlapping, can broadly be grouped into: pool based methods, density based methods, and data expansion methods. Pool based methods select samples for labeling from an unlabelled pool, based on either the uncertainty of the current trained model on samples U^{n-m} , the diversity of samples in the labelled set L^m used to train the current model, or a combination of both [29,31,12]. Pool-based methods are simple in their formation and implementation, but can be computationally expensive for large datasets of high dimensional data such as images. Since pool-based methods largely rely on metrics evaluated on the entire unlabeled dataset to select new candidates, they is not ideal for applications that require low latency. Density based methods seek to capture key characteristics of the underlying data distribution. This is done by selecting a core-set of samples for labeling that are sufficiently representative of the entire dataset, and lead to good generalization [46,39,38]. More recent literature blends pool and core-set methods to take advantage of each approach's benefits. These methods thus lead to efficient and robust models trained on core-sets containing diverse samples that maximize the margins between class margins [22,15]. Some methods in this approach use the hidden layer representations from training a self-supervised task on the image data, instead of the raw pixels. These include pre-training on image orientation: random $(90, 180, 270, 360)^\circ$ rotation classification, or self-supervised constrastive learning, where the target is an arbitrary patch of adjacent pixels in the image [7,14,50]. Data expansion methods seek to expand the training dataset, by generating reasonably realistic synthetic data samples for each target class at only a computational cost than that of obtaining human labeled samples, while enhancing the learning algorithm's performance on the real test dataset [?]. Since their introduction, generative adversarial networks (GANs) and their variations, [17,16,47], were the go-to method for generating synthetic data. However, the samples tend to be unrealistic, the training unstable, and lacking intrinsic evaluation metrics [4,32].

2.3 Label Noise

Label Noise refers to the scenario in which data labels are corrupted, with or without intention, so that we do not have 100% confidence in their correctness. Label noise is different from feature noise which is normally used to refer to adding Gaussian noise to feature values. Label noise impacts learning algorithms more adversely than feature noise does, and is harder to deal with [9,51,1,11]. Label noise is inherent in the data collection and processing life-cycle. Most real world datasets are subjected to a number of label noise sources based on how the data is collected, curated and stored. Label noise in practice broadly stems from: (1) incorrect crowd sourced labels where the annotators are non-experts such as is the case with Clickworker, and Amazon Mechanical Turk. (2) Incorrect expert annotations due to the complexity of the data, as is common in medical fields [18]. (3) Labelling errors introduced by automatic labelling by web crawling software and other AI labelling systems such as such as Scale AI. (4) Noise introduced by multiple experts or non experts labelling the same sample differently.

Learning noisy labels is especially hard due to the fact that cost functions are generally significantly less complex than feature extraction layers are. Label noise can be grouped, and is mostly treated based on what is known about the noise generating distribution [33]. Some datasets contain label noise from a known and quantifiable generative distribution, while in other cases, too little or nothing is known about the noise transition matrix to model. Label noise can be class independent or class dependent. Class independent label noise is the easiest to generate, for each sample, the label is swapped with with any other label of a different class, with a fixed probability $1/N$ where N is the number of classes [36]. Class-dependent label noise is normally a result of expert human annotation. It results from pairs of closely related or indistinguishable classes being occasionally swapped [20], e.g *True large sized cat occasionally labelled as a small dog, and visa versa.* Common methods, for training deep learning methods include first filtering out samples with a high probability of being noisy and iteratively training on a dataset with trusted labels until a thresh-hold is reached. The filtering process in most literature involves training two different neural networks with a custom loss, and monitoring samples on which they disagree on predictions. This methods works well since it has been shown the networks train on stronger signal first, which is the case in a dataset of mostly clean labels. representative methods in this approach, trained in a none-active learning manner include Decoupling [30], and Co-teaching [21]. The main implementation difference between the two approaches is in how the two network weights are updated, Decoupling updates each network's weights based on its on error term when the networks have a prediction disagreement. Co-teaching on the other hand, cross updates the weights with the error signal from the other network. Unlike Decoupling, Co-teaching addresses noisy labels explicitly by enabling the networks to peak into each other's training signal.

We have introduced deep learning, active learning, and learning on label noise. the next section is the main focus of this manuscript, we explore methods leveraging the versatility of deep neural networks, in the active learning framework where labelling budget is an important metric, and we are faced with a noisy label challenge.

3 Deep Active Learning Algorithms For Noisy Labels

In this section we focus on the main contribution of this manuscript: exploring literature on deep active learning algorithms used for vision tasks in the presence of label noise.

3.1 Noise Prior Induced Algorithms

Algorithms that seek to deal with label noise by modelling the underlying noise generating distribution.

3.2 Noise Robust Deep Active Learning Algorithms

algorithms that are independent of the noise distribution, and seek robust active training by architectural or loss function manipulation.

4 Evaluation Datasets and Metrics

In this section, datasets and their links are provided that pertain to active learning and deep learning on label noise. The benchmarks in evaluating algorithms on such datasets are also explored.

4.1 Datasets

4.2 Evaluation Metrics

5 Conclusion and Future Research Directions

The section summarizes imtactful literature on the topic of interest, lists future directions from key literature on the problem and opinions from this study

6 literature Tree

A visual depiction of the literature summary to deep active learning on noisy labels for computer vision.

References

1. Algan, G., Ulusoy, I.: Image classification with deep learning in the presence of noisy labels: A survey. ArXiv **abs/1912.05170** (2021)
2. Arik, S., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6679–6687 (2021). https://doi.org/10.1609/aaai.v35i8.16826
3. Baohua, S., Lin, Y., Wenhan, Z., Michael, L., Patrick, D., Charles, Y., Jason, D.: Supertml: Two-dimensional word embedding for the precognition on structured tabular data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2973–2981 (06 2019). https://doi.org/10.1109/CVPRW.2019.00360
4. Barnett, S.: Convergence problems with generative adversarial networks (gans). ArXiv **abs/1806.11382** (2018)
5. Breiman, L.: Random forests. Machine Learning Journal **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (08 2016). https://doi.org/10.1145/2939672.2939785
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. ICML'20, JMLR.org (2020)

8. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model. In: Arxiv (2022)
9. Chicheng, Z., Kamalika, C.: Active learning from weak and strong labelers. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper/2015/file/eba0dc302bcd9a273f8bbb72be3a687b-Paper.pdf
10. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units(elus). In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings (2016), http://arxiv.org/abs/1511.07289
11. Cordeiro, F., Carneiro, G.: A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In: The 33rd SIBGRAPI Conference on Graphics, Patterns and Images. pp. 9–16 (11 2020). https://doi.org/10.1109/SIBGRAPI51738.2020.00010
12. C.Shui, F.Zhou, C.Gagn'e, B.Wang: Deep active learning: Unified and principled method for query and training. In: International Conference on Artificial Intelligence and Statistics (2020)
13. Denil, M., Bazzani, L., Larochelle, H., de Freitas, N.: Learning where to attend with deep architectures for image tracking. In: Institute of Electrical and Electronics Engineers, Neural Computation. vol. 24, pp. 2151–2184 (2012)
14. Du, P., Zhao, S., Chen, H., Chai, S., Chen, H., Li, C.: Contrastive coding for active learning under class distribution mismatch. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8907–8916 (2021). https://doi.org/10.1109/ICCV48922.2021.00880
15. Geifman, Y., El-Yaniv, R.: Deep active learning over the long tail. ArXiv **abs/1711.00941** (2017)
16. Gonog, L., Zhou, Y.: A review: Generative adversarial networks. In: The 14th IEEE Conference on Industrial Electronics and Applications (ICIEA). pp. 505–510 (2019). https://doi.org/10.1109/ICIEA.2019.8833686
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
18. Górriz, M., Carlier, A., Faure, E., i Nieto, X.G.: Cost-effective active learning for melanoma segmentation. ArXiv **abs/1711.09168** (2017)
19. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
20. Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. In: Advances in Neural Information Processing Systems (05 2018)
21. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems (2018)

22. Har-Peled, S., Roth, D., Zimak, D.: Maximum margin coresets for active and noise tolerant learning. In: International Joint Conferences on Artificial Intelligence (2007)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
24. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: Neural Computation. vol. 9, pp. 1735–1780 (1997)
25. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., Zhai, X.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021 (2021), https://openreview.net/forum?id=YicbFdNTTy
26. Konyushkova, K., Sznitman, R., Fua, P.: Learning active learning from data. In: NIPS (2017)
27. LeCun, Y., Bengio, Y.: Convolutional Networks for Images, Speech, and Time Series, p. 255–258. MIT Press, Cambridge, MA, USA (1998)
28. Lecun, Y., Bottou, L., Bengio., Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (09 1998). https://doi.org/10.1109/5.726791
29. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: SIGIR '94. pp. 3–12. Springer London, London (1994)
30. Malach, E., Shalev-Shwartz, S.: Decoupling "when to update" from "how to update". In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 961–971. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
31. McCallum, A., Nigam, K.: Employing em and pool-based active learning for text classification. In: International Conference of Machine Learning (1998)
32. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: The 35th International Conference on Machine Learning (2018)
33. Nagarajan, N., Inderjit, D., Pradeep, R., Ambuj, T.: Learning with noisy labels. In: Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf
34. Nair, V., Hinton, G.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning (2010)
35. Novák, A., Laki, L., Novák, B.: CBOW-tag: a modified CBOW algorithm for generating embedding models from annotated corpora. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4798–4801. European Language Resources Association (09 2020), https://aclanthology.org/2020.lrec-1.590
36. Patrinin, G., Rozza, A., Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2233–2241 (07 2017). https://doi.org/10.1109/CVPR.2017.240
37. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (10 2014). https://doi.org/10.3115/v1/D14-1162, http://www.aclweb.org/anthology/D14-1162

38. Phillips, J.: Coresets and sketches. CoRR **abs/1601.00617** (2016), http://arxiv.org/abs/1601.00617
39. Phillips, J., Tai, W.: Near-optimal coresets of kernel density estimates. In: 34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary. LIPIcs, vol. 99, pp. 66:1–66:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2018). https://doi.org/10.4230/LIPIcs.SoCG.2018.66, https://doi.org/10.4230/LIPIcs.SoCG.2018.66
40. Popov, S., Morozov, S., Babenko, A.: Neural oblivious decision ensembles for deep learning on tabular data. ArXiv **abs/1909.06312** (2020)
41. Prokhorenkova, L., Gleb, G., Vorobev, A., Dorogush, A., Gulin, A.: Catboost: unbiased boosting with categorical features. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Proceedings of the 32nd International Conference on Neural Information Processing Systems. vol. 31, p. 6639–6649. Curran Associates, Inc. (2018)
42. Roman, L., Valeriia, C., Avi, S., Arpit, B., Bruss, C., Tom, G., Andrew, W., Micah, G.: Transfer learning with deep tabular models (06 2022). https://doi.org/10.48550/arXiv.2206.15306
43. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. In: Psychological Review. vol. 65, pp. 386–408 (1958)
44. Schäfl, B., Gruber, L., Bitto-Nemling, A., Hochreiter, S.: Hopular: Modern hopfield networks for tabular data. ArXiv **abs/2206.00664** (2022)
45. See, A., Liu, P., Manning, C.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 1073–1083 (2017)
46. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. International Conference on Learning Representations (Poster) (2018)
47. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5971–5980. IEEE Computer Society, Los Alamitos, CA, USA (03 2019). https://doi.org/10.1109/ICCV.2019.00607, https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00607
48. Srinivas1, A., Lin, T., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: 2021 Conference on Computer Vision and Pattern Recognition (2021)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://arxiv.org/pdf/1706.03762.pdf
50. Wang, C., Singla, A., Chen, Y.: Teaching an active learner with contrastive examples. In: Advances in Neural Information Processing Systems (2021)
51. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. 10th International Conference on learning Representations (2022)
52. Wortsman, M., Ilharco, G., Gadre, S., Roelofs, R., Gontijo-Lopes, R., Morcos, A., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International Conference on Machine Learning. pp. 23965–23998. PMLR (2022)

53. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research **abs/2205.01917** (2022)
54. Zihang, D., Hanxiao, L., Quoc, L., Mingxing, T.: Coatnet: Marrying convolution and attention for all data sizes. In: 35th Conference on Neural Information Processing Systems (06 2021)