TRANSFER LEARNING WITH CONVOLUTIONAL NETWORKS FOR ATMOSPHERIC PARAMETER RETRIEVAL

David Malmgren-Hansen[1], *Valero Laparra*[2], *Allan Aasbjerg Nielsen*[1], *Gustau Camps-Valls*[2]

[1]Technical University of Denmark, Lyngby, Denmark
[2]Image Processing Lab (IPL), Universitat de València, València, Spain

ABSTRACT

The Infrared Atmospheric Sounding Interferometer (IASI) on board the MetOp satellite series provides important measurements for Numerical Weather Prediction (NWP). Retrieving accurate atmospheric parameters from the raw data provided by IASI is a large challenge, but necessary in order to use the data in NWP models. Statistical models performance is compromised because of the extremely high spectral dimensionality and the high number of variables to be predicted simultaneously across the atmospheric column. All this poses a challenge for selecting and studying optimal models and processing schemes. Earlier work has shown non-linear models such as kernel methods and neural networks perform well on this task, but both schemes are computationally heavy on large quantities of data. Kernel methods do not scale well with the number of training data, and neural networks require setting critical hyperparameters. In this work we follow an alternative pathway: we study *transfer learning* in convolutional neural nets (CNNs) to alleviate the retraining cost by departing from proxy solutions (either features or networks) obtained from previously trained models for related variables. We show how features extracted from the IASI data by a CNN trained to predict a physical variable can be used as inputs to another statistical method designed to predict a different physical variable at low altitude. In addition, the learned parameters can be transferred to another CNN model and obtain results equivalent to those obtained when using a CNN trained from scratch requiring only fine tuning.

Index Terms— Transfer Learning, Convolutional Neural networks, Infrared measurements, parameter retrieval

1. INTRODUCTION

Predicting atmospheric variables from satellite measurements is a key point in Earth observation. While Numerical Weather Prediction (NWP) methods are based on well tested physical models, sometimes the computational load makes their use prohibitive. Statistically-based methods can help either to substitute the NWP methods in some applications, or to provide predictions that can be used as a first guess for the NWP methods.

Atmospheric profiles of physical variables are important for weather forecasting and atmospheric chemistry studies. However predicting them is one of the most challenging tasks due to their high dimensionality. On the other hand measurements from high spectral resolution instruments, like the Infrared Atmospheric Sounding Interferometer (IASI) sensor implemented on the MetOp satellite series, provide useful information for the estimation of these profiles [1, 2]. Designing efficient statistically-based algorithms for this task is challenging not only because the high dimensionality of the input but also the high dimensionality of the output space. Both show high correlation and structure that should be exploited.

Previous approaches based on kernel methods and neural networks actually exploited either spectral or vertical correlations [3–6], but spatial redundancy was not explicitly included in the models. In this way, convolutional neural networks (CNNs) offer a good alternative to exploit relations in all three dimensions jointly [7]. Neural networks offer an additional advantage to our multivariate regression problem: models are intrinsically multi-output and account for the cross-relations between the state vector at different altitudes. This allows us to attain smoothness, and hence consistency, across the atmospheric column in a very straightforward way.

Nevertheless, estimating multiple physical variables simultaneously is complicated. Multiple models for different variables are necessary and they are heavy to train. In this sense, tools from *transfer learning* could be very useful in this setting, which we evaluate in this work. Here we investigate the use of two simple strategies of transfer learning which have been very successful in computer vision problems [8]. The basic idea is to reduce the cost of retraining neural nets by departing from proxy solutions obtained from previously trained models for related variables.

The remainder of the paper is organized as follows. In §2 we introduce the methodology conducted in the experimental section. Section §3 gives empirical evidence in terms of accuracy, bias and smoothness of the estimates. Conclusions and future developments are given in §4.

2. METHODOLOGY

The experiments conducted here are based on the L2 processing presented in [4]. This scheme was proposed to predict physical atmospheric parameter profiles from the IASI data using statistically-based algorithms. We used data collected in 13 consecutive orbits within the same day, 17-08-2013. Each orbit consists of approximately $92,000$ samples. We use the first 7 orbits to train and the latter 6 for testing. Note

This work was partly supported by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423).

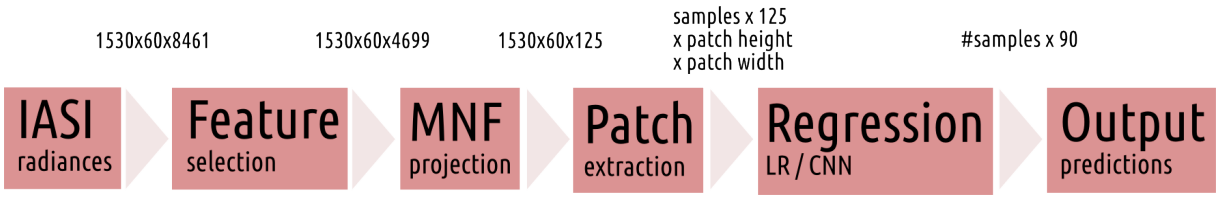

Fig. 1: Pipeline schematic: first selecting 4699 channels from the original 8461 spectral according to noise specifications, then performing a *spectral* dimensionality reduction based on linear projection to 125 features, and then a nonlinear regression based on statistically-based algorithms. Here we use linear regression (LR) and convolutional neural networks (CNN) for retrieval of atmospheric parameters at 90 altitudes simultaneously.

that $6-7$ orbits cover a great part of the Earth and the radiance signal in our dataset includes variances from different geographical locations.

In [9] it was shown how increasing the number of spatial neighboring pixels largely improves the prediction performance when included as additional input features to standard regression models. In [7] we analyzed different Ordinary Least Squares Linear Regression (OLS) and CNN configurations to predict atmospheric temperature profiles. Table 1 shows the summary of results when using different sizes of spatial neighborhoods. It can be seen that the non-linear properties of the CNN further improves performance.

Table 1: Summary of the averaged RMSE [K] across the atmosphere profile for temperature prediction.

Patch Size	1×1	3×3	5×5	7×7	10×10	15×15	25×25
CNN	–	2.48	–	–	2.43	2.20	2.28
OLS	3.30	3.00	2.91	2.86	2.84	2.85	–

3. EXPERIMENTAL RESULTS

Despite the benefits of using CNNs for the physical variable prediction, the training process is rather expensive. The concept of transfer learning within deep learning has proven useful to overcome this problem for a range of computer vision tasks [8]. For example, deep CNNs trained on large databases of natural images can be transferred to smaller datasets for specific applications with high end performance. In this work, we analyze two simple yet useful strategies for transfer learning. One strategy uses a CNN trained to predict a particular physical variable as a feature extraction method. The other strategy uses the CNN parameters of the trained network as initial parameters when training a CNN to predict a different physical variable.

3.1. Feature extraction

Here we explore the ability of exploiting the most successful model from previous section that was trained to predict temperature profiles to design a method to predict dew point temperature (DT) profiles. The straight forward way to do this uses the CNN trained to predict temperatures (*CNN-T*) as a feature extraction algorithm and train a linear model to predict DT using these features as inputs. While extremely simple, this strategy has become very popular in computer vision due to its good performance [8].

Note that the last layer of the *CNN-T* model is a fully connected layer without rectifier, i.e. a linear combination of the outputs from the previous layer. Here we use the outputs of the second last layer (just before the last linear combination) of *CNN-T* as a feature extractor, and put a linear predictor on top of it. We use this new model to predict DT, and name it *CNN-DT-F(T)*.

Results obtained with such approach are presented in Fig. 2. It is interesting to compare the proposed strategy, *CNN-DT-F(T)*, with a linear regression model (*OLS*) since both present the same training complexity. While the proposed strategy has been previously used successfully in several works in the computer vision community, it is clear that in this case it is not efficient for all the pressure levels. In particular between [300-700] hPa the simple *OLS* outperforms its performance. However for high pressure levels, the *CNN-DT-F(T)* model obtains better performance than *OLS*, and achieves almost the same accuracy as a CNN trained from scratch *CNN-DT*.

3.2. Parameter initialization

Although extremely simple, the previous approach has the problem of not having the warranty of obtaining an performance similar to the one obtained when using a CNN trained from scratch. Here we analyze a different strategy to use the information contained in the trained models when training new ones. We used the same network architecture as for predicting temperatures to design CNN models that predict DT. We analyze the effect of initializing the parameters for the training procedure with the ones of *CNN-T*, and using the ones from a new CNN trained to predict ozone concentration [ppm] *CNN-O*. Results for the *CNN-O* network compared with a linear regression model can be seen in Fig. 3. This network was trained from scratch and obtained a good performance to predict ozone: RMSE [ppm] is smaller than the OLS and in the range to the ones presented in [4] with little extra effort.

In order to compare the results, we trained a CNN from scratch using randomly initialized parameters. Therefore, we trained three new CNN models, all of them optimized to predict DT but with different parameters initialization: (1) initializing the parameters randomly, *CNN-DP-P(R)*, (2) initializing with the *CNN-T* model parameters, *CNN-DP-P(T)*, and

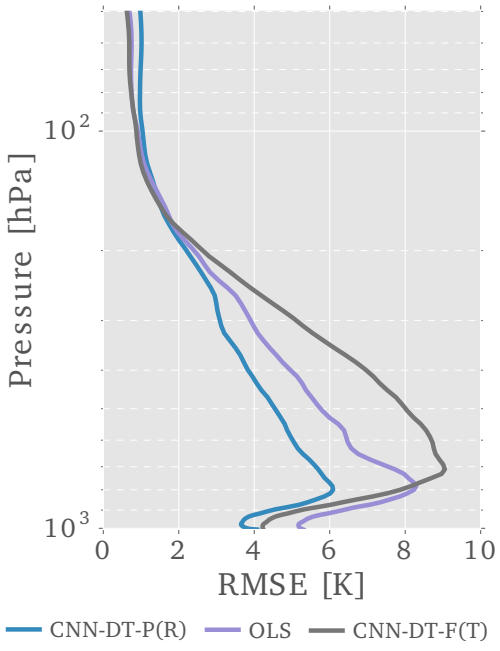

Fig. 2: Feature extraction approach. Dew point temperature RMSE [K] profiles for different models.

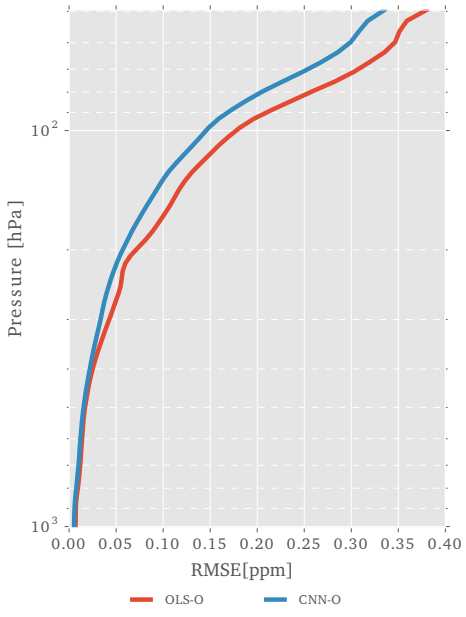

Fig. 3: Ozone RMSE [K] profiles of CNN and OLS regression models.

(3) initializing with the *CNN-O* model parameters, *CNN-DP-P(O)*.

Figure 4 shows the results for the different networks. Several conclusions can be extracted. On the one hand, results are much better with this strategy than with the strategy used in the previous section. It is clear how the *CNN-DT-P(T)* and *CNN-DT-P(O)* models achieve similar performance than the

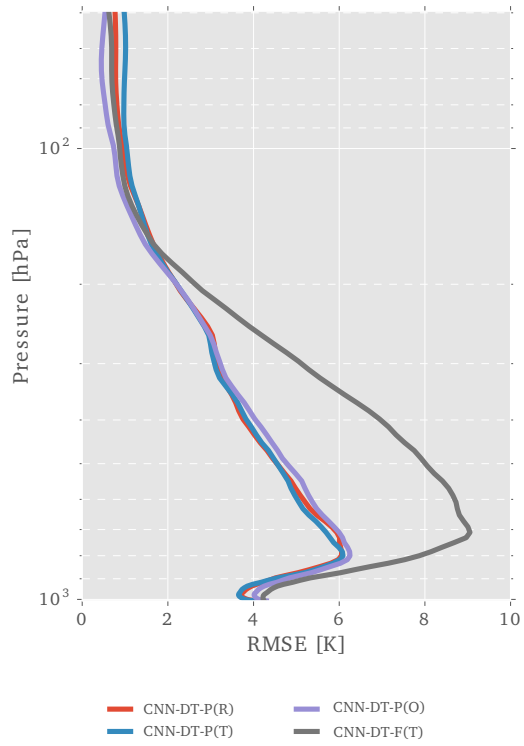

Fig. 4: Parameter initialization approach: RMSE profiles for DT of different regression models based on transfer learning.

CNN-DT-P(R) model. This gives us the idea that, although the parameter space is huge, either the solutions reach the same global minimum, or the different local minima with similar performance. Both options are equivalent for practical purposes though. This is particularly interesting in the case of ozone concentration, since it has a very different nature than DT.

Figure 5 shows specifically how the objective function is minimized during the training procedure and should be used to see the convergence of each model. Regard the convergence speed, the *CNN-DT-P(T)* model is much faster than the other two. However, note that *CNN-DT-P(O)* converges to its minimum in a similar way as when initializing the parameters randomly *CNN-DT-P(R)*.

Although the predictions from the three different models show a similar performance in RMSE terms, solutions can fall into completely different local minimum. An interesting property of the model is the consistency of the predictions with regard other physical variables. Figure 6 shows the cross-correlation matrices between the temperature predictions and the DT predictions obtained using either the *CNN-DT-P(R)*, or the *CNN-DT-P(T)* model. The DT predictions of the model *CNN-DT-P(T)* model are similarly aligned with the temperature predictions (the cross-covariance matrix symmetry is similar) than the ones from the *CNN-DT-P(R)* model. The non-diagonality (ND) has been computed as the Frobenius norm of the matrix minus its transpose, i.e. ND=$\|\mathbf{A} - \mathbf{A}^\top\|_F^2$. Surprisingly, the mutual information between the variables is smaller for the *CNN-DT-P(T)* model.

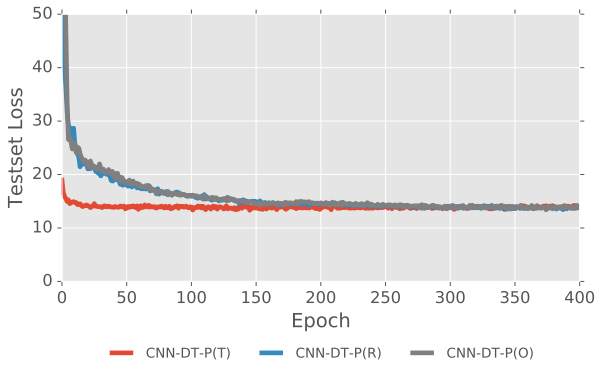

Fig. 5: Objective minimization during the training procedure. Performance of CNN models trained to predict DT initialized using different sets of parameters.

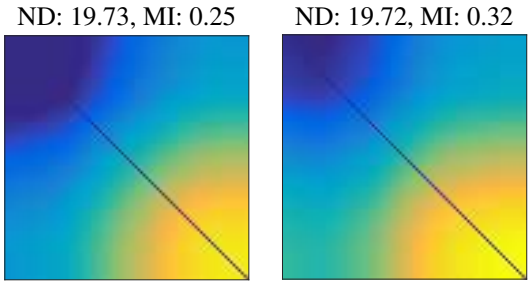

Fig. 6: Correlation matrices (yellow means higher) between the predicted DT using *CNN-DT-P(T)* and T using *CNN-T* (left); and between the predicted DT using *CNN-W-P(R)* and T using *CNN-T* (right).

This is a safety check to ensure that, even that the model has been initialized with the parameters of a model trained to predict other physical variable, the predictions do not need to be dependent of the behavior of this model. Mutual information has been computed in a similar way as in [9] using the method introduced in [10].

4. CONCLUSIONS

We analyzed the problem of transfer learning in multi-output physical parameter retrieval when using CNN models. In general we found that some benefits can be obtained from the transfer learning methodology. We analyzed two different strategies. Firstly, we used a model trained for predicting temperature profiles as a feature extraction method for the previous stage to a simple linear regression algorithm. We found that this strategy is not ideal but it can be helpful in some aspects. At low altitudes, we get higher accuracy than the shallow linear regression model. At higher altitudes though, the transfer learning approach does it worse. Fine tuning for a specific output variable is necessary in order to achieve good predictions.

The second strategy was using the parameters of an already trained model as initial parameters to train a new model to predict a different physical variable, in our case dew point temperature (DT). We initialized the model using the parameters of two already trained models: one trained to predict temperature and one to predict ozone. We compared the performance also with a model trained when using randomly initialized parameters. We found that initializing the parameters using an already trained model helps in time convergence terms and achieves a similar result as training the model from scratch when the model used as initialization was trained to predict a similar variable. The performance reached by CNN initialized from random weights can be reached in less than around $\frac{1}{8}$ of the training time if the weights are transferred (initialized) from the models trained for temperature, while the one initialized with the parameters of the ozone model provides no advantage with regard the random initialization. Note that features of T and DT are more similar in structure and even in units with regard to ozone.

5. REFERENCES

[1] EUMETSAT, *IASI Level 1: Product Guide*, 2014, EUM/OPS-EPS/MAN/04/0032,.

[2] B. Tournier, D. Blumstein, F. Cayla, , and G. Chalon, "IASI level 0 and 1 processing algorithms description," in *Proc. of ISTCXII Conference*, 2002.

[3] W.J. Blackwell, "A neural-network technique for the retrieval of atmospheric temperature and moisture profiles from high spectral resolution sounding data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 11, pp. 2535–2546, Nov 2005.

[4] G. Camps-Valls, J. Munoz-Mari, L. Gomez-Chova, L. Guanter, and X. Calbet, "Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1759–1769, 2012.

[5] G. Camps-Valls, J. Verrelst, and Muñoz-Marí, "A survey on gaussian processes for earth observation data analysis: A comprehensive investigation," *IEEE Geoscience and Remote Sensing Magazine*, , no. 6, June 2016.

[6] V. Laparra, J. Muñoz-Marí, L. Gómez-Chova, X. Calbet, and G. Camps-Valls, "Nonlinear statistical retrieval of surface emissivity from iasi data," in *IEEE International and Remote Sensing Symposium (IGARSS)*, 2017.

[7] D. Malmgren-Hansen, V. Laparra, A. Aasbjerg Nielsen, and G. Camps-Valls, "Statistical retrieval of atmospheric profiles with deep convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing (Submitted)*, vol. 56, 2018.

[8] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[9] David Malmgren-Hansen, Valero Laparra, Allan Aasbjerg Nielsen, and Gustau Camps-Valls, "Spatial noise-aware temperature retrieval from infrared sounder data," *IEEE International Geoscience and Remote Sensing Symposium*, 2017.

[10] V. Laparra, G. Camps-Valls, and J. Malo, "Iterative gaussianization: From ICA to random rotations," *IEEE Transactions on Neural Networks*, vol. 22, no. 4, pp. 537–549, 2011.