Continual Active Adaptation to Evolving Distributional Shifts

Amrutha Machireddy, Ranganath Krishnan, Nilesh Ahuja, Omesh Tickoo
Intel Labs
{amrutha.machireddy, ranganath.krishnan, nilesh.ahuja, omesh.tickoo}@intel.com

Abstract

Building neural network models that are adaptable to evolving data distributions without suffering catastrophic forgetting is important for real-world deployment in many applications. In real-world setting, the observed data distribution changes over time due to non-stationary environment. In this paper, we consider the problem of evolving covariate shift and propose source-free active adaptation method to fine-tune the neural networks to continually evolving data without catastrophic forgetting. We evaluate the model performance with respect to adaptation as well as forgetting under sequential evolution of data based on fifteen different common corruptions and perturbations from CIFAR10-C related to shift in lighting, weather, noise etc. We demonstrate the proposed method improves model accuracy to the continually evolving data by 21.3% *on an average over the different covariate shifts without catastrophic forgetting.*

1. Introduction

In the supervised learning framework, the goal is to learn the underlying input-output mapping based on the available training samples and to predict the output for unseen inputs based on the learnt mapping. The training and testing data are assumed to follow the same distribution for the models to work well in the supervised learning paradigm [1]. However, this assumption does not hold true always as data in real world is continually evolving and a model trained on a certain distribution might not work well on the dynamically changing input distributions [2]. We are interested in such scenarios where the data distribution evolves with time. There are many research areas that fall under such distributional shift including out-of-distribution detection [3, 4], novel class detection [5], anomaly detection [6] and adversarial detection [7, 8]. However, in this work, the focus is on covariate shift [9] present in data where the training and testing input data follow different distributions but the output labels remain the same. We consider the number of classes to be fixed based on the initial data.

The model accuracy of existing state-of-the-art methods degrades under distributional shift [10], however, uncertainty-aware models can detect these distributional shifts using uncertainty estimation techniques [11]. In this paper, we use uncertainty estimates of the model on the new data to identify the most informative samples which can be used for fine-tuning the model to learn the shifted distribution.

While adapting the model to learn the new data distribution, the model might forget the previously learnt distribution leading to catastrophic forgetting [12], which is a well-known problem in continuous learning. Catastrophic forgetting has primarily been studied in class-incremental [13] or task-incremental [14] setup, whereas we are interested under continually evolving covariate shift. One of the well-known approach to prevent catastrophic forgetting is the replay method [15], in which a subset of the past data is stored. However, retaining past data information is not feasible in certain applications due to privacy concerns and also requires additional memory for storage. To overcome this, we propose a source-free adaptation approach with the use of batch normalization [16, 17] to adapt the model to the continually evolving data distribution.

The goal of this work is to adapt the model to covariate shift in data while preventing catastrophic forgetting on the previously learnt distributions.

The main contributions in this paper include:

- We propose a source-free batch-normalization adaptation approach to the continually evolving data without the need of retaining the past sample information (source data). We identify a subset of informative samples from the observed data through uncertainty-based selective sampling for active labelling, which is used to fine-tune the model to adapt to the shifted distribution.

- We evaluate model performance under sequential evolution of data w.r.t. adaptability and catastrophic forgetting. The proposed approach is capable of continually adapting to new data distribution while not forgetting the previously learnt distributions.

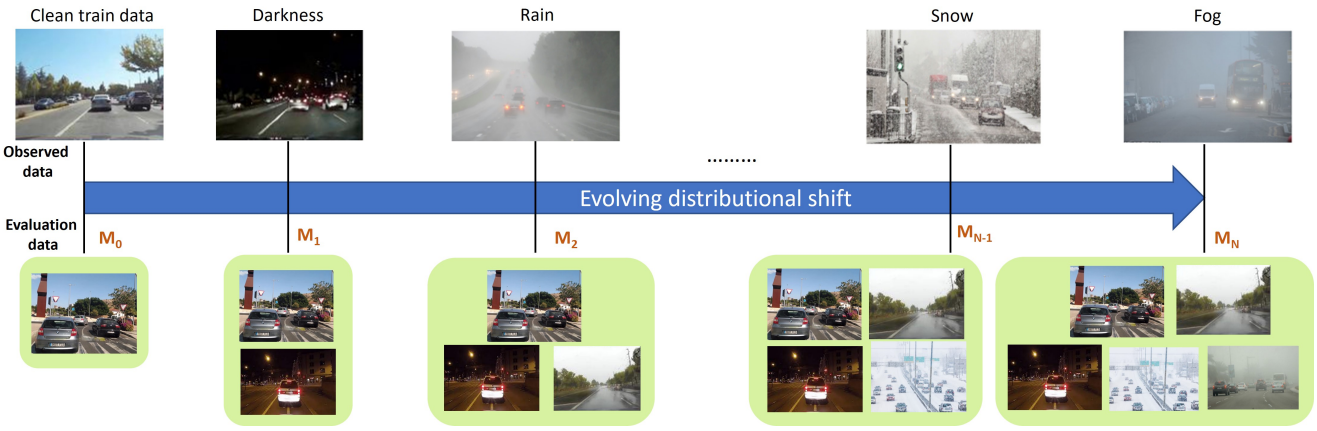

Figure 1. Illustration of continually evolving distributional shifts in real-world: The initial model trained on the clean data can observe continually evolving distributional shifts and must adapt itself while not forgetting the past learning. At each instance, the model must be able to perform well on the newly learnt distribution as well as all the data distributions it has learnt in the past.

2. Related work

There have been several works in the direction of reducing catastrophic forgetting in the field of continual learning [18]. In the regularization based approach, the past data is not required but a drift from the already learnt tasks is penalized [19,20]. In the case of memory-based approaches, a small subset of data is stored which is used to constrain the optimization process such that the loss on the past data does not increase [21]. In experience replay, a subset of samples is stored and used for retraining the model [15]. In the iCARL [22] approach, a distillation loss is used to remember the past samples. The GEM [21] and A-GEM [23] algorithms update the gradient such that the loss on the samples in the buffer does not increase. All these methods consider the class-incremental setup where at each instance, a new class is introduced to the model. In this work, we are interested in the setting where the number of classes are fixed, but the data distribution of the classes evolve. An online continual learning approach for adapting to distributional shifts was proposed based on experience replay method to remember the past distributions [24]. The replay methods work well in practice to prevent catastrophic forgetting, however, they require a small subset of the past data to be stored for fine-tuning the model. However, in applications where a subset of the past data information cannot be stored due to data privacy reasons, such methods cannot be used.

The batch normalization [16] technique was proposed to improve the training convergence by reducing the internal covariate shift and in a way help in eliminating the need for regularization in the model. Various methods have been proposed to use the batch normalization technique to adapt the model to new data distributions. In literature, predictive time batch normalization has been looked from the domain adaptation [25] and robustness perspective [26]. In test-time adaptation [27, 28], the batch norm parameters are updated based on the target data distribution via entropy minimization. The test-time adaptation has an additional compute overhead of training during inference that will raise challenges in real-time-critical applications. However, these methods focus only on model adaptation to target distribution as they do not consider the continually evolving distributional shift setup and catastrophic forgetting problem. The work in this paper explores batch normalization adaptation without the need for past data samples and characterizes the model performance under continually evolving distributional shift.

3. Problem setup

In supervised learning, when a model is trained with data from a given distribution, it performs well as long as the test data is from the same distribution. However, when there is a distributional shift in data, the model performance degrades [10, 11]. Let $P_{\text{train}}(x)$ and $P_{\text{test}}(x)$ be the probability distribution of the training and testing input data respectively. $P_{\text{train}}(y|x)$ and $P_{\text{test}}(y|x)$ correspond to the conditional output distribution of the labels for the training and testing data respectively. In this work, the focus is on covariate shift present in data where the training and testing input data follow different distributions $(P_{\text{train}}(x) \neq P_{\text{test}}(x))$, but the output distributions remain same $(P_{\text{train}}(y|x) = P_{\text{test}}(y|x))$.

The given model is trained with samples from the initial available data distribution. In the continually evolving distribution shift setup, we assume a sequential order in presentation of the shifted data to the model such that samples corresponding to only one type of shift are available at each instance. During inference, the model observes samples from the shifted distribution. However, the labels as-

sociated with the shifted data are not known and have to be labelled by an oracle upon detection of distribution shift. The cost associated with labelling the samples by an oracle is expensive. Therefore, it is important to identify a small subset of informative samples that are used to adapt the model. The goal in this paper is to build a model that can identify a subset of informative samples to be labelled from the shifted data and continually adapt itself while not forgetting the past learning as shown in Figure 1.

4. Proposed approach

To start with, the given model is trained using data from the initial distribution. During inference, only the shifted samples are observed and their corresponding labels are unknown. In active learning, the samples to be labelled are prioritised such that the accuracy of the model is improved using a smaller subset of data [29]. As the cost associated with human annotation of the samples is expensive, a subset of informative samples need to be chosen based on a query strategy. As discussed earlier, uncertainty estimation techniques can be used to detect distributional shifts in data [11]. We use the predictive entropy of the model to quantify the uncertainty [30, 31] which is defined as

$$\mathbb{H}(y|x, D) := -\sum_{k=1}^{K} p(y = c_k|x, w) \log p(y = c_k|x, w), \tag{1}$$

where D corresponds to the data the model has been trained on, K corresponds to the total classes, and $p(y = c_k|x, w)$ corresponds to the output from the classifier with weights w. A high entropy value implies the model is not certain about the class the data belongs. Therefore, to identify the most informative samples that can be chosen for active labelling, the samples are ordered based on the decreasing order of entropy. The subset of samples are chosen based on the entropy are labelled and the model can be adapted to the shifted data. However, by repeatedly adapting the model to the new data, the model may forget the representation of the past and perform poorly on the initial data distributions.

We build upon the insights from [27] that shows the effectiveness of updating the batch normalization statistics for test-time adaptation. We propose to update the batch normalization statistics in the continual adaptation setting while optimizing the cross-entropy objective on the samples that were detected to be distributionally shifted through predictive uncertainty estimation. During fine-tuning, only the batch normalization parameters of the model are updated and the weights of all the layers are frozen and not updated after the initial training. We call this approach as source-free batch norm (BN) adaptation as illustrated in Figure 2. The model is trained with the initial training data $\{X_0, Y_0\}$. As the data evolves, the model observes samples $\{X_i\}$ from a different distribution. The samples from the new distribution identified based on uncertainty estimation, the samples with high predictive entropy $\{\overline{X}_i\}$ are selected to obtain the corresponding labels $\{\overline{Y}_i\}$ at every sequential step. The labelled informative samples $\{\overline{X}_i, \overline{Y}_i\}$ are used to update the batch norm parameters of the model while optimizing the cross-entropy loss without the access to source data $\{X_0, Y_0\}$. This process continues sequentially for the evolving data distributions.

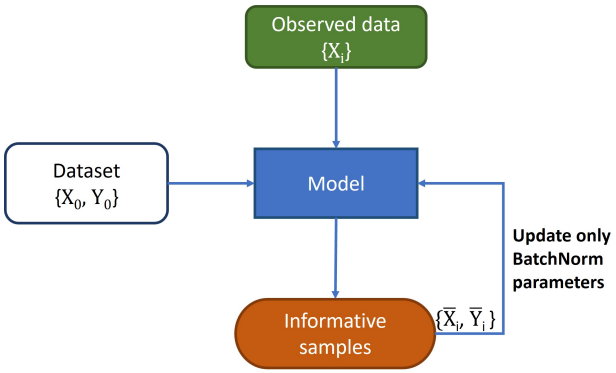

Figure 2. Source-free BN adaptation: The model is trained with the initial training data $\{X_0, Y_0\}$. As the data evolves, the model observes samples $\{X_i\}$ from a different distribution. A subset of the observed samples $\{\overline{X}_i\}$ are chosen based on the predictive entropy and the corresponding labels $\{\overline{Y}_i\}$ are obtained. The labelled data $\{\overline{X}_i, \overline{Y}_i\}$ are used to update the batch norm parameters of the model to adapt to the new data distribution without the access to source data $\{X_0, Y_0\}$. This process continues for the different shifts observed by the model.

5. Experiments

5.1. Adaptation to continually evolving data

We perform experiments on the CIFAR10 dataset [32]. To simulate the evolution of data, we consider 15 different common corruptions and perturbations on the data introduced in CIFAR10-C [33]. We use the corruption severity level 5 for experimentation. We choose 45000 random samples from the CIFAR10 dataset to train the initial model. The remaining 5000 samples in the CIFAR10 training data are held out to generate corrupted samples corresponding to the 15 corruptions of severity level 5. These held out samples are used for fine-tuning the model. For testing, the CIFAR10 and CIFAR10-C test data consisting of 10000 samples for each corruption are considered to evaluate the model performance.

We perform experiments with the ResNet-18 model architecture [34] for all the methods. The baseline model is trained with the stochastic gradient descent optimizer for 100 epochs with a learning rate 0.1, momentum 0.9 and the weight decay 0.0005. To adapt the model to the shifted data, the model is fine-tuned with the Adam optimizer for

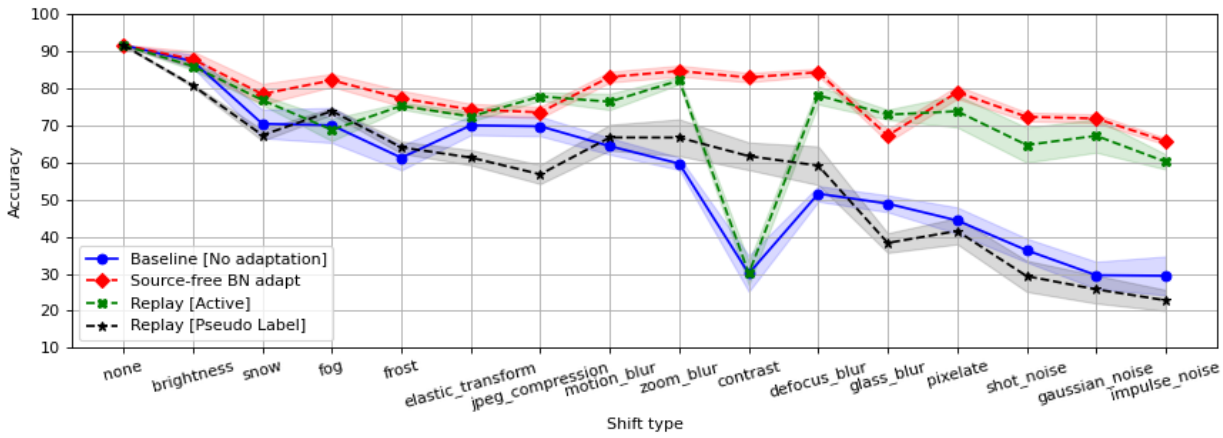

Figure 3. Covariate shift adaptation: Comparison of accuracy of the methods after adapting to each corruption in the continually evolving setup. The X-axis denotes the order in which the corruptions are introduced to the model. The source-free BN adaptation improved the accuracy by 21.3% on average from the baseline.

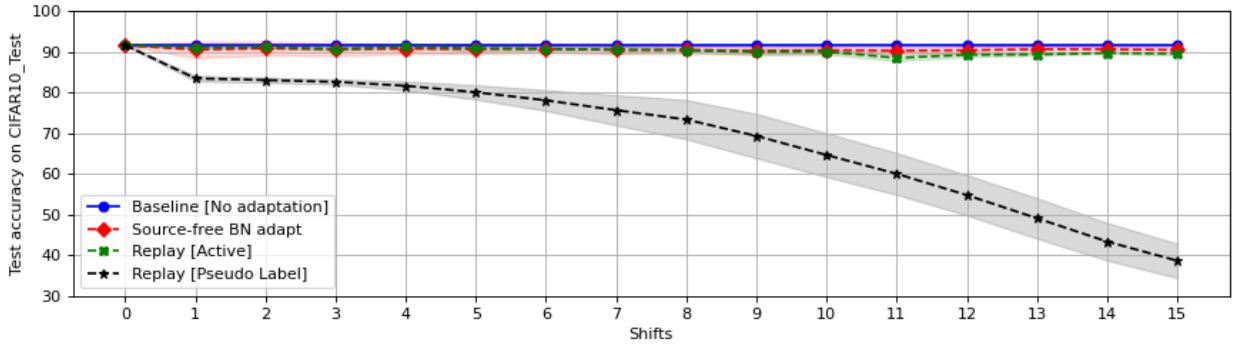

Figure 4. Catastrophic forgetting on clean CIFAR10: The accuracy of all the methods on the clean CIFAR10 test data is computed after the model is adapted for each corruption. The X-axis corresponds to the number of corruptions the model has been adapted to until now. It is seen that the source-free BN adaptation performs well even in the absence of the past sample information.

10 epochs with learning rate 0.0001. The same hyperparameters are used for all the methods for fair comparison.

The baseline model is trained using the 45000 clean CIFAR10 samples. During inference, the 5000 corrupted samples generated for each corruption are introduced. The shifted data is introduced one after the other to mimic the continual learning setup. As the model is trained only on the clean data, the performance of the model degrades with the introduction of corrupted samples. Also, the samples that are introduced during inference do not have the associated label information. We choose to actively label the data using an oracle which has an added cost associated with it. To reduce the cost of labelling all the samples, we selectively choose the most informative samples that can be given to the oracle for labelling. The predictive entropy of the model during inference is used to identify a subset of 2000 informative samples which are then labeled by the oracle. The labeled subset of data is used to fine-tune the model for each corruption. The accuracy of the model on the CIFAR10-C test data comprising of 10000 samples is computed after fine-tuning the model based on the actively labelled samples corresponding to the chosen corruption type to evaluate the extent to which the model adapts to the shifted data.

Figure 3 shows the performance of the algorithms as the data distribution continuously evolves with the different shifts. We present the results from five independent trials. Baseline corresponds to the model with no adaptation and its model accuracy degrades significantly under evolving distribution shift. The replay method is an existing approach in the class incremental [15] continuous learning setup to prevent catastrophic forgetting which we experiment in the covariate shift setting. A buffer of 5000 samples consisting of a subset of the past samples is used. As the shifts in the data increase, the number of samples cor-

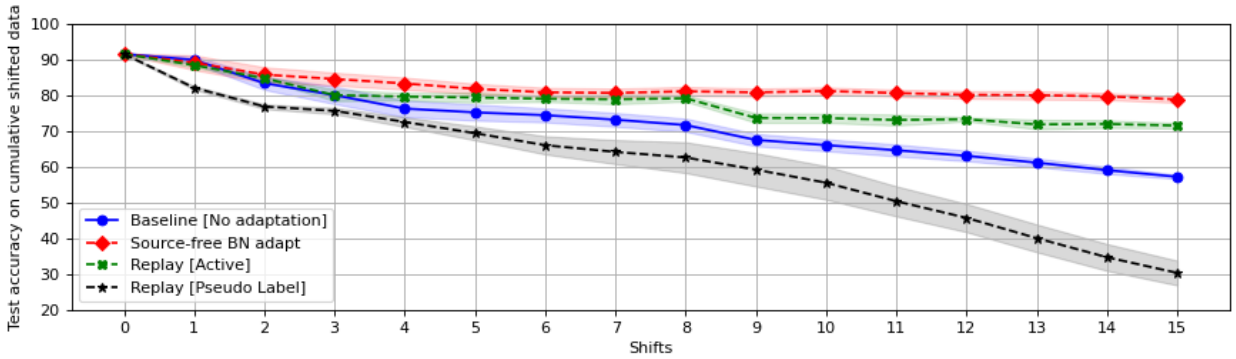

Figure 5. Performance on cumulative test data (combined evaluation of adaptation and forgetting): When a model is adapted, it should perform well on the current distribution shift as well as all the distribution shifts it has observed until now. We compute the accuracy of each model on the cumulative test data consisting of the clean CIFAR10 test data and the test data corresponding to the shifts it has observed until now. It can be seen that the source-free BN adaptation is able to retain the past information.

responding to each shift in the buffer decrease. All the parameters in the model are fine-tuned based on the samples in the buffer as well as the subset of samples labeled using the oracle. We call this approach as active replay. Although this method is able to adapt to the shifted data, it requires the past samples during adaptation which might not be available in many applications. We also compare the performance of the algorithm by pseudo-labeling the samples based on the model prediction. Since the model predictions are likely to be incorrect as it does not have knowledge of the corrupted data distribution, we see that the model performance degrades badly. Our proposed method of source-free batch norm adaptation with active labelling improved the model accuracy by 21.3% on average over the different shifts in the data compared to the baseline, performing better than replay methods. The proposed method is computationally fast and requires less memory as only the batch normalization parameters are updated and does not require any past sample information.

5.2. Catastrophic forgetting evaluation

In the previous section, we have shown the performance of the algorithm to adapt to shifted data. However, over the process of adapting to continually evolving data, the model performance should not degrade on the initial test data which is the clean CIFAR10 test data in this case. For this purpose, we evaluate the performance of each of the above mentioned methods to overcome catastrophic forgetting. Figure 4 shows the accuracy on the CIFAR10 test data by the models that are obtained after being adapted to each shifted data. The X-axis corresponds to the number of distributional shifts the model has been adapted to. The model performance over the clean test data will degrade if the model suffers from catastrophic forgetting. Here baseline corresponds to the initial model trained on the clean CIFAR10 data without any adaptation to new data. In the active replay, the presence of the the initial data distribution in the form of past samples helps alleviate catastrophic forgetting while adapting the model to new distributions. The source-free BN method adapts without catastrophic forgetting with an average accuracy drop of only 1% from the baseline model, even though it does not require the previous source data for learning. The replay method with pseudo labeling suffers severe catastrophic forgetting due to the accumulation of possibly incorrect labels of the shifted data.

5.3. Evaluation of adaptation retaining capacity

We next perform combined evaluation of model performance with respect to adaptation and forgetting towards continually evolving data. Figure 5 shows the accuracy obtained on the cumulative test data comprising the test data of all the shifts it has observed until now. Each number on the X-axis corresponds to the number of corruption types the test data includes along with the clean test data. For example, 2 corresponds to the cumulative test data of the first and second corruption type observed by the model along with the test data corresponding to the clean CIFAR10 data. The active replay method performed well on the clean CIFAR10 data as seen in Figure 4, however, the model performance degrades when the accuracy is computed over the cumulative shifted test data. Replay with pseudo-labelling follows similar trend of catastrophic forgetting on the cumulative dataset also. The source-free BN method improved the accuracy by 10% on average as compared to the baseline which shows that the model is able to retain the shifted data information even in the absence of the past data during adaptation.

6. Conclusion

In this work, we propose the source-free batch-normalization technique for detecting and adapting the model to continually evolving distribution shift. We show the effectiveness of the proposed approach by evaluating the model performance to adaptability to new data distributions and remembering the past data distributions. For each new evolution in the data distribution, a subset of informative samples were selected based on the predictive entropy of the model which reduces the compute burden and cost of labelling the shifted data for fine-tuning. The results show that the proposed approach is capable of adapting to new data distributions while not forgetting the past information without having access to the source data.

References

[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 1

[2] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012. 1

[3] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1

[4] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1

[5] Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani M Thuraisingham. Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):859–874, 2010. 1

[6] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 1

[7] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *Proceedings of the Thirty-fourth Conference on Uncertainty in Artificial Intelligence*, 2018. 1

[8] Nilesh A Ahuja, Ibrahima Ndiour, Trushant Kalyanpur, and Omesh Tickoo. Probabilistic modeling of deep features for out-of-distribution and adversarial detection. *arXiv preprint arXiv:1909.11786*, 2019. 1

[9] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012. 1

[10] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 1, 2

[11] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020. 1, 2, 3

[12] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[13] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 1

[14] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 1

[15] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 1, 2, 4

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1, 2

[17] Muhammad Awais, Md Tauhid Bin Iqbal, and Sung-Ho Bae. Revisiting internal covariate shift for batch normalization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):5082–5092, 2020. 1

[18] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[21] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 2

[22] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2

[23] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 2

[24] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8281–8290, 2021. 2

[25] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 2

[26] Wonju Lee, Seok-Yong Byun, Jooeun Kim, Minje Park, and Kirill Chechil. Unsupervised model drift estimation with batch normalization statistics for dataset shift detection and model selection. *arXiv preprint arXiv:2107.00191*, 2021. 2

[27] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2020. 2, 3

[28] Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[29] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 3

[30] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 3

[31] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021. 3

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[33] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 3

[34] 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, 2016. 3