Marcel Dettling

Impact of Mislabelling on Deep Learning Methods and Strategies for Improvement

*Marcel Dettling[1] (marcel.dettling@zhaw.ch)
Martin Frey[1]; (martin.frey@zhaw.ch)
Manuel Walser[1] (manuel.walser@zhaw.ch)
Patrick Haas[2] (patrick.haas@tracktics.zone)

[1] Institute for Data Analysis and Process Design, ZHAW School of Engineering, Zurich University of Applied Sciences, Winterthur, Switzerland
[2] Tracktics GmbH, Zürich, Switzerland

Abstract

This contribution revolves around classifying football player actions with 1-dimensional convolutional neural networks (CNNs) based on 6-channel inertial motion unit (IMU) data arising from tracking devices worn by the players. Our training and test data consist of eight games, where humans labelled ball actions by inspecting video records. Unfortunately, these labels are far from perfect due to various reasons (e.g., sloppiness, not all players respectively ball actions visible in the record, ambiguity what a ball action is, etc.). Such mislabelled data provide challenges on several levels. First, performance evaluation with poorly annotated data can be strongly misleading, indicating inferior performance than what is truly achieved. Second, the question is what amount of mislabelled data deep artificial neural networks can tolerate before they break down. We try to shed some light on the magnitude of these effects by simulation studies on the football data, as well as some standard machine learning datasets such as MNIST (numbers) and Fashion-MNIST (clothes). Third, we present some efficient strategies to overcome the issue with imperfect labels and aim to provide some guidelines how to efficiently invest effort in labelling data.

Keywords
Classification; Mislabelling; Deep Learning; Neural Network; Sports Analytics

1. Introduction

Until recently, the evaluation of motion data from sports was out of reach for amateur teams, because the equipment was expensive and required manual work by experts [1]. With the upraise and easy accessibility of global positioning systems (GPS) and inertial measurement unit (IMU) technologies, wearable tracking devices became readily available. The Swiss start-up company TRACKTICS has developed and distributes the first commercially available tracking system for amateur football teams (see http://www.tracktics.com). It comes at affordable cost and consists of a wearable tracking device that records position (GPS) and movement (IMU). Equally important is the accompanying software for processing the data and routines for evaluating the games. Modern machine learning has shown to be useful for the latter task [2].

An important metric of team and player performance is the number and temporal distribution of ball actions. These are not straightforward to detect from the tracking data, as neither information on the ball position nor from the opponent are available. For a set of eight 90-minute amateur games, humans labelled ball actions by analysing video records. It comes as

little surprise that these labels are imperfect. Only one camera was available, hence, not all ball actions are visible, players may be unidentifiable, there is ambiguity in the definition and most certainly, there are blatant errors from sloppiness as it is tedious work. Nevertheless, we trained 1-dimensional convolutional neural networks ([3], see below for details) for detecting ball actions from these labelled tracking data, which reached a convincing preliminary accuracy of 96.87%. However, the preliminary false discovery rate for ball actions at 38.51% was substantial.

Video analysis of misclassified snippets revealed that many originated from incorrectly labelled data. This raised two major questions, namely what the true performance of the network is, and what loss in performance due to the incorrectly labelled training data we need to expect. For providing a sound answer to these questions, correctly labelled data are indispensable. Hence, we developed an application for efficient relabelling of data (see Figure 1) and with these, ran simulation studies for quantifying the true performance. In the literature, various contributions on the effect of mislabelled data are available, see e.g. [4] and [5] for recent work. However, with neural networks, much of the effect remains specific for the architecture that was employed; hence, our investigations were clearly necessary. To complement the picture, we extended our simulations using the MNIST handwritten digits database [6] and the Fashion-MNIST dataset showing different types of clothes [7].

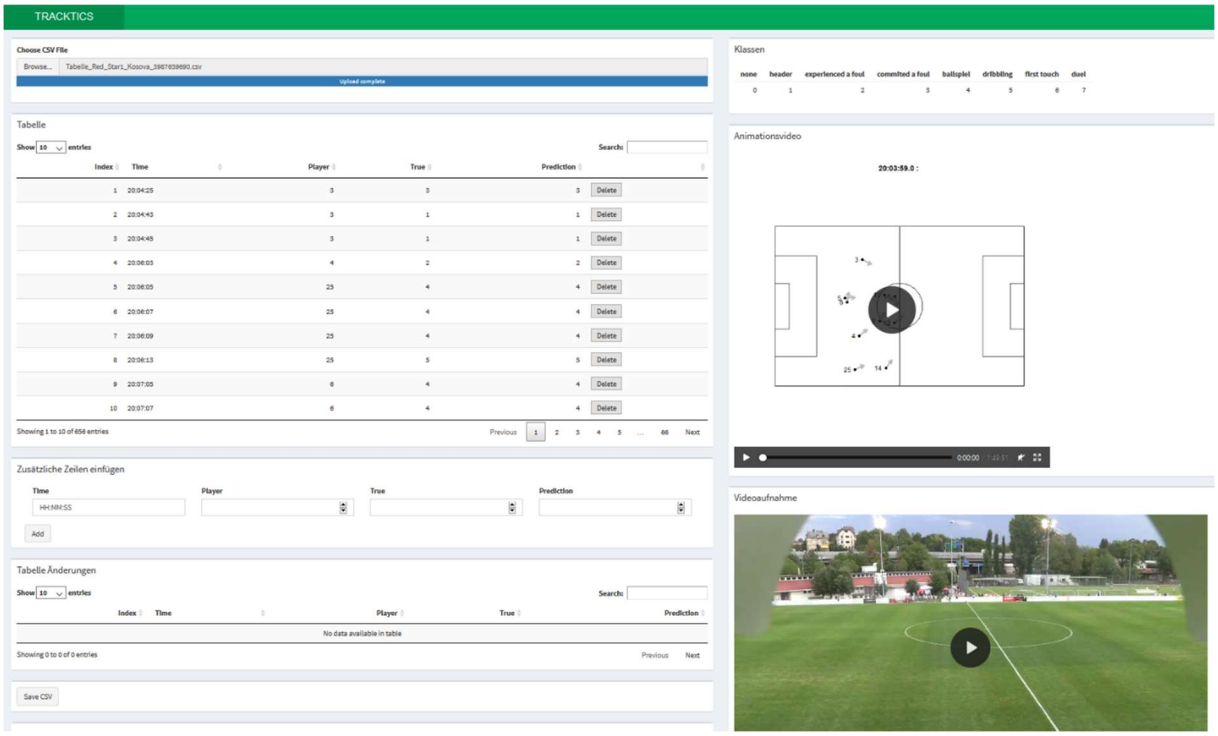

Figure 1: Screenshot of relabelling tool for improving the annotation of ball actions.

2. Methodology

The ball-action-classification-method exclusively builds on the 6 channel-IMU data that are available at a frequency of 200Hz. Position data from GPS provided little added value, neither does the algorithm incorporate any information from teammates. Hence, it is on a fully individual level. The IMU tracking data for each player were cut into 5s pieces with 3s of overlap each. A convolutional neural network with seven 1-dimensional layers (1d-CNN, see [8]) was fitted using the data of 5 games, while 2 further games were used as a validation set and 1 game was retained as the test set. The test data are strongly imbalanced with 1463 ball actions vs. 31186 non-actions, yielding an a-priori-rate of 4.48%. While the accuracy with 96.87% was reasonable, we observed many false positives (38.51%) among the predicted

ball actions. A close inspection of the video records showed that many of these were true ball actions that for some reason had been mislabelled. Thus, we programmed the relabelling tool in Figure 1 where game footage, an animation of the GPS position, human and predicted labels plus a tab for data modification were present.

For further enhancing our understanding the effect of mislabelling on network training and evaluation, we ran simulation studies. The one for the football games involves a big effort in relabelling the data, which at present is not complete. Hence, we cannot provide any conclusive results for this application of CNNs, we here elaborate on additional simulations that we performed on benchmark machine learning datasets. These are parts of the MNIST handwritten digit database [6] and the Fashion-MNIST data [7]. Both datasets consist of 8000 data points per class that we split into 75% training and 25% test.

For the MNIST numbers dataset, we configured an “easy” problem where the digits 1 and 5 needed to be classified, as well as a “hard” problem with digits 3 and 8. In each problem and round, a fraction of labels (e.g. 35%, 36%, …, 65%) were swapped, before the network was trained and evaluated. Please note that these problems are easy to handle with correct labels, achieving close to 100% test set accuracy with the method that we used. The easy problem is achieving slightly better results. We studied the effect of the mislabelled training data by evaluating against 1) the true, unfalsified labels, 2) the false, swapped labels. We applied three different models to understand how different architectures react to mislabelling. The first model is a shallow CNN, detailed in Figure 2. The second is a deeper CNN and finally for comparison, we fitted a Random Forest. In this paper, we only present the results of the shallow CNN.

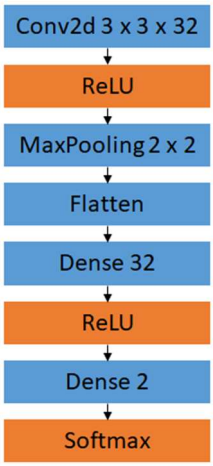

Figure 2: Architecture for the simulation study

3. Results

We here show the results from our simulation study on the MNIST data mainly in the form of graphical presentations. In each plot, the y-axis reports the test set accuracy, depending on the percentage of mislabelled data that is on the x-axis. Both the traces for the “easy” problem (1/5) and the “hard” problem (3/8) are shown in the same plot. We performed three runs each in exactly the same configuration, so that we can provide standard errors. In Figure 3 we observe that the CNN classifier is largely insensitive against even massive amounts of mislabelling in the training data. With 35% of swapped labels, the classification on the test data remains largely error-free. Only when >45% of the labels are altered, the method begins to break down. From this we can conjecture, that in balanced and relatively simple classification tasks, CNNs are highly robust. However, in practice as with our football games, we usually cannot count on error-free test set labels. This has a much more devastating effect on the observed performance, as Figure 4 shows.

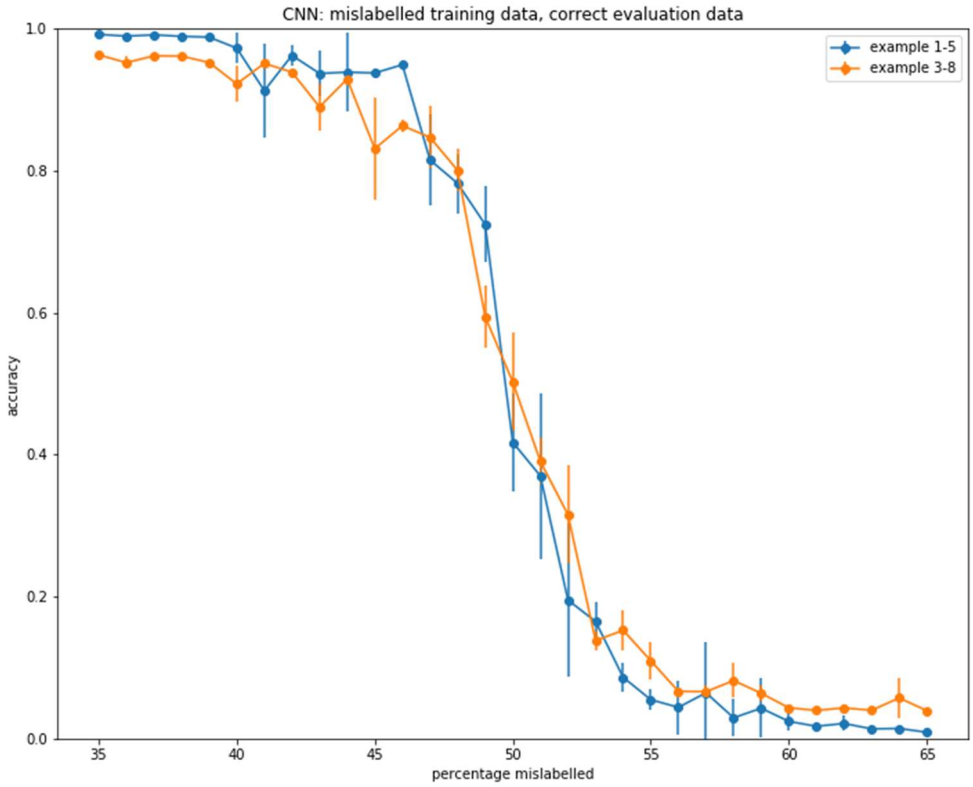

Figure 3: Accuracy of CNN for 1 vs. 5 resp. 3 vs. 8 classification (MNIST data) with different fractions of mislabelled training data, evaluated on test data with correct (non-swapped) labels.

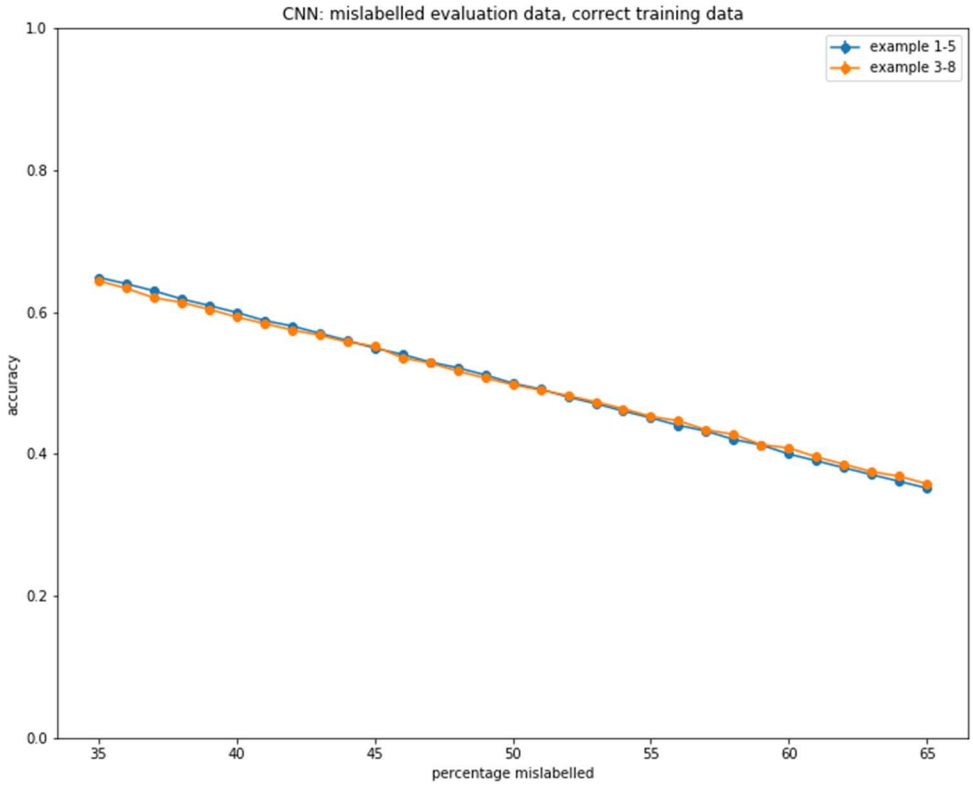

Figure 4: Accuracy of CNN for 1 vs. 5 resp. 3 vs. 8 classification (MNIST data) with correctly labelled training data, but mislabelled test data where different fractions of the test set had swapped labels.

If the data are correctly labelled (both training and test), the CNN achieves a virtually perfect classification of two-class MNIST digit distinctions. However, if there are label errors in the test data, they directly and almost linearly degrade the measured test set accuracy. In verbatim, if 35% of the test data have their labels swapped, we recognize many of them being falsely predicted – which they are not, as the predicted label is correct, but it is only the test label, which is wrong and indicates a pseudo-false classification. Hence, even if the CNN is very robust against the effect of mislabelling, the standard way of evaluating the performance is not. Figure 5 finally shows the "real life situation" where both the training data and the test data have imperfect labels. Again, the perceived performance in terms of test set accuracy is poor. This has nothing to do with the CNN being irritated by the falsely labelled training observation, but almost exclusively stems from the swapped test labels.

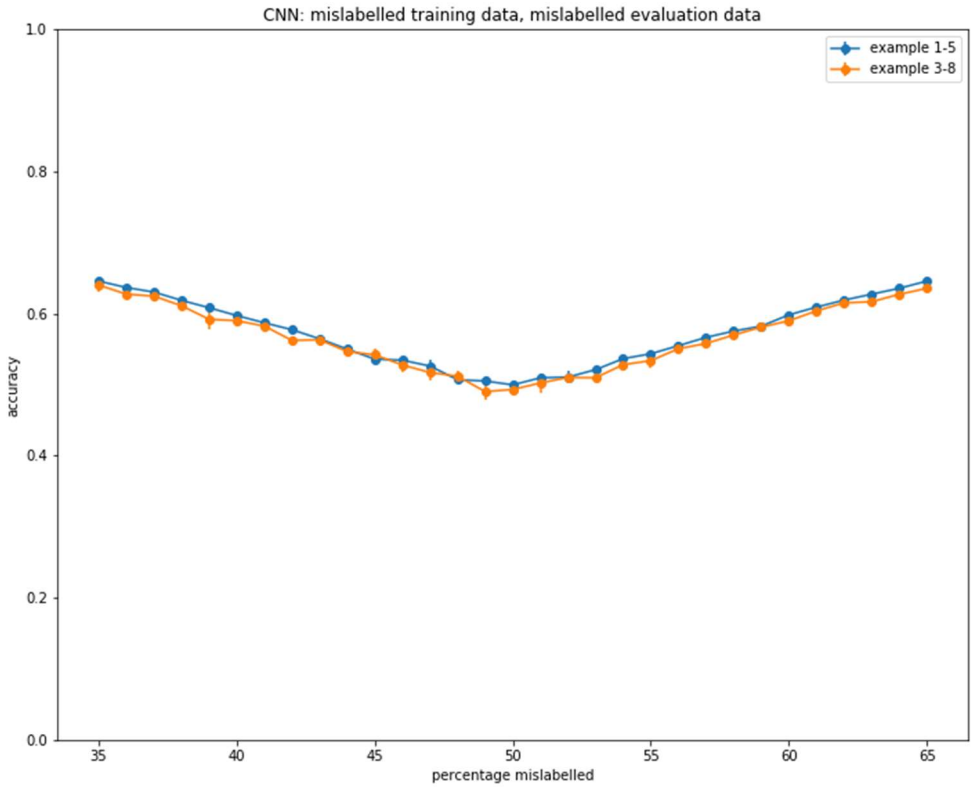

Figure 5: Accuracy of CNN for 1 vs. 5 resp. 3 vs. 8 classification (MNIST data) with correctly labelled training data, but mislabelled test data where different fractions of the test set had swapped labels.

4. Discussion and Conclusion

Our research points out a few important lessons. First, neural networks are surprisingly robust against wrongly labelled training data. At first, this may seem surprising for such a flexible, highly parametrized method, for which one could suspect a tendency to overfit. However, these networks are also based on efficient regularization strategies that apparently can cope with major fractions of mislabelled data. In practice however, the insensitivity of a neural network classifier against imperfect training labels cannot easily be backed up. Yet, the performance may even look very poor if evaluated against equally mislabelled test data. While the classifier performs its work flawlessly and predicts correct class labels, they do not match with the erroneous test labels. In practice, this means that if one suspects "poor performance" due to mislabelled data, it will mostly be sufficient to verify the labels for a relatively small amount of test data, resp. one can even focus on the misclassified test observations and double-check only these.

What remains to be done is to verify the observed phenomenon on the ball action data. It has a few characteristics that are different from the MNIST handwritten digit data. In particular, the problem is harder, i.e. does most likely not allow for near-perfect classification even without labelling issues. Furthermore, the classes are imbalanced which may lead to marginalization of the rarer class with its associated issues for performance measurement under mislabelling.

References

[1] J. Castellano, D. Alvarez-Pastor and P.S. Bradley (2014). Evaluation of Research using Computerised Tracking Systems (Amisco® and Prozone®) to Analyse Physical Performance in Elite Soccer: A Systematic Review. Sports Medicine, 44(5), 701-712.

[2] R. Rein and D. Memmert (2016), “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science” in SpringerPlus 5:1410, DOI 10.1186/s40064-016-3108-2.

[3] S. Bai, J.Z. Kolter and V. Koltun (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. ArXiv preprint (https://arxiv.org/abs/1803.01271).

[4] G. Algan and I. Ulusoy (2020). Label Noise Types and Their Effects on Deep Learning. ArXiv preprint (https://arxiv.org/abs/2003.10471).

[5] B. Frenay and M. Verleysen (2014). Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, No. 5, P. 845-869. DOI: 10.1109/TNNLS.2013.2292894.

[6] MNIST: Handwritten Digit Database. Y. LeCun, C. Cortes and C.J. Burges. http://yann.lecun.com/exdb/mnist.

[7] Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. H. Xiao, K. Rasul and R. Vollgraf. https://arxiv.org/abs/1708.07747.