Utilizing convolutional neural networks for categorising ground-based cloud observations – Supplementary material

Markus Rosenberger - markus.rosenberger@univie.ac.at

First of all: If you have any questions, comments, or suggestions regarding our work, please do not hesitate to contact us via the above given address. We are happy about any contributions.

Summary

Summed up, we have proven that Residual Neural Networks can be trained for skillful and reliable cloud classification from ground-based pictures. The most prominent obstacle during the training process were the highly imbalanced ground truth observation frequencies. In order to reduce these biases we carried out class specific data augmentation by adding real pictures to less abundant classes. However, results indicate that this only reduced prediction biases but did not completely resolve them. Further work will be needed on this issue. Still, MLCM and reliability diagrams indicate skillful performance in almost all cloud classes when compared to climatology or random guessing as forecast methods.

Applications of automated cloud classifiers, as the one we have trained, could be for example as a cloud monitoring system close to a solar power plant to obtain information on upcoming irradiation deficiencies in real time. Cloud class information could also be used for data assimilation in operational NWP.

Supplementary Material

Following research questions are tackled in our work:

1. Can machine learning methods discriminate between 30 SYNOP cloud classes from ground based RGB pictures?

2. How can we overcome biases due to observation imbalances?

3. How reliable are predicted probabilities of occurrence?

Different approaches have already been tried to automatically retrieve cloud types from satellite data during the past decades. However, since the WMO defined cloud observation standards in their cloud atlas via visual properties seen from the Earth's surface, the results of these previously

Figure 1: Example picture of the northward facing camera.

mentioned methods can hardly be compared directly to human observations. Thus in this work we try to find a way to automatically retrieve cloud classes from conventional RGB pictures taken at the Earth's surface. Our dataset consists of pictures taken at one location in Vienna with four cameras, each pointing in a different main cardinal direction in order to cover the whole sky. The main advantage of taking images from a single camera system rather than from several different sources, is that the model benefits from a homogeneous data set. An example picture is shown in Fig. 1.

Pictures are available in 2 distinct periods, from 05.10.2016 — 18.02.2019 and from 04.05.2022 onwards. Since for operational human cloud observations the whole visible sky is considered, we also use all 4 pictures at once for every instance. Moreover, the order of the sub-images varies randomly for each training epoch to increase robustness of our model. Hourly operational cloud observations at the station Vienna Hohe Warte are used as ground truth since this station and the cameras are less than 2 km apart. Ground-truth observations consist of 30 classes, 10 per cloud height level. In the lowest level there is always an observed class, in each of the upper two levels not more than one cloud class is reported. Thus the ground-truth vector of a single instance consists of 1–3 out of 30 cloud classes making this a multi-label classification problem.

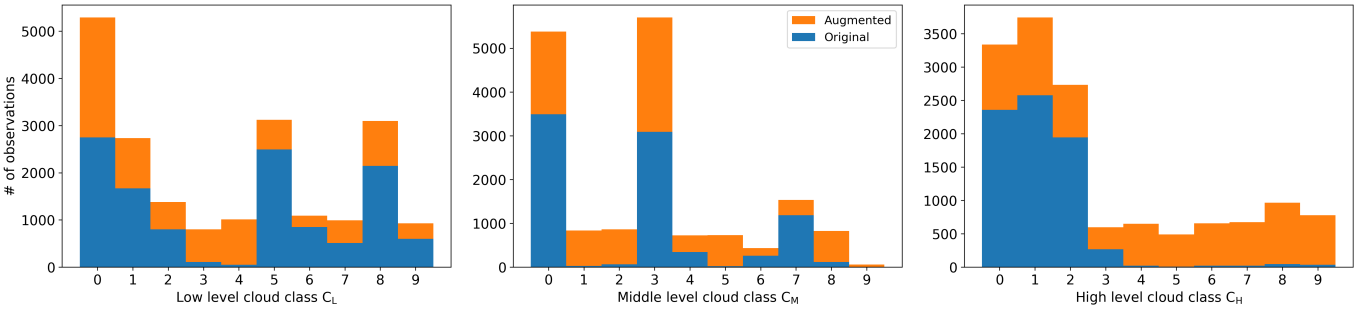


Figure 2: Cloud class observation frequencies in the raw data set (blue bars) and after class specific data augmentation (orange).

As can be inferred from the second research question, probably the most challenging task is to overcome the large biases introduced by large observation frequency imbalances between different classes. We tackle this problem with a class specific data augmentation method, where we add real pictures to the dataset, which contain cloud classes that are least abundant in the raw data.

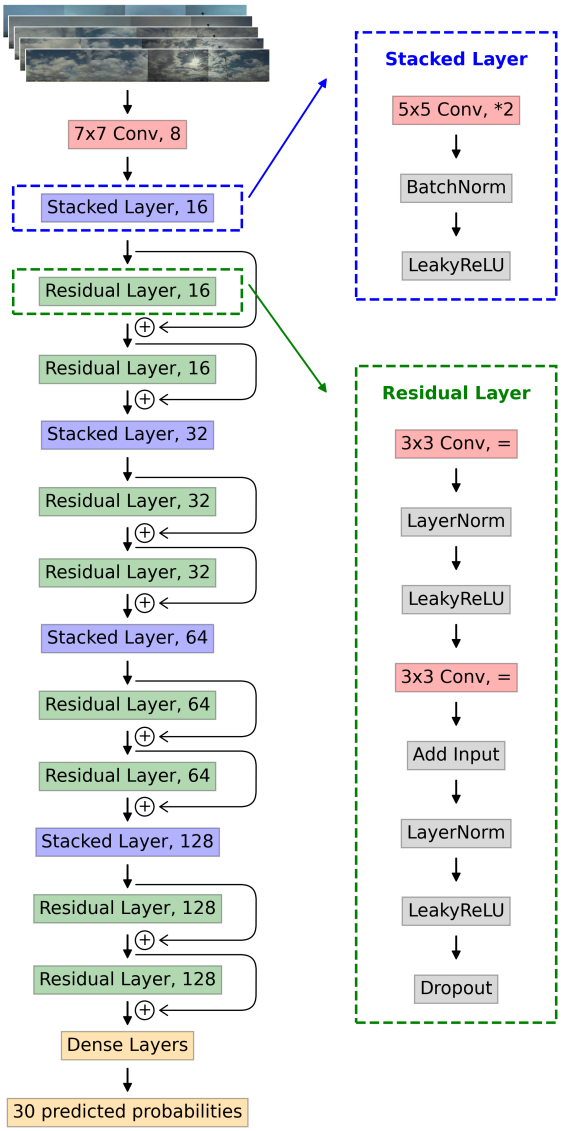


Figure 3: Schematic representation of our model archtitecture.

Blue bars in Fig. 2 show the cloud class distribution of the raw dataset and orange bars show the distribution of the final data set after the class specific data augmentation process.

Convolutional neural networks (CNNs) proved to deliver sufficient results in a broad range of image classification tasks. He et al. (2015) showed that Residual Neural Networks have the potential to outperform classical Deep CNNs. Hence, in this work we follow their approach and train such a Residual Neural Network from scratch. Our model consists of several blocks of convolution layers followed by a normalization layer and an activation function. The exact architecture is shown schematically in Fig. 3.

In order to evaluate the performance of our model on the classification task we calculate confusion matrices. However, since more than one cloud class can in principle be observed and thus also predicted per instance, we use the Multi-Label Confusion Matrix (MLCM; Heydarian et al., 2022). The MLCM is evaluated on the validation dataset and true positive (TP) classifications are located on the main diagonal. Apart from the main diagonal, false negative (FN) rates for each observed class are shown in the corresponding row and false positive (FP) rates are shown in the columns of the matrix. True Negative (TN) classifications can be found in the main diagonal apart from the TP

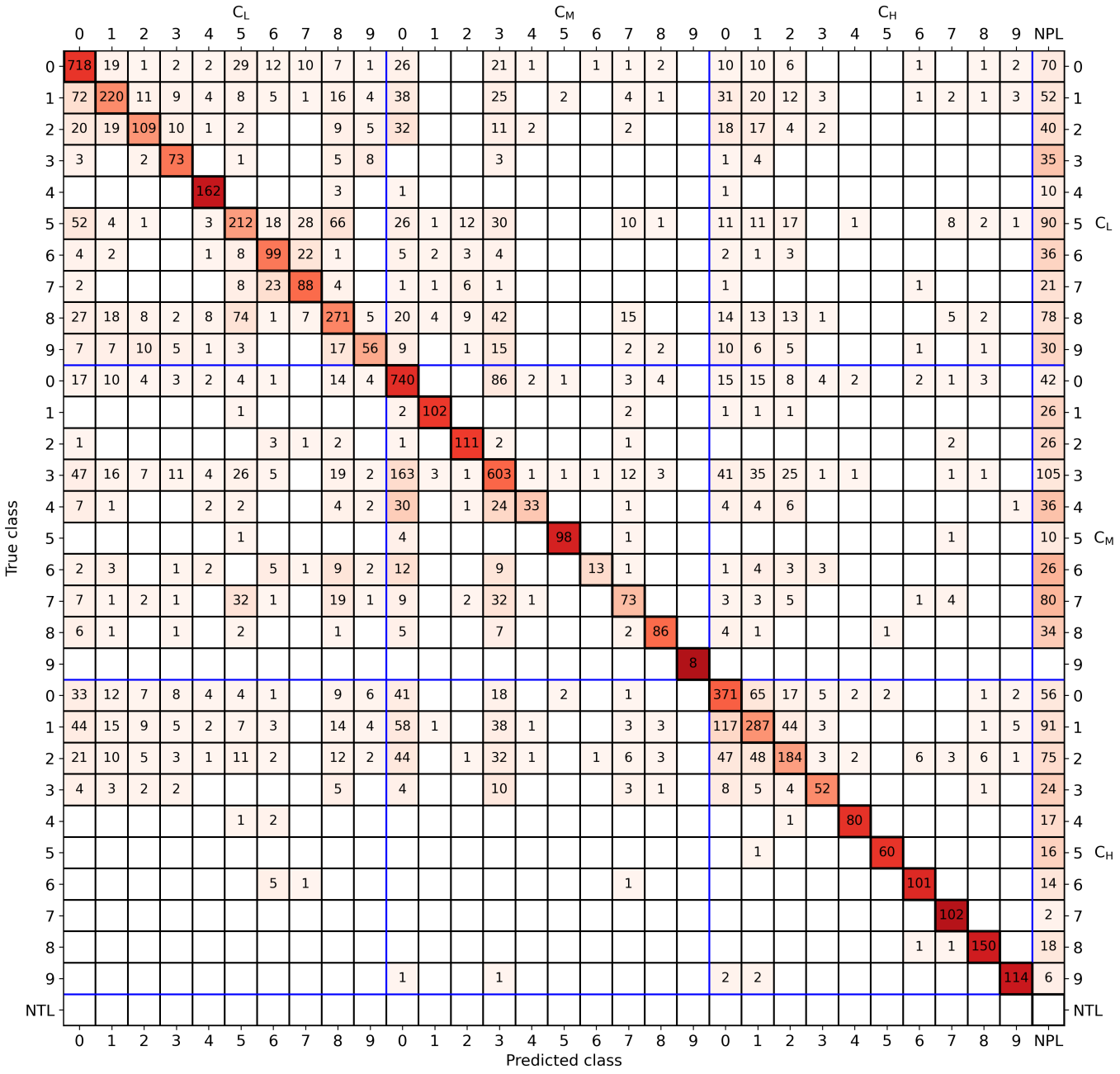
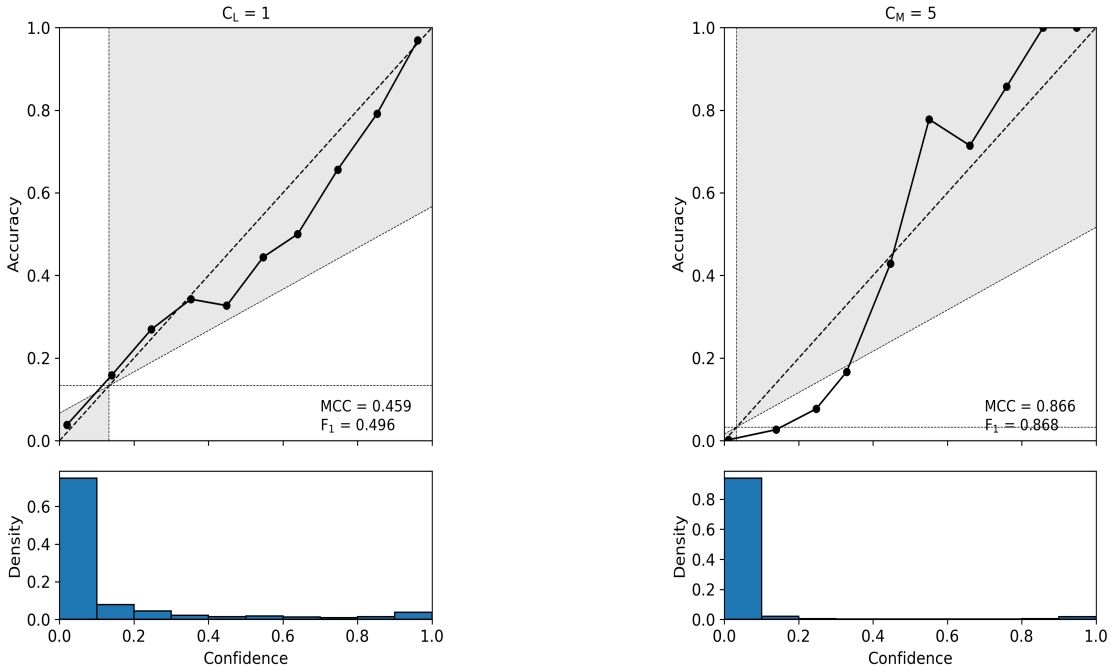


Figure 4: Multi-Label Confusion Matrix indicates that our model classifies the vast majority of instances correctly.

entry.

The only substantial difference between the MLCM and a classical confidence matrix is the last column, called *NPL* which stands for *No predicted label*. This column contains all the instances where none of the predicted probabilities was >0.5, hence no class was considered to be positively predicted. Also the last row, *NTL (No True Label)*, is different to classical confusion matrices, though no instance is present in this row in our work since each instance has at least one ground truth label. Fig. 4 clearly shows that the vast majority of instances is classified correctly by our model. However, there is also a substantial amount of instances where our model struggles to find the correct cloud classes. These FP and FN cases can be grouped into several categories:

- Prediction bias due to observation imbalances. Highly abundant classes in the raw dataset



(a) (b)

Figure 5: Reliability diagrams that are representative for the performance on highly abundant classes (a, very good reliability and resolution) and aggressively augmented classes (b, underconfident but still well performing).

are also predicted more often than others, e.g. classes $C_L = 5$, $C_M = 3$.

- Too small predicted probabilities. These instances can be found in the *NPL* column and are present for all classes.
- Misclassifications due to visible similarities of the classes, e.g. $C_L = 5$ and $C_L = 8$.
- Misclassifications due to evolution of the cloud. Some classes have to be chosen when a cloud underwent an evolution from a specific mother cloud type. This temporal evolution is often not visible from a single timestamp frame and thus leads to errors, e.g. $C_M = 6$.

Apart from classification accuracy, for probabilistic forecasts also the reliability is a very important parameter. A probabilistic forecast model is said to be reliable or well calibrated, when the predicted probability of a specific outcome represents the actual probability of occurrence of this event. To assess this property, reliability diagrams are used. For a well calibrated forecast, points in this diagram lie closely to the diagonal 1:1 line. Fig. 5 shows 2 representative reliability diagrams for our model. Fig. 5a indicates very good reliability and resolution of our model, since all dots lie very closely to the diagonal line. This is the case for those classes with high observation frequencies. On the other hand, Fig. 5b represents those classes that have been augmented rather aggressively. There, the model tends to show what is called underconfidence, i.e. small probabilities of occurrence are overestimated by the model and high probabilities of occurrence are underestimated. This leads to the S-shape of the black solid line. However, overall our model is highly skillful in almost all classes when compared to climatology, i.e. to the natural probabilities of observation. This can be inferred from the gray areas in the reliability diagrams. Each dot in such a gray area indicates that the model's performance is more skillful than using climatology as forecast.

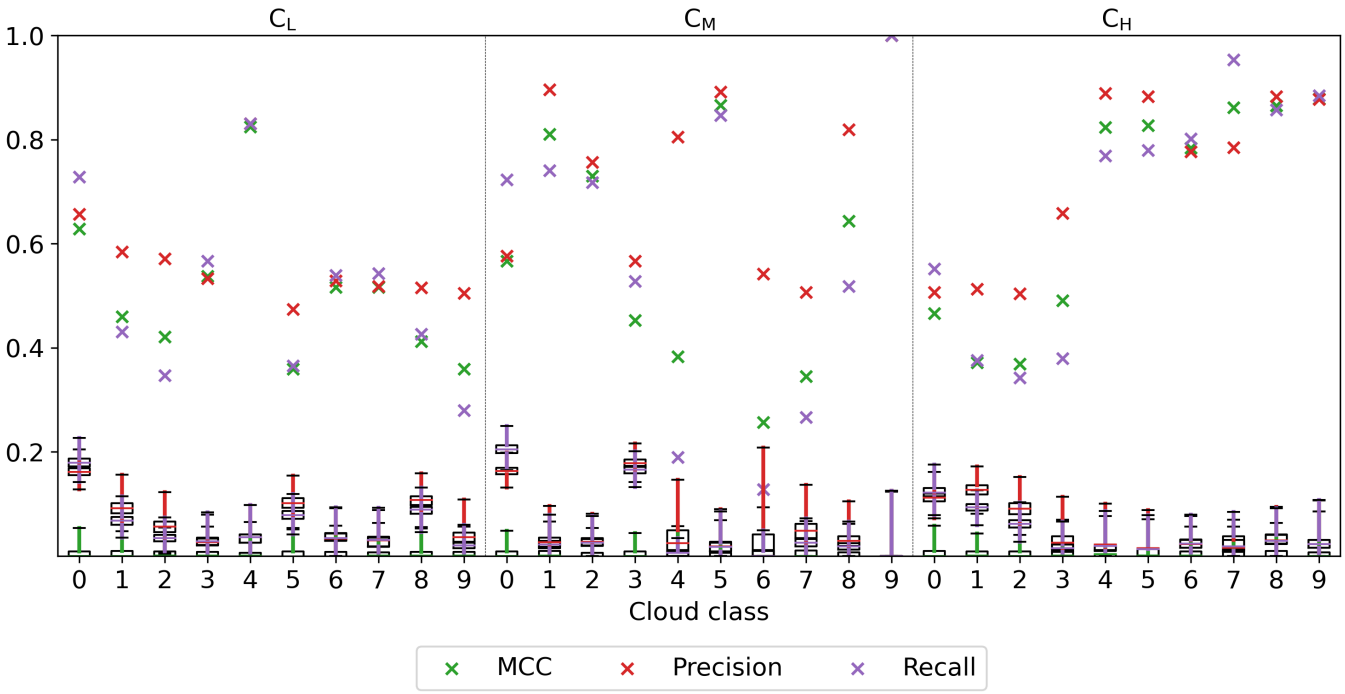


Figure 6: Matthews Correlation Coefficient (MCC), Precision, and Recall for each cloud class compared to bootstrap results from random allocation of instances to the confusion matrix.

Fig. 6 shows the most prominent and important classification metrics, namely Precision, Recall, and Matthews Correlation Coefficient (MCC) for each of the 30 cloud classes. The prior two metrics can take values between 0 and 1, while MCC is bounded by -1 and 1. For each metric, +1 indicates the perfect forecast. Boxplots for each class show distributions of bootstrap experiments with random allocation of instances to the confusion matrix. Fig. 6 shows that measured values of all 3 metrics are well outside bootstrap distributions of the respective metric in each class, clearly indicating that our model outperforms random forecasts.

References

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

Heydarian, M., Doyle, T. E., and Samavi, R. (2022). Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.