Life Is Not Black and White - Combining Semi-supervised Learning With Fuzzy Labels

A concept paper

Lars Schmarje[1], Reinhard Koch[1]

[1]*Multimedia Information Processing Group, University of Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany*

Abstract

The required amount of labeled data is one of the biggest issues in deep learning. Semi-Supervised Learning can potentially solve this issue by using additional unlabeled data. However, many datasets suffer from variability in the annotations. The aggregated labels from these annotation are not consistent between different annotators and thus are considered fuzzy. These fuzzy labels are often not considered by Semi-Supervised Learning. This leads either to an inferior performance or to higher initial annotation costs in the complete machine learning development cycle. We envision the incorporation of fuzzy labels into Semi-Supervised Learning and give a proof-of-concept of the potential lower costs and higher consistency in the complete development cycle. As part of our concept, we discuss current limitations, futures research opportunities and potential broad impacts.

Keywords

Semi-Supervised Learning, uncertainty, variability, fuzzy, classification, clustering, annotation, consensus

1. Introduction

Deep Learning was successfully applied to many computer vision problems over the last years. One of the biggest issues with deep learning is the required amount of labeled data for training. Thus, many Semi- and Self-Supervised algorithms have been proposed which can decrease the required labeled data by using additional unlabeled data [1, 2, 3, 4, 5, 6]. These methods aim for the clear vision that we feed the complete data into the network and additionally provide only few labeled samples per class. This step can then be used in a general machine learning development cycle (MLDC, Figure 1a) to reduce the annotation cost. We describe the MLDC as a three step cycle (data collection, model training and model evaluation) based on [7].

However, this vision is missing two important facts because life is not just black and white. Firstly, we as humans have to provide such labels. We will make mistakes and suffer from inconsistency in the form of intra- and interobserver variability [8, 9]. This means that the labels we provide may be wrong or even differ over time or between annotators. Secondly, in the real-world, we often encounter an ambiguous situation where a true label is either difficult to obtain or not existing. For example, the cross bread of two different dogs can not be classified as one of its parents. These two issues have been summarized before as *fuzzy* labels [10, 4]

and have been discussed as problematic for many Semi-Supervised Learning algorithms. The issue of fuzzy labels is common in many life science [11, 12, 13] or medical [14] datasets. Even in curated datasets like Imagenet [15] and CIFAR10 [16, 17] these issues can exist for example due to correlated label noise with visual similar classes [18]. In current research, these issues are often countered by expensive label cleaning [19] or with noise estimation [20]. We believe these perspectives are often too narrow and lead to unnecessary prerequisites of other steps in the MLDC and/or are expensive with regard to the required annotation time. In this paper, we will give an alternative perspective of incorporating Semi-Supervised Learning into a MLDC which is aware of the issue of fuzzy labels (Figure 1c).

1.1. The issue of fuzzy labels

A *fuzzy* label l is a probability distribution over the classes C for an image x. A label is the aggregation of the annotations $a_1, \ldots, a_n$ for x. These annotations are one-hot encoded estimates of the label l by human annotators. Fuzzy labels are aggregated annotations which are not consistent with each other (e.g. $\forall i : a_i \neq \frac{1}{n} \sum_j a_j$). For example, if you have 10 annotations and 5 annotations are for class A while the other 5 are for class B, the aggregated fuzzy label could be 0.5 for each class. If all annotations are consistent with each other, we call the label *consistent*. This definition is similar to soft and hard labels. However, every hard label could also be represented by a one-hot encoded soft-label but a label is either fuzzy or consistent. Throughout this paper, we will call also the corresponding image x fuzzy or consistent depending on its label. A dataset is called consistent if all images are consistent otherwise it is fuzzy.

These fuzzy labels pose two issues we need to overcome to achieve the vision of Semi-Supervised Learning. The first issue is that we need to incorporate the knowledge about the existence of fuzzy labels into the training. Many Semi-Supervised Learning (SSL) algorithms assume that only consistent labels exist in their dataset. Schmarje et al. [10] showed that SLL algorithms have a inferior performance if they do not consider the existence of fuzzy labels. An assumption for the reason for this performance was that fuzzy images are not easily classifiable into the existing classes and thus confuse the algorithm. The fuzzy labels can be incorporate in the training procedure with for example S2C2 which is an extension to most existing SSL algorithms [4]. The second issue is that most algorithms aim at good hard classifications as output for the neural network. This hard classification makes sense for consistent labels but does not work properly for fuzzy labels as they may have no best hard label. If we want to integrate the output of an algorithm into a MLDC, we need to consider this. For example, it could be necessary to resort fuzzy data either into an existing class, a new class or exclude the data.

1.2. Concept

Our main concept is illustrated in Figure 1c. We will first repeat the issue of most Semi-Supervised Learning (SSL) in the MLDC on uncurated data and than show the difference to our idea.

As stated above, SSL aims at reducing the annotation cost by leveraging unlabeled data which is often expected to be consistent or resulting in inferior performance [10]. However, if we have

a uncurated dataset, we argue that the requirement of consistent labels is not given. It is costly to detect these consistent labels in the given dataset and determine their label. The costs are so high because we need to use strategies e.g. a consensus process or strict protocols like [21] to counter intra- and interobserver variability. Moreover, we need a post-processing to decide what we do with difficult, fuzzy images. If we have for example an image with a label of 50% for two classes we have to select a hard label or ignore this image. After this costly dataset cleaning, we can apply any SSL algorithm with a relative low cost (Figure 1b). The generated predictions can than be used to annotate more data more easily. The main issue is that the SSL algorithms expect also cleaned unlabeled data e.g. no fuzzy intermediate images between classes as [10] hypothesises. Otherwise, some main assumptions like the separation of classes in a higher feature space might not be valid.

We envision a machine learning cycle that includes the knowledge of fuzzy labels into the SSL algorithm (Figure 1c). Thus, the initial annotation could then be limited to a small portion of the data which we leverage as labeled data. No assumptions and annotations are then need for the unlabeled data. This could for example be achieved by additional clustering of the data or an automatic distinction between consistent and fuzzy images. A consensus step could still be required in the MLDC to use these output predictions as additional input. But this step could be implemented based on predictions of the network and thus make the annotation and consensus cheaper. Overall, we expect a lower number of required annotations and thus cost while achieving a higher consistency in the data for our envisioned development cycle in comparison to the previous one. We give a proof-of-concept in the next section.

2. Proof-of-Concept

2.1. Dataset

We use the Mice Bone dataset proposed in [4] and show examples in Figure 1d. This dataset consists of gray-scale images of collagen fibers of mice bone. The classification problem has three classes: similar, dissimilar and not relevant fiber structures. We follow [4, 10] and use only consistent images for training and validation and enforce a class balance in this data. We call this consistent and balanced training data *seed* and use the rest of the data as unlabeled data.

We use two different seeds for the Mice Bone data. The first seed is based on the segmentation masks from [22]. Due to the dimension reduction from a segmentation to a classification label and the uncertain segmentation, the labels show high variability and only one annotation is available. We treat all labels as consistent and this leads to an inconsistent seed. The second seed is based on three independent annotations of each image. These annotations allow us to estimate the fuzziness of each image and thus lead to a valid seed.

2.2. Metrics

We aim at faster and more consistent annotations throughout the complete MLDC. Every presented experiment is executed on the same raw data with one of above-described seeds. An experiment consists of three independent annotations of the same person over time for the complete dataset. For ease of referencing, we view these independent annotations over time as
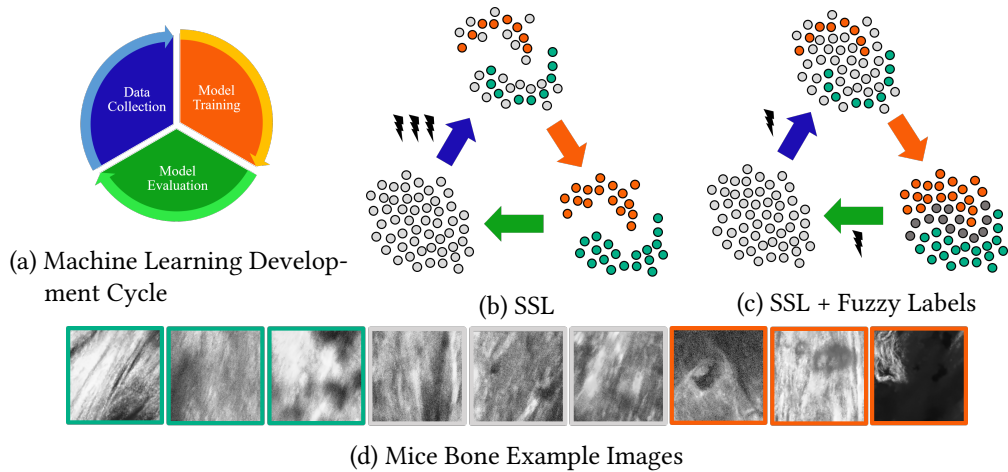
(a) Machine Learning Development Cycle

(b) SSL

(c) SSL + Fuzzy Labels

(d) Mice Bone Example Images

Figure 1: The left image (a) illustrates a simplified machine learning development cycle. The middle (b) and right (c) image how this process can be visualized with Semi-Supervised Learning (SSL) without and with the consideration of fuzzy labels. The light gray circles represent unlabeled images. The colored circles represent labeled images or the prediction of these labeled images. Dark gray circles represent datapoints which are identified as fuzzy. The number of thunderbolts indicates the number of required annotations by humans in each step. In the lower figure (d), we show 9 real-world Mice Bone examples from [4]. The images of the classes similar (green) and dissimilar (orange) fiber orientations are easy to consistently annotate. The middle images (grey) are more difficult to consistently annotate and thus are fuzzy.

three (intra-)annotators. The annotations were either done without any additional information or with different outputs of an SSL algorithm given as *support (predictions)*. These support predictions could be accepted or manually corrected. This support can also be an overclustering of the data. An overclustering is a clustering with more cluster than used classes [10, 4]. We give in total 6 different evaluation metrics which are either average over the three annotators or the cross-combinations between them.

Cohen's *kappa* coefficient (κ) is a statistical metric that is often used to measure the intra- and inter-observer variability [23]. The coefficient measures the agreement between two annotators for a classification task. *Accuracy* (Acc.) measures the true positives in relation to all annotations between two annotators. *F1-Score* aggregate the precision and recall between the annotation of two annotators. *Total Accuracy* (T.Acc.) counts the consistent annotations between all annotators and divides them by the total number. This metric is like Acc. but not for a pair of annotators but for all three annotators in parallel. *Consistency* (Cons.) divides how many predictions were considered consistent by the total number of images. An image is consistent with the given support predictions, if no manual correction was applied. *Time* is the time it took the annotator to label the complete dataset.

2.3. Results

We discuss six different experimental setup with support predictions as proof-of-concept results in Table 1. The first experiments used no support predictions. These results of this experiment

Table 1
The first three rows describe the experiment in each row. An x marks the usage of a clean seed and — means not applicable because no support predictions were used. The abbreviations for the other columns are defined in subsection 2.2. The best result per column is marked bold.

Method for support predictions	Valid Seed	Used Output	Scores					
			$\kappa \uparrow$ [%]	Acc. $\uparrow$ [%]	F1 $\uparrow$ [%]	T.Acc. $\uparrow$ [%]	Cons. $\uparrow$ [%]	Time $\downarrow$ [min]
None		None	71.35 ± 2.57	84.53 ± 1.23	80.23 ± 2.05	77.21	—	13.95 ± 2.25
FOC [10]		p_n	48.14 ± 11.9	71.78 ± 7.37	64.48 ± 8.25	59.24	65.15	9.09 ± 0.57
FOC [10]		p_o	56.4 ± 8.67	73.76 ± 5.55	68.69 ± 6.54	62.85	69.66	7.08 ± 0.44
S2C2 [4]	x	p_n	79.95 ± 1.03	87.92 ± 0.56	85.66 ± 1.14	82.01	**85.26**	$\mathbf{5.15 \pm 0.60}$
S2C2 [4]	x	p_o	74.82 ± 0.17	84.15 ± 0.15	83.57 ± 0.24	76.74	84.96	5.51 ± 0.32
S2C2 [4]	x	p_n & p_o	$\mathbf{83.17 \pm 2.16}$	$\mathbf{89.59 \pm 1.29}$	$\mathbf{88.08 \pm 1.62}$	**84.58**	84.70	5.18 ± 0.63

were used to calculate the above-mentioned clean seed. The second and third experiments used support predictions from the inconsistent seed with the method FOC [10]. Either a classification head p_n or an overclustering head p_o was used as output resulting in a classification or over-clustering of the data respectively. The last three experiments used a support calculated on the clean seed with the method S2C2 [4]. As output either classification head p_n, a overclustering head p_o or both heads (p_n & p_o) based on the fuzziness estimation were used.

We want to highlight three important aspects. Firstly, the usage of support reduces the required annotation time while not necessarily improving its consistency. Secondly, the consistency with support is worse with an inconsistent seed and the method FOC in comparison to the others. Due to the similar architecture of FOC and S2C2, we attribute this worse performance rather to the inconsistent seed than to the method. Thirdly, S2C2 (with both outputs) more than halves the required annotation time while increases almost all other metrics.

3. Discussion

3.1. Future Research

Our proof-of-concept experiments show that using the output SSL algorithms that are aware of fuzzy labels can improve the consistency of acquired labels and reduce their annotation time. However, this benefit is not always achieved when using network predictions which enforces the importance of careful seed selection and/or SSL algorithm. As a first proof-of-concept, these results lack additional experiments such as different cross-combinations and comparisons to noise estimation and other algorithms. While these issues have to be addressed in future research, the results illustrate the potential benefits of our concept for a MLDC that considers fuzzy labels. Future research could also incorporate uncertainty estimates from humans and neural network predictions into the annotation process. A promising research direction would also be large user studies and the simulation of such studies for detailed ablations.

3.2. Broader Impact

Semi-Supervised Learnig (SSL) aims at decreasing the required amount of labeled data to a few samples per class. We envision a MLDC which allows us to apply SSL to most classification problems with relatively low cost. Especially, in domains like medical imaging, the required labeled data is a severe limiting factor. If we could resolve this issue with SSL which considers fuzzy labels, we would open a vast variety of new research opportunities at a large scale which is currently not feasible.

If we incorporate model predictions into our annotation process, we need to be aware that we might suffer from a confirmation bias. In the worst cases scenario, we could create a self-fulfilling prophecy which would lead to a degeneration of the complete MLDC. However, by carefully monitoring our processes as part of the development cycle we can detect these issues early. Additionally, we assume that cautiously leveraging this bias can lead to more high-quality data which can improve future algorithms and research.

Moreover, we advocate a change of evaluation perspective. While it is important to evaluate on highly curated datasets for algorithm development, we also have to look at the complete MLDC. Otherwise we can not detect the issues described in this paper that fuzzy labels are a limiting factor for applying SSL to different real-world data. We already see great benefits when applying deep learning to other research fields but decreasing such limiting factors further can potentially increase the impact of deep learning even more.

4. Conclusion

Semi-Supervised Learning (SSL) has great potential by solving one big issue of deep learning: The required amount of labeled data and its cost. However, many SSL algorithms do not take fuzzy labels into account and thus require expensive prepossessing steps of the data. By including the knowledge of fuzzy labels into our machine learning development cycle (MLDC), we envision that we can decrease the overall annotation time and thus its costs while achieving a higher consistency. We give a proof-of-concept of this concept on a Mice Bone dataset with fuzzy labels. We highlight that a lot of future research opportunities exist to validate and improve the presented ideas. We advocate to broaden our view from the algorithm development to the the complete MLDC to detect and overcome issues like fuzzy labels. This could spark a variety of before-infeasible research opportunities and thus lead to new breakthroughs in science in general.

References

[1] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, arXiv preprint arXiv:2002.05709 (2020) 1597–1607. arXiv:2002.05709.

[2] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big Self-Supervised Models are Strong Semi-Supervised Learners, Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) (2020). arXiv:2006.10029.

[3] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) (2020). arXiv:2001.07685.

[4] L. Schmarje, M. Santarossa, S.-M. Schröder, C. Zelenka, R. Kiko, J. Stracke, N. Volkmann, R. Koch, S2C2 - An orthogonal method for Semi-Supervised Learning on fuzzy labels (2021).

[5] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent: A new approach to self-supervised Learning, Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) (2020). arXiv:2006.07733.

[6] L. Schmarje, M. Santarossa, S.-M. Schroder, R. Koch, A Survey on Semi-, Self-and Unsupervised Learning for Image Classification, IEEE Access (2021) 1–1. arXiv:2002.08721.

[7] S. Biderman, W. J. Scheirer, Pitfalls in Machine Learning Research: Reexamining the Development Cycle (2020). arXiv:2011.02832.

[8] P. F. E. Addison, D. J. Collins, R. Trebilco, S. Howe, N. Bax, P. Hedge, G. Jones, P. Miloslavich, C. Roelfsema, M. Sams, R. D. Stuart-Smith, P. Scanes, P. Von Baumgarten, A. McQuatters-Gollop, A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting, ICES Journal of Marine Science 75 (2018) 941–952.

[9] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, Medical Image Analysis 65 (2020). arXiv:1912.02911.

[10] L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, R. Koch, Beyond Cats and Dogs: Semi-supervised Classification of fuzzy labels with overclustering, arXiv preprint arXiv:2012.01768 (2020). arXiv:2012.01768.

[11] P. Culverhouse, R. Williams, B. Reguera, V. Herry, S. González-Gil, Do experts make mistakes? A comparison of human and machine identification of dinoflagellates, Marine Ecology Progress Series 247 (2003) 17–25.

[12] J. Brünger, S. Dippel, R. Koch, C. Veit, 'Tailception': using neural networks for assessing tail lesions on pictures of pig carcasses, Animal 13 (2019) 1030–1036.

[13] J. M. Durden, B. J. Bett, T. Schoening, K. J. Morris, T. W. Nattkemper, H. A. Ruhl, Comparison of image annotation data generated by multiple investigators for benthic ecology, Marine Ecology Progress Series 552 (2016) 61–70.

[14] E. Ooms, H. Zonderland, M. Eijkemans, M. Kriege, B. Mahdavian Delavary, C. Burger, A. Ansink, Mammography: Interobserver variability in breast density assessment, The Breast 16 (2007) 568–576.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, volume 60, Association for Computing Machinery, 2012, pp. 1097–1105.

[16] A. Krizhevsky, G. Hinton, Others, Learning multiple layers of features from tiny images, Technical Report, Citeseer, 2009.

[17] J. Peterson, R. Battleday, T. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, Proceedings of the IEEE International Conference on Computer Vision

2019-Octob (2019) 9616–9625. `arXiv:1908.07086`.

[18] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, J. Berent, Correlated Input-Dependent Label Noise in Large-Scale Image Classification, CVPR (2021) 1551–1560. `arXiv:2105.10305`.

[19] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, A. van den Oord, Are we done with ImageNet?, arXiv preprint arXiv:2006.07159 (2020). `arXiv:2006.07159`.

[20] J. Li, R. Socher, S. C. H. Hoi, DivideMix: Learning with Noisy Labels as Semi-supervised Learning, in: International Conference on Learning Representations, 2020, pp. 1–14. `arXiv:2002.07394`.

[21] V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, B. C. Wintle, A practical guide to structured expert elicitation using the IDEA protocol, Methods in Ecology and Evolution 9 (2018) 169–180.

[22] L. Schmarje, C. Zelenka, U. Geisen, C.-C. Glüer, R. Koch, 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy, in: DAGM German Conference of Pattern Regocnition, volume 11824 LNCS, Springer, 2019, pp. 374–386. `arXiv:1907.12868`.

[23] M. L. McHugh, Interrater reliability: the kappa statistic, PubMed Biochemia (2012) 276–82.