36th International Conference on Machine Learning (ICML), 2019

Bayesian Generative Active Deep Learning

Toan Tran[1] Thanh-Toan Do[2] Ian Reid[1] Gustavo Carneiro[1]

[1]University of Adelaide, Australia [2]University of Liverpool

Wed Jun 12th 04:30 – 04:35 PM @ Room 201

Deep learning (DL)

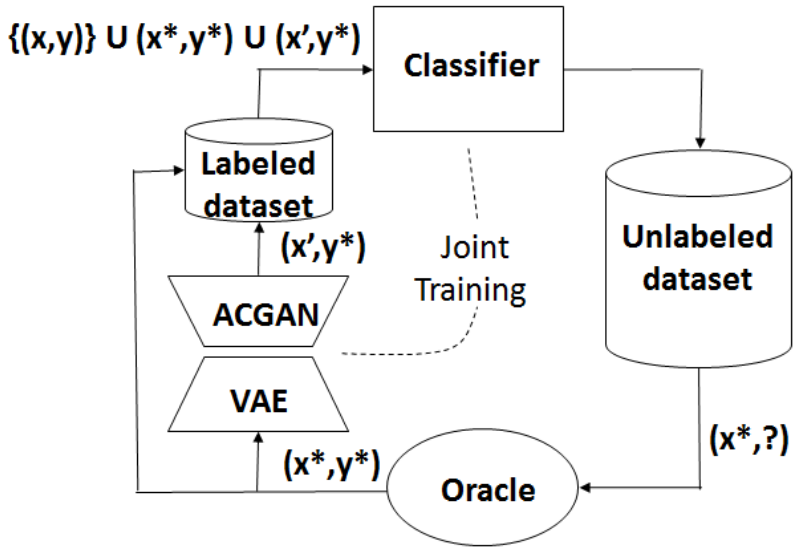
- The dominant machine learning methodology [Huang et al., 2017, Rajkomar et al., 2018].
- **Issue:** significant human effort for the labeling; considerable computational resources for large-scale training process [Sun et al., 2017].

How to address these training issues? Two popular approaches:

1. (Pool-based) active learning (AL): Challenging to be applied in DL: AL may overfit the (small) informative training sets
2. Data augmentation (DA): the generation of new samples is done without regarding their informativeness $\Longrightarrow$ the training process takes longer than necessary and relatively ineffective

Main goals of this paper

- Propose a novel Bayesian generative active deep learning method
- Targets the augmentation of the labeled data set with informative generated samples
- **Key technical contribution**: theoretically and empirically show the informativeness of this generated sample.



c) Bayesian Generative Active Deep Learning

Figure 1: Our proposed method depicted by VAE-ACGAN model

Generative adversarial active learning (GAAL) [Zhu and Bento, 2017]

- Relies on an optimization problem to generate new informative samples
- Can generate rich representative training data with the assumptions:
  - GAN model has been pre-trained, and
  - The optimization during generation is solved efficiently

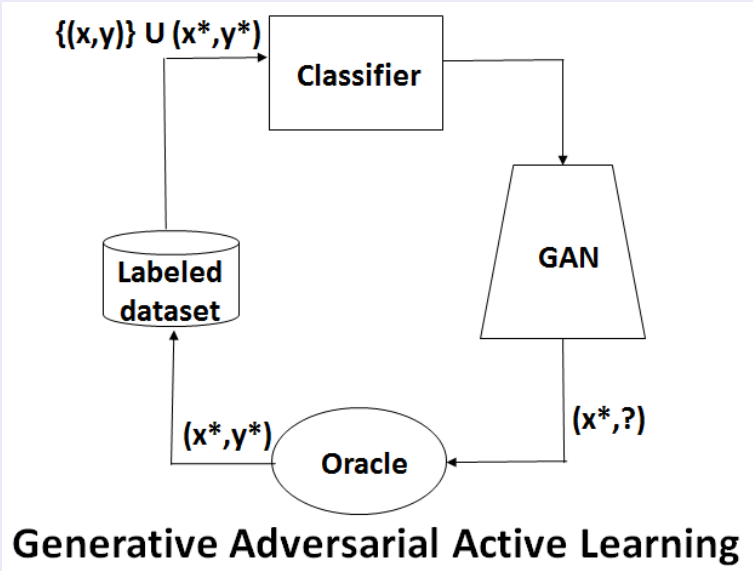


Figure 2: Generative adversarial active learning (GAAL) [Zhu and Bento, 2017]

Comparison between our proposed method and GAAL [Zhu and Bento, 2017]

	GAAL [Zhu and Bento, 2017]	Ours
acquisition function	simple (binary classifier)	more effective (deep models)
training of the generator (G) and classifier (C)	2-stage GAN model is pre-trained	G and C are jointly trained Allowing them to "co-evolve"
classification results	not competitive enough	

Main technical contribution:

Combining BALD and BDA for generating new labeled samples that are informative for the training process.

- Initial labeled data: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is the data sample labeled with $\mathbf{y}_i \in \mathcal{C} = \{1, 2, \ldots, C\}$ ($C = \#$ classes).
- Bayesian active learning by disagreement (BALD) scheme [Gal et al., 2017, Houlsby et al., 2011]:
- The most informative sample $\mathbf{x}^*$ is selected from the (unlabeled) pool data $\mathcal{D}_{\text{pool}}$ by [Houlsby et al., 2011]:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{D}_{\text{pool}}} a(\mathbf{x}, \mathcal{M}), \tag{1}$$

where the acquisition function $a(\mathbf{x}, \mathcal{M})$ is estimated by the Monte Carlo (MC) dropout method [Gal et al., 2017]

$$a(\mathbf{x}, \mathcal{M}) \approx -\sum_c \left(\frac{1}{T}\sum_t \hat{p}_c^t\right) \log\left(\frac{1}{T}\sum_t \hat{p}_c^t\right) + \frac{1}{T}\sum_{c,t} \hat{p}_c^t \log \hat{p}_c^t, \tag{2}$$

where T is the number of dropout iterations,
$\hat{\mathbf{p}}^t = [\hat{p}_1^t, \ldots, \hat{p}_C^t] = \text{softmax}(f(\mathbf{x}; \theta^t))$, with f is the network function parameterized by $\theta^t \sim p(\theta|\mathcal{D})$ at the t-th iteration.

- The generated sample $\mathbf{x}'$:

$$\mathbf{x}' = g(e(\mathbf{x}^*)), \tag{3}$$

where a variational autoencoder (VAE) [Kingma and Welling, 2013] contains an encoder $e(.)$ and a decoder $g(.)$

- VAE training: minimizing the "reconstruction loss", where if # training iterations is sufficiently large, we have:

$$\|\mathbf{x}' - \mathbf{x}^*\| < \varepsilon, \tag{4}$$

$\varepsilon > 0$ (arbitrarily small) – see Fig. 3.

- $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{x}^*, \mathbf{y}^*), (\mathbf{x}', \mathbf{y}^*)\}$, which are used for the next training iteration.

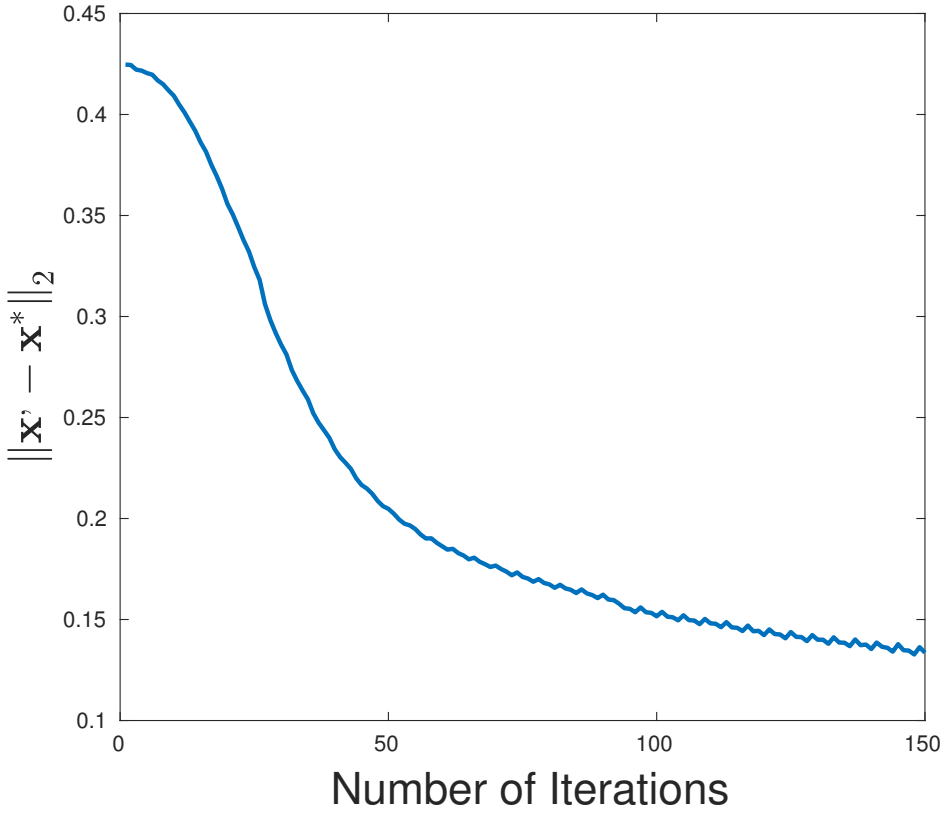


Figure 3: Reduction of $\|\mathbf{x}' - \mathbf{x}^*\|$ as the training progresses of the VAE model.

Key Question

How about the "information content" of the generated sample $\mathbf{x}'$ measured by $a(\mathbf{x}', \mathcal{M})$?

Proposition 2.1

Assuming that there exists the gradient of the acquisition function $a(\mathbf{x}, \mathcal{M})$ with respect to the variable $\mathbf{x}$, namely $\nabla_{\mathbf{x}} a(\mathbf{x}, \mathcal{M})$, and that $\mathbf{x}^*$ is an interior point of $\mathcal{D}_{pool}$, then $a(\mathbf{x}', \mathcal{M}) \approx a(\mathbf{x}^*, \mathcal{M})$ (i.e., the absolute difference between these values are within a certain range). Consequently, the sample $\mathbf{x}'$ generated from the most informative sample $\mathbf{x}^*$ by (3) is also informative.

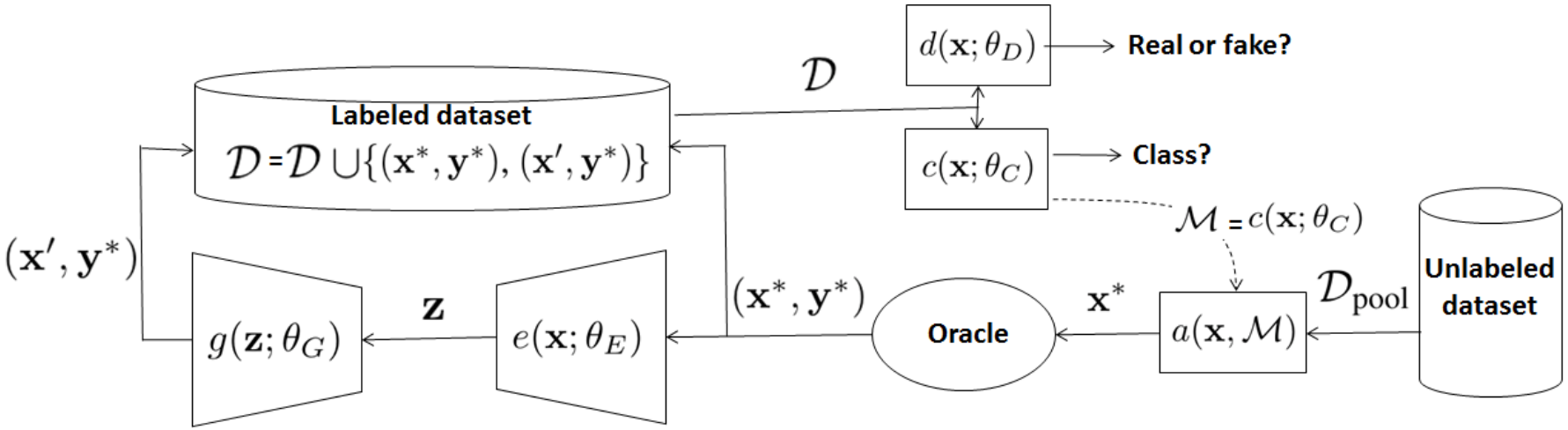
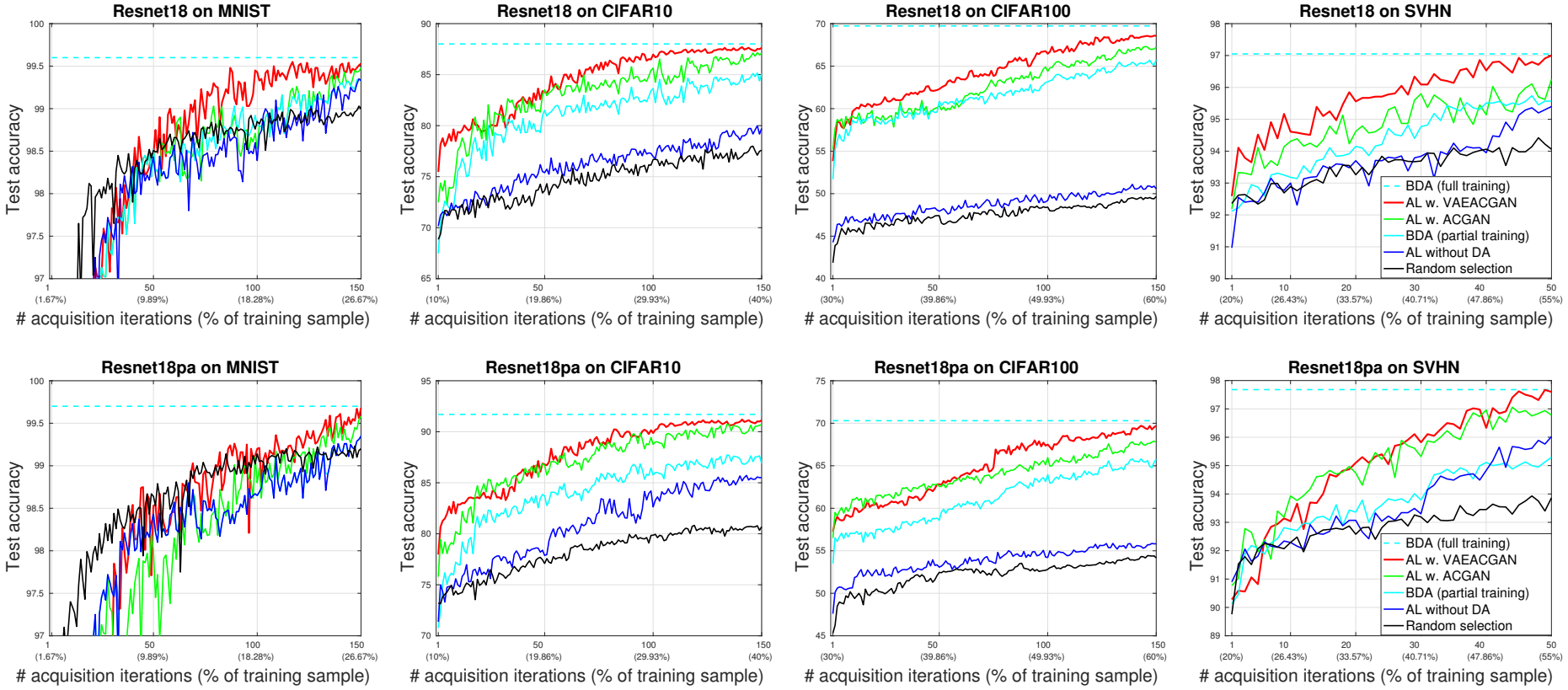


Figure 4: Network architecture of our proposed model.

- Classification performance measured by the top-1 accuracy as a function of the number of acquisition iterations and the percentage of training samples.
- Our proposed algorithm, active learning using "information-preserving" data augmentation (AL w. VAEACGAN) is compared with:
  - Active learning using BDA (AL w. ACGAN)
  - BALD [Gal et al., 2017] without using data augmentation (AL without DA),
  - BDA [Tran et al., 2017] without active learning (BDA) (using full and partial training sets)
  - Random selection
- Benchmark data sets: MNIST [LeCun et al., 1998], CIFAR-10, CIFAR-100 [Krizhevsky et al., 2012], and SVHN [Netzer et al., 2011].
- Baseline classifiers: ResNet18 [He et al., 2016a] and ResNet18pa [He et al., 2016b]



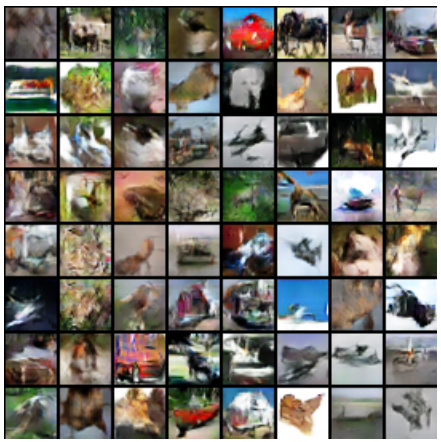
(a) MNIST (b) CIFAR-10 (c) CIFAR-100 (d) SVHN

Figure 5: Training and classification performance of the proposed Bayesian generative active learning (AL w. VAEACGAN) compared to other methods. This performance is measured as a function of the number of acquisition iterations and respective percentage of samples from the original training set used for modeling.

Table I: Mean ± standard deviation of the classification accuracy on MNIST, CIFAR-10, and CIFAR-100 after 150 iterations over 3 runs

	AL w. VAEACGAN	AL w. ACGAN	AL w. PMDA	AL without DA	BDA (partial training)	Random selection
			MNIST			
ResNet18	**99.53 ± 0.05**	99.45 ± 0.02	99.37 ± 0.15	99.33 ± 0.10	99.33 ± 0.04	99.00 ± 0.13
ResNet18pa	**99.68 ± 0.08**	99.57 ± 0.07	99.49 ± 0.09	99.35 ± 0.11	99.35 ± 0.07	99.20 ± 0.12
			CIFAR-10			
ResNet18	**87.63 ± 0.11**	86.80 ± 0.45	82.17 ± 0.35	79.72 ± 0.19	85.08 ± 0.31	77.29 ± 0.23
ResNet18pa	**91.13 ± 0.10**	90.70 ± 0.24	87.70 ± 0.39	85.51 ± 0.21	86.90 ± 0.27	80.69 ± 0.19
			CIFAR-100			
ResNet18	**68.05 ± 0.17**	66.50 ± 0.63	55.24 ± 0.57	50.57 ± 0.20	65.76 ± 0.40	49.67 ± 0.52
ResNet18pa	**69.69 ± 0.13**	67.79 ± 0.76	59.67 ± 0.60	55.82 ± 0.31	65.79 ± 0.51	54.77 ± 0.29

(a) MNIST



(b) CIFAR-10

(c) CIFAR-100

(d) SVHN

Figure 6: Images generated by our proposed (AL w. VAEACGAN) approach for each data set.

- Consistently shows that our proposed approach outperforms other methods.
- The classification results by AL w. VAEACGAN are statistically significant w.r.t. BDA (partial training) on all those data sets, and w.r.t. AL w. ACGAN on CIFAR-$\{10, 100\}$ for both models (i.e., $p \leq .05$, two-sample t-test).
- Proposed a novel Bayesian generative active deep learning approach that consistently shows to be more effective than data augmentation and active learning in several classification problems.
- Our proposed approach is (active learning) model-agnostic, it therefore can be combined with several currently sate-of-the-art active learning methods [Ducoffe and Precioso, 2018, Gissin and Shalev-Shwartz, 2018, Sener and Savarese, 2018].

Future work

- To investigate how to generate samples directly using complex acquisition functions
- To work on the efficiency of our proposed method (its empirical computational cost is slightly higher than BDA [Tran et al., 2017] and BALD [Gal et al., 2017, Houlsby et al., 2011]).
- Imbalanced data sets

Poster presentation:

Wed Jun 12th 06:30–09:00 PM
@ Pacific Ballroom #263

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841, 2018.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In International Conference on Machine Learning, pages 1183–1192, 2017.

Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630–645. Springer, 2016b.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, volume 1, page 3, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In NIPS workshop on deep learning and unsupervised feature learning, volume 2011, page 5, 2011.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Peter J Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, et al. Scalable and accurate deep learning for electronic health records. arXiv preprint arXiv:1801.07860, 2018.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In International Conference on Learning Representations, 2018. URL `https://openreview.net/forum?id=H1aIuk-RW`.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 843–852. IEEE, 2017.

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In Advances in Neural Information Processing Systems, pages 2797–2806, 2017.

Jia-Jie Zhu and Jose Bento. Generative adversarial active learning. arXiv preprint arXiv:1702.07956, 2017.