Advanced Data Mining Reading Course

Final Report

By
Vangjush Komini

vangjush@kth.se

June 8, 2021

-

Contents

1 Deba's Work

1.1 Motivation

There has been an extensive outline of the importance of having a calibrated response from the machine learning models in this presentation. Although calibration is not a very well-established concept, it is intrinsically related to the notion of uncertainty. A model is calibrated correctly if the model properly represents the uncertainty conveyed by the data. Thus, the likelihood of the predictive response for a given test item should be the same as the likelihood of this item in the pool of the training data. This behavior is vital in sensitive machine learning applications such as healthcare, self-driving car, and finance, where the model should not be under- or overconfident in the predictive output. In an underconfident scenario of the model, the likelihood for the test item given by the machine learning model is systematically lower than the actual representation in the pool of data. This under-representation by the model would eventually make the predictive output more susceptible to noisy input data. On the other hand, having an overconfident model would yield predictive output at a higher likelihood value than relative to the likelihood of the test item in the pool of the training data. In this scenario, the model is prone to adversary examples as it will associate the response without the necessary precautions. In a self-driven car example, the model should provide the correct amount of confidence in road segmentation. Sensitive cases are road occlusions from pedestrians where the model should exhibit the necessary amount of caution when segmenting the borders of the street. Similar cases are also present in ventricular segmentation of the heart when physicians plan interventions. Calibration of the machine learning model possesses the potentials to mitigate the drawbacks mentioned above.

1.2 Self Review

In the first paper, [3] the presenter made an excellent introduction of the calibration notion and how it affects the adaption of these technologies in daily routines. Furthermore, from the technical perspective, a range of critical key drivers that affect the calibration while at the same time accuracy needs special attention. The presenter discussed such key drivers include model capacity, batch normalization, weight decay, and negative log-likelihood. Furthermore, an elegant and novel visualization method dedicated at debugging the scale of calibration is outlined. From the proposed reliability diagram, three different metrics that measure the calibration are derived. Namely, expected calibration error (ECE) maximum calibration error, Brier score, and negative log-likelihood. This is followed up with a review of popular calibration methods that have been actively utilized in the field for quite a long time. Such methods are isotonic regression, histogram binning, Bayesian binning into quantiles, Platt scaling. The last one has been further simplified in this paper into a temperature scale. The final predictive output has been optimized from the maximum entropy perspective.

In the second paper, [7] the maximum entropy principle shows potentials at making calibrating the model during the training process. Unlike the other methods where the calibration happens after the training, in this paper, the same principle as temperature scaling is adopted during the training process instead. Furthermore, an interesting observation between accuracy and the prediction's confidence is incorporated in the loss function. Once the confidence is integrated into the cross-entropy loss, a regularization term that makes the predictive output as flat as possible maximizes the entropy of the predictive output. The flatness of the predictive output should be as close to the uniform distribution while using Kullback-Leiber divergence for the comparison.

The last paper, [5] is a thorough outlook of the calibration methods and definition. This paper is the first study that proves an underestimation of the calibration error (CE) for a binning version of model $f : X \rightarrow [0, 1]$, $CE(f_{binned}) \leq CE(f)$.

Similarly, the authors provided a proof of severe miscalibration scale due to the binning effect. Inside a single bin, there are scenarios where the model predictive output is underestimated half the time and overestimated the other half. Consequentially, bin-averaging process will show zero CE which in fact is wrong since the model should have binned in a smaller portions in order to get the averaging correct.

Furthermore, the sample complexity of the scaling and binning method is not at the same proportions. The Platt scaling method scales for a error ϵ the sample size required are $O(1/\epsilon^2)$. Whereas for the the binning method has a complexity of $O(B/\epsilon^2)$ where B is the number of bins. Platt scaling is not as effective method but has low sample complexity, on the other hand, binning method is more effective at estimating CE but has high sample complexity. To combine the best of the two models, the authors proposed a fusion of these two models. A scaling technique follows the binning of the model and the new CE comes from averaging the newly fitted values at each bin instead of raw values. This has lower accuracy than Platt scaling but higher the binning methods, and the new sample complexity is $O(B + 1/\epsilon^2)$.

Last but not least, a verification process for the calibration is first proposed that provides a unbiased metric to compute the bounds for the estimated CE.

1.3 Opposition

The presenter provided the audience with a very good introduction to calibration in machine learning methods. There was a very good explanation of why calibration is such an essential property in sensitive applications. Furthermore, there was excellent coherence in the slides when outlining the parametric and non-parametric methods. The presenter did an in-depth explanation of the advantages and drawbacks of each method.

To benefit the audience, a unifying theory on calibration would be a significant key driver for a better understanding of early adopters of such methods. Further, a more elaborative explanation of calibration that explains the relation to other mathematical approaches, such as probabilistic methods. Treating calibration as an isolated metric as done in the presentation would limit the number of methods that could be developed and hinder the complete understanding of the traditional developed metrics. Calibration error and Brier score along with negative log likelihood require a more systematic revision, which tell the differences and commonualities in calibartion. Last but not least, a preleminary outline of the calibration for regression method is also of great importance but this requires some discussion of the presenter with other peers.

2 Abubakrelsedik Work

2.1 Motivation

Deep learning has revolutionized machine learning; however, this success has been chiefly attributed to the annotation in a massive dataset. Annotation of the dataset is quite an expensive pursuit. Moreover, it is more than often suboptimal due to the lack of expert knowledge in the field and high variance among individuals. This subprime characteristic is even more emphasized in the medical field, where doctors operate on an expensive hourly basis. Furthermore, they are pretty often in disagreement with one another.

Annotation of the training data is not always available. A way around this drawback is to learn semantic features from the raw data without labels. Self-supervised learning (SSL) is an approach that can deliver the same discriminative capabilities without the need for annotated labels. SSL works by withholding part of the data in some form to predict the rest. Similar to SLL, contrastive learning (CL) is a different approach that aims to learn representative features without labels. CL utilizes augmentation to produce two different versions of the same items. The loss function is constructed to maximize the agreement between augmented items originating from the same data and minimize the agreement between those arising from different data. Unsupervised learning also operates without labels, but the ultimate goal of this approach is to learn the latent structure of the global outlook of the dataset, which is not necessarily helpful at providing the best discriminative features. In the subsequent empirical studies, SSL and CL have shown a promising research direction towards reducing or diminishing the need for annotation.

2.2 Self Review

The last paper [2] in the presentation is quite an essential contribution towards SSL as it introduces probably the most elegant setup for performing learning representative features without labels. This method performs a set of random rotations to the input images. The model tries to predict the degree of rotation. Other SSL such as colorization, image completion, and jigsaw puzzles do not ensure that the training process will tune the method to learn semantic representation. Pixel correlation values will also suffice to perform both of these SSL pretext tasks but not rotation prediction. Rotation is quite more dependent on the semantic presentation rather than just generalizing over pixel correlation.

Furthermore, in rotation prediction, the contextual information is kept intact, making the model more sensitive to the contextual information. As a result, the model must have a global outlook of the items residing inside the image. As there is rotation prediction of the image, the most feasible way the model can perform such a task is by understanding the contextual information of the items and tracking their position in the image. This is not necessarily the case in other pretexts such as colorization, which can be performed without the need of the model to understand the items. Colorization is a task that might suffice a suitable generalization of the correlation of pixel gradients. This study is even more critical because the best performance angles are not random but the orthogonal axes, namely 0, 90, 180, 270. This superior performance is empirically proven to be legitimate not only in classification but also in detection and segmentation. It is crucial to notice that an increase in model depth does not necessarily translate into better performance when using the rotation prediction. The model performance increases up to a particular depth and starts decaying from that point onwards.

In the second paper, [4] another critical key driver for the success of deep learning using SSL is the architectural choice. Not having the proper architectural choice can essentially diminish the efforts invested in the SSL training. Architectures that possess skip connections do have better preservation of the performance as the depth of the network increases. This is not the case when skipping connections are not present in the architecture, as is the case for VGG likes. Furthermore, increasing the number of filters that perform the feature distillation and the higher dimensionality of the embeddings increases the quality of the discriminative features quite considerably when performing SSL training. Good accuracy in the pretext task does not necessarily translate into a good performance on the downstream task in all the models. VGG models tend to be less effective compare to Resnet. Last but not least, the faster the learning rate decays, the better the accuracy on the downstream task for a given model.

Another recent approach [6] presented is the improvement of the SSL using context. The primary justification for this approach is the so-called chromatic aberration in the lenses, which is a direct result of frequencies leaving the lens at different angles. Therefore, colors such as magenta and green are radially offset from the center of the image. This offset information is then easily picked by the training process in an SSL setting to find the position of the patches in an image. Consequently, there is a possibility that SSL might not generate context information but instead learn the matching the top and the bottom segment line. Another finding similar to the other paper is the performance of mid-layers while using SSL might not generalize over representative features as good as the first or the last layers. To circumvent this drawback, the learning rate varies across different layers. The authors of the paper proposed chromatic blurring (blurring across channels) to suppress the aberration effect. The random aperture of patches is the other method that creates a specific aperture within an arbitrary position of a given patch. Forcing the network to predict the patches' rotation makes the network less prone to aperture drawbacks.

2.3 Opposition

The presentation was overall smooth, with significant slides indicating the insight from each of the papers. To benefit the people, the presenter should have included more annotated images to explain the concept without borrowing figures from the papers. This way is possible for the presenter to reflect his proper understanding of the method and not be biased from the paper's context. The presenter should have also given an fair comparison between the methods, such as it can force the listeners to gather more insights from each of the methods.

3 Ahmed Work

3.1 Motivation

Graph representation learning (GRL) has attracted considerable attention recently, and this is being mainly attributed to their permutation invariance. Furthermore, GRL is quite a fitting technology when the data are being presented in a non-euclidean form. This capability is significant because most of the data produced are not fully compliant with the Euclidean axioms. Popular datasets where the application of GRL is making the difference are a social network, transportation strategy (i.e., google maps), protein-protein interaction, and so forth. Graph convolutional networks (GCN) are the most popular methods to provide a good representation of the graph data. In a traditional convolutional neural network (CNN), a single item from the pool of data is the one that influences the entire training without taking into account the influence of other items in the collection. At GCN with a single node from the graph, its proximity will also count towards the loss computation. Among many other things, this is the main difference between graph and traditional machine learning (ML). Contextually, traditional ML takes a local outlook solely when being presented with a collection of training data. In contrast, graph ML takes a combination of the local and global perspectives of the training dataset.

3.2 Self Review

HIN2vec [1] is a successful approach where heterogeneous nodes are generalized better than other traditional GCN methods. This method outperformed all the other techniques in node prediction in many famous graph datasets such as Yelp, US patents, etc. Nevertheless, the success of this method is closely tight to the engineering of several factors. The most important thing during the training process is the regularization of meta-path vectors and selecting the node type in negative sampling. If both of these methods' key drivers are not set appropriately, the technique could efficiently behave suboptimally.

Last but not least, this method is very prone to cycles in random walks, and to mitigate this drawback, there is a need for direct prior knowledge in the application domain.

Graph attention network (GAT) has revolutionized the way GRL operates on a structured dataset. To capitalize upon this success, recent work on heterogeneous graphs managed to apply attention mechanism [8]. This method is different from its vanilla predecessor as it can perform hierarchical attention. Namely, the attention coefficient can target node-level and semantic-level, while the vanilla GAT can perform attention solely at the node level. By semantic-level attention, the author refers to the meta path attention, which provides a more compact GRL when combined with nodel level. This method performs exceptionally well in numerous heterogeneous datasets in supervised learning (node, link prediction) and unsupervised learning (clustering).

Contextualized nodes in a graph setting do possess potentials at providing a better node embedding [9]. In traditional GCN, a node has a single embedding which in many cases is not representative enough to convey all the information presented in the interaction model. Having contextual information prior increases the quality of the node embeddings at the downstream task as it presents the node into multiple different perspectives. The interaction between neighboring nodes results from the masking operation for the features and attention coefficients at the node level, where the intermediate nodes. Eventually, multiple different interactions will result as the consequence of multiple different perspectives of the same node. This ensemble of embeddings allows the model to accommodate more information than the case where all these different embeddings are collapsed into a single one.

3.3 Opposition

The presenter did a very good and did explain the main contribution of each paper. Furthermore, the presenter was quite careful at making sure that the main contributions from each paper are appropriately presented to the audience. Nevertheless, the presenter should have put some extra effort into the following points to benefit the audience the most. Throughout the entire presentation, the presenter was spending a considerable amount of time on a single slide. This could make the audience lose their focus quite fast. As a result, the audience might quickly lose the ability to follow the presentation until the end. The presenter did not put additional effort into producing higher-quality figures than necessary. The majority of the figures were taken in the presentation point-blank from the papers, which in many cases do not help present the idea of the paper based on the presenter's understanding.

Last but not least, the presenter should have made a more elaborative introduction of the GRL as this will accelerate the understanding of the presentation even for people who are outside the field.

References

[1] Tao-Yang Fu, Wang-Chien Lee, and Zhen Lei. Hin2vec. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017.

[2] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[4] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *CoRR*, abs/1901.09005, 2019.

[5] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. *CoRR*, abs/1909.10155, 2019.

[6] T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. *CoRR*, abs/1711.06379, 2017.

[7] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Confidence calibration in deep neural networks through stochastic inferences. *CoRR*, abs/1809.10877, 2018.

[8] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, Philip S. Yu, and Yanfang Ye. Heterogeneous graph attention network. *CoRR*, abs/1903.07293, 2019.

[9] Carl Yang, Aditya Pal, Andrew Zhai, Nikil Pancha, Jiawei Han, Charles Rosenberg, and Jure Leskovec. Multisage: Empowering gcn with contextualized multi-embeddings on web-scale multipartite networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 2434–2443, New York, NY, USA, 2020. Association for Computing Machinery.