



Adaptation of state-of-the-art neural network architectures to interference fringe reduction in absorption spectroscopy

Lenard L. Röder¹

Received: 18 April 2023 / Accepted: 17 April 2024 / Published online: 9 May 2024
© The Author(s) 2024

Abstract

State-of-the-art neural network architectures in image classification and natural language processing were applied to interference fringe reduction in absorption spectroscopy by interpreting the data structure accordingly. A model was designed for temporal interpolation of background spectra and a different model was created for gas concentration fitting. The networks were trained on experimental data provided by a wavelength modulation spectroscopy instrument and the best performing architectures were analyzed further to evaluate generalization performance, robustness and transferability. A BERT-styled fitter achieved the best performance on the validation set and reduced the mean squared error of fitted amplitude by 99.5%. However, analysis of the de-noising behavior showed large biases. A U-Net styled convolutional neural network reduced the mean squared error of the interpolation by 93.2%. Evaluation on a test set provided evidence that the combination of model interpolation and linear fitting was robust and the detection limit was improved by 52.4%. Transferring the trained interpolator model to a different spectrometer setup showed no chaotic out-of-distribution effects. Additional fine-tuning further increased the performance. Neural network architectures cannot be generally applied to all absorption spectroscopy tasks. However, given the right task and the data representation, robust performance increase is achievable.

1 Introduction

Artificial neural networks and deep learning have contributed to major breakthroughs in several applications like image classification, segmentation, generation of images and text, natural language processing (NLP) and many more. These new frameworks outperform conventional machine learning algorithms in many disciplines [1–5].

However, they cannot be trivially applied to regression tasks in natural sciences. In contrast to most conventional data analysis methods, hyperparameter optimization often needs complete retraining of the neural network and is therefore associated with high computational effort. Error estimation of the output is a research field on its own [6] and the model prediction can react chaotically to tiny deviations of the input [7]. Deep learning applications often differ from regression tasks with respect to their goal, constraints and requirements [2].

Nonetheless, several studies have successfully applied neural networks to spectroscopy tasks and reported performance increases compared to conventional approaches. Applications feature classification of spectroscopic data obtained from Raman spectroscopy and other spectroscopic techniques [8, 9], speeding up expensive calculations via surrogate models for nonlinear tomography [10, 11] and spectrum prediction [12], and concentration estimation [13–15]. Nicely et al. [13] used a shallow neural network for fringe reduction in direct absorption spectroscopy using simulated data. Tian et al. [15] report a good linearity of a direct fit performed by a neural network for high SNR input spectra. In our recent study we observed good performance of a neural network based noise reduction scheme for a specific noise structure where all tested conventional methods fell short [16].

This study will focus on absorption spectra obtained via wavelength modulation spectroscopy but the main considerations should be transferable to direct detection methods or other data acquisition schemes, as well. The main noise sources of wavelength modulation spectroscopy instruments can be split into relative and absolute contributions. The relative contributions cause disturbances proportional to the concentration of the measured species in the cell, e. g.

✉ Lenard L. Röder
lenard.roeder@mpic.de

¹ Department of Atmospheric Chemistry, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany

variations in pressure, temperature, laser power and detector sensitivity. The main absolute limitation is often caused by notorious etalon fringe patterns that emerge from reflective surfaces of the optics [17, 18]. Other possible noise sources can be interference with absorption of other molecules, laser and detector stability or signal processing electronics. This study will, however, focus mainly on the reduction of noise resulting from interference fringe patterns. The procedures described may also be able to remove other noise types, as long as the requirements in Sect. 2.3 are fulfilled.

1.1 Neural networks overview

The basic architecture of an artificial neural network (ANN) consists of iterated layers y_k of linear transformations implemented via matrix multiplications W_k and a following non-linearity σ [2]:

$$y_{k+1} = \sigma(W_k y_k) \quad (1)$$

Theoretically this architecture can approximate any function using only a single intermediate ("hidden") layer [19–21]. Advanced architectures try to optimize the ability of the model to learn and to generalize while being very efficient in time and memory.

A first breakthrough in image recognition tasks was achieved via convolutional neural networks (CNN) [22]. The matrix multiplications are replaced with convolutions of small filters whose weights are shared between all positions in the image [23]. This operation also ensures translation equivariance of the model output with regards to the input image [2]. Together with normalization schemes to overcome internal covariate shift [24], shortcut paths to decrease gradient decay [25] and randomization patterns to lower the chance of overfitting [26], models were constructed that outperformed human predictions for image classification tasks.

Vaswani et al. [27] introduced the transformer, a new architecture for NLP that builds on a mechanism called attention, where a query sequence gates the input sequence to focus on important parts. This method led to other similar architectures that have become state-of-the-art in NLP, such as the bidirectional encoder representation from transformers (BERT) [5]. Others adjusted the transformer-based architectures for image classification and achieved comparable results to state-of-the-art convolutional architectures [28, 29].

The field of deep learning is fast evolving and state-of-the-art candidates are constantly changing. Researchers seeking to apply deep learning to their discipline may not need to design whole neural network topologies themselves, but rather adjust already tested state-of-the-art networks to their need. Results may benefit from focusing on finding the appropriate already available state-of-the-art model. A

drawback of this approach, though, is the vast size of state-of-the-art models.

1.2 Opportunities for neural networks

These fringe patterns can interfere with the frequency region of the signal. In our previous study we reported different behavior and performance of several noise reduction methods depending on this interference. Many conventional methods are based on frequency separation of signal and background and therefore only show great performance if this interference is small. Otherwise the method will produce a high bias. We reported, however, an improved performance of a neural network based approach in a region of high interference [16].

This behavior can be motivated in an example: Suppose an absorption spectrum is obstructed by a pure sinusoidal background modulation and some white noise. When fitted, the resulting spectrum will show high variations depending on the phase of the sine background. Conventional numerical noise reduction schemes that act on the frequency domain will not be able to reconstruct the original signal but will most likely dampen the result towards zero. However, if the structure of the background is known *a priori*, noise reduction is very efficient. In this example the structure of the background can be extracted from the edge of the spectrum by a curve fit and the obtained sine can be subtracted from the noised spectrum. Of course, such a simple example of background can simply be removed by applying established experimental techniques, e.g., lock-in amplification. However, the background structure in a real experiment is often more complex. It cannot be easily removed via experimental techniques and the fitting parameters cannot be easily fit. A machine learning algorithm, though, could learn the distribution of background structures and infer the interference with the signal. This approach requires the background structure to have a sparse distribution and be stable over the whole absorption spectrum.

Another conventional approach to reduce the impact of etalon fringe patterns is regular determination of background spectra and an interpolation scheme [30, 31]. If the background structure changes point-wise and slow enough, this approach will give precise estimates of the underlying background structure of a measured spectrum. The interpolated spectrum can be subtracted from the measurement and will yield great noise reduction. In a second example a problem of this approach will be discussed: Assume the background structure does not change point-wise but along the frequency axis. If e.g. the phase changes in the order of half a cycle between two background measurements, the result of the interpolation will vastly differ from the true background structure. Given *a priori* knowledge about the speed of the phase shift, the background can be reconstructed again with

high precision. In practice obtaining this *a priori* knowledge can be very hard as a realistic background has a much more complex structure and could consist of several phase changing periodic structures or beat interferences. Also the true shape of the fringe pattern could be obscured by aliasing effects [32]. Again a machine learning algorithm could learn the distribution of these structures and yield a better interpolation scheme to reconstruct the background.

In Sect. 2 possible interpretations of the data in order to motivate appropriate model architectures are discussed. In Sect. 3 the spectrometer is described in more detail and the training data acquisition and processing is presented. In Sect. 4 the performances are discussed for the training process, validation set evaluation, test set evaluation and further transfer applications.

2 Neural network architectures

As motivated in the introduction, two different machine learning objectives will be considered: concentration fitting (2) and temporal interpolation (3), with spectrum size N , number of spectra for each measurement C and sequence interval T . The fitter F directly fits an absorption spectrum to obtain the trace gas concentration. The interpolator I gets a temporal sequence that contains regular background measurements and reconstructs the evolution of the background in the inbetween region.

$$F : Z \longrightarrow AZ \in R^{N \times C} A \in R^1 \quad (2)$$

$$I : X \longrightarrow YX \in R^{T \times N \times C} Y \in R^{T \times N \times C} \quad (3)$$

2.1 Possible fitting architectures

Although the input to the fitting task is one-dimensional (disregarding the feature axis) increasing the dimension to two by a redundant linear transformation might be beneficial. This enables the application of models that have proven themselves on popular image classification benchmarks. In this study the continuous wavelet transform (CWT) with the DOG2 wavelet is used to scale up the spectra to $R^{N \times J \times C}$ with $J = 128$. The continuous wavelet transform has been very successful in environmental sciences [33, 34], but similar behavior can be expected for other linear transformations. An example plot of a CWT-transformed spectrum can be found in Fig. 1.

Now, the data can be interpreted in different ways: a $N \times J$ image of the whole wavelet spectrum or an N -dimensional sequence of J -dimensional vectors.

If data is considered as image-like, a CNN architecture can be used. A novel family of efficient and effective

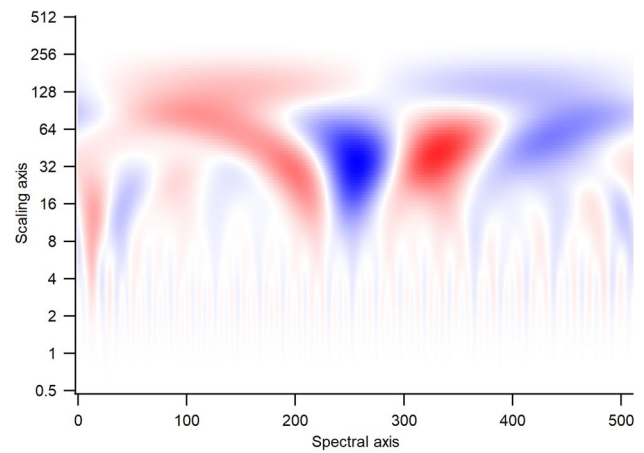


Fig. 1 Example plot of a CWT-transformed spectrum, taken from the validation set. Values smaller than zero are colored red, values larger than zero are colored blue. The plot suggests how to interpret the data as image-like. The anisotropy of the image is visible, with high frequency noise at the lower end of the scaling axis, signal at the higher end of the scaling axis and left-right-centered along the spectral axis, and obstructing fringes in the same scaling range but isotropic along the spectral axis. The SNR of the underlying spectrum is 3.2

CNNs is EfficientNetV2 [35], from which the smallest iteration EfficientNetV2B0 (EffV2) will be used for performance reasons. A CNN architecture iteratively extracts features of the image y_k and scales down using convolutional (Conv) blocks that contain pooling operations and strided convolutions.

$$y_{k+1} = \text{Conv}(y_k) \quad (4)$$

In classification tasks the final layer is globally reduced and fed through a shallow fully connected neural network (FCNN). The result is then softmaxed to retrieve the probability for each class. In this case, though, the output needs to represent the concentration of the trace gas, therefore the final output is not passed through an activation layer. This way the output can be any real number.

However, as mentioned earlier, a CNN is translation invariant with respect to its input - in image processing the algorithm should not depend on the global position of the object. In the data described there is an anisotropy: The signal is left-right-centered and occupies a distinct part of the wavelet scaling, while noise is isotropic and can appear differently throughout the scaling dimension. The vision transformer (ViT) [29] decomposes a picture of shape $R^{N \times J \times C}$ into patches of shape $R^{P \times P}$ and flattens the pixels to obtain data of shape $R^{N/P \times J/P \times P^2 \times C}$. These patches are embedded via a linear transformation. The patch embedding is then added to a learned positional encoding that represents the position of the patches with regards to the complete image. Afterwards the architecture is very

similar to BERT [5] and contains of iterated transformer (Trans) blocks containing attention layers:

$$y_{k+1} = \text{Trans}(y_k) \tag{5}$$

With the attention mechanism, the model can learn global dependencies and structure of the data due to a higher receptive field [28]. The final layer of the transformer can again be globally reduced and fed into the FCNN.

Additionally, a hybrid architecture will be considered composed of an EffV2 backbone consisting of the first 5 blocks of the model, whose output is linearly transformed and directly fed to a ViT as patch embeddings. Dosovitskiy et al. reported similar performance of both approaches and found a similarity in function of the first transformer layers to the CNN backbone. The backbone will extract local features of the image that can be processed globally by the transformer [29].

If the data is considered as sequence-like, the transformer architecture can be applied directly. In this case the data is simply reshaped to $R^{N \times (J \cdot C)}$ and fed into a BERT-styled architecture [5]. Table 1 summarizes these approaches and gives a small overview of the chosen size.

2.2 Possible interpolation architectures

The interpolator also gets 2D inputs: temporal sequences of spectra. The spectra at the beginning and the end and also at regular distances throughout the sequence are considered pure background and are point-wise interpolated to fill the intermediate temporal region. Also the spectral edges of this intermediate region is kept as no absorption signal is to be expected here. An example can be found in Fig. 2. Again the data can be interpreted image-like or sequence-like but this time with a different objective.

If data is considered as image-like, the U-Net variant can be used, that has been utilized for segmentation tasks [3] or image noise reduction [36]. The down-scaling part of the U-Net is constructed from a subset of the EffV2 and the inverse of the initial EffV2. In the inverse part every down-scaling operation is replaced by an upscaling operation to obtain outputs with the original dimension.

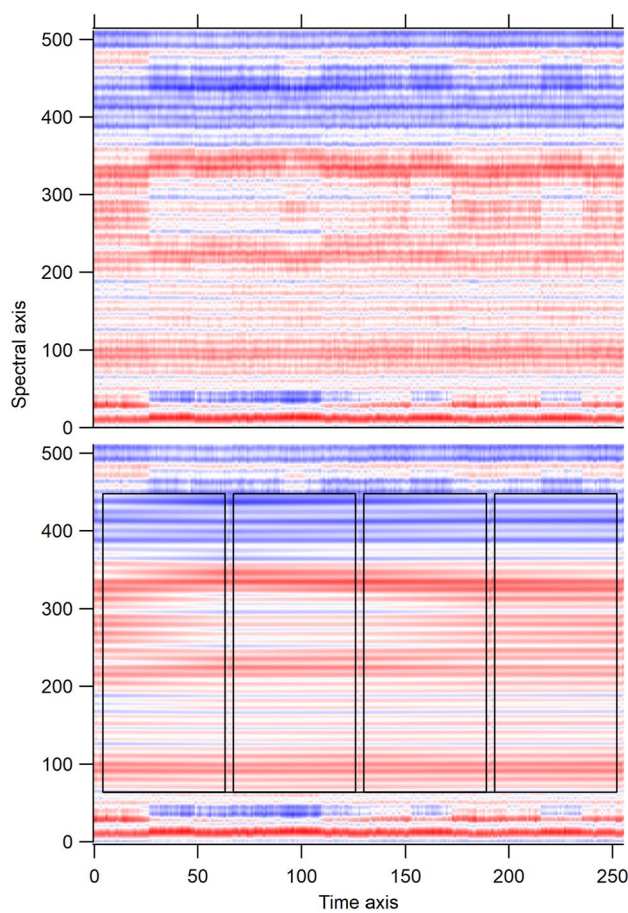


Fig. 2 Example plot of a temporal background spectrum sequence (top) and estimation via linear interpolation along the time axis (bottom). The black boxes mark interpolated regions. The temporal interpolation anchors and spectral edges are left unchanged. Values smaller than zero are colored red, values larger than zero are colored blue. The plot suggests how to interpret the task similar to image reconstruction. The objective is to reconstruct the original sequence (top) from the linear interpolation (bottom)

Another approach would be again a ViT styled segmentation followed by a transformer architecture. The final patches can be linearly transformed and concatenated back to the original shape of the picture. A hybrid architecture is not

Table 1 Summary of utilized model architectures and number of parameters

Task	Model	Data interpretation	Mechanism	Parameters
Fitter	EffV2 [35]	Image	Convolution	6.3 M
	ViT [29]	Image	Transformer	10.2 M
	Hybrid [29]	Image	Convolution/transformer	5.2 M
	BERT [5]	Sequence	Transformer	10.1 M
Interpolator	U-Net [3, 35]	Image	Convolution	11.6 M
	ViT [29]	Image	Transformer	9.9 M
	BERT [5]	Sequence	Transformer	10.1 M

considered here since the U-Net architecture already requires a lot of memory.

If data is again considered sequence-like, a linear transformation can be applied to each $N \times C$ vector. In contrast to the CWT a learnable linear transformation is used here since the dimension does not need to be increased in this case. This linear embedding can then be input to a BERT styled model. The output is again fed to a linear transformation and gets reshaped to match the desired output. In their original paper the authors introduced a masked language model (MLM) where words were randomly replaced by a *missing token* and their model was pre-trained in an unsupervised fashion to learn natural language structure [5]. In the example given here most of the data is missing, but the target output can be slowly transformed to the desired input by increasing the number of interpolated spectra in the intermediate region to speed up the initial learning period. This procedure is also applied to the image-like representations.

2.3 Remarks

In summary several ways to interpret the data have been considered and appropriate established neural network architectures have been chosen for each interpretation. It is important to emphasize here that different performances of these models do not indicate advantages of one model architecture or data interpretation over the other. No proofs or evidence of a specific model type to be favored can be given, as these are only single random examples. The performance depends very much on the choice of hyper-parameters like optimizer setup, learning rate value and schedule or number of parameters. Taking into account several ways to understand data may lead to good performance without sensitive variations of hyper-parameters and emerging biases as a consequence.

All procedures motivated in this section and in the introduction require several characteristics of the underlying background structure for the neural network approach to work accordingly. All of these requirements are fulfilled for the specific spectrometer setup used in this study with regards to interference fringe patterns as a main noise source. The procedures might show similar performance on other noise sources, if these requirements are also met there:

- **Sparse noise distribution.** All possible noise shapes must follow a sparse probability distribution compared to independent white noise. Otherwise no prior information about noise structure can be extracted from background measurements.
- **Local stability.** The noise structure needs to be locally stable so reconstruction is possible from the absorption-free parts of the data.
- **Global stability.** The noise structure needs to be stable over time, otherwise the prior information needed

for reconstruction changes and the network needs to be retrained. If the structure changes too fast, retraining is needed before desirable performance can be achieved.

3 Experimentation and network training

The experiment conducted for this study is based on the instrument TRISTAR reported in [30, 37]. It is driven by a room temperature quantum cascade laser [38] from Alpes Lasers operated near the formaldehyde (HCHO) transition at 1759.72 cm^{-1} [39]. The laser beam is guided into a 50 cm long White Cell [40] where it gets reflected 128 times to yield an effective path length of 64 m. The beam exits the cell and is split by a beam splitter into two separate paths. One beam is guided through a reference cell filled with the substance of interest at high concentration. Both beams are focused on identical infrared detectors from VIGO Systems. The laser is modulated by a slow triangle wave that scans through the absorption spectrum and a fast sine wave that is demodulated at twice frequency by the data acquisition FGPA. The resulting spectrum is similar to the second derivative of the absorption profile, depending on the modulation depth. In the experiment each data point consists of $C = 2$ spectra, one from the increasing part of the triangle wave and one from the decreasing part, with $N = 512$ points each.

Experimental data of the instrument is gathered for 14 days. The instrument inlet is connected to an air purifier to remove the substance of interest, in this case HCHO. The reference cell is filled with a high concentration of HCHO. The detector channel without reference cell detects absorption-free spectra that only consist of the interference fringe background structure. Spectra acquired at this detector will be called BGD. The detector channel with inserted reference cell can be assumed fringe-free due to the high concentration signal. Spectra acquired at this detector will be called REF. Only the last 82% of this data is used for training.

Training data for the interpolator is created using sequences of BGD. A sequence X is input to a network that gets 4 s of spectra every 63 s and does point-wise linear interpolation in the intermediate region. The interpolation anchors and 64 points at both edges of the spectrum are left unchanged. Then this array of shape $R^{T \times N \times C}$ is normalized for each subarray in C while mean and standard deviation are stored. This prepared matrix is fed through the interpolator and the output is rescaled by the standard deviation and translated according to the mean that was stored before. The loss is calculated as mean-squared-error (MSE) loss between this final matrix Y and the input sequence X .

Training data for the fitter is created using random pairs of BGD and REF. Two log-uniform distributed values A , S are drawn for the concentration and the signal-noise-ratio

(SNR). The input to the network Z is a linear combination of BGD and REF with

$$Z = A \text{ REF} + \frac{A}{S} \text{ BGD} \quad (6)$$

First this input is normalized and the standard deviation is stored. Then the CWT is performed that transforms to a dimension of $R^{N \times J \times C}$. The CWT result is fed through the fitter to obtain a single value. This value is rescaled by the standard deviation stored. The loss is calculated as MSE loss between this rescaled output value and the value A . For additional investigation two instances of fitters are trained for different SNR ranges.

For validation, sequences of BGD and pairs of BGD and REF along with pre-determined random values are taken from the first 9–18% of the measurement data. The best iteration of each interpolator or fitter, respectively, is then applied to the test set. The test set consists of a sequence of BGD and injections of calibration gas into the measurement cell from the first 9% of the measurement time. Variation and point-wise accuracy of background and calibration signal can then be determined and compared to a conventional approach.

Each network is trained on a HPC cluster hosting several NVIDIA V100 GPUs. The training process is parallelized through the distributed learning scheme Horovod [41].

4 Results and discussion

In this section the performance of the described model architectures will be discussed.

4.1 Training performance

As already mentioned in Sect. 2, the interpolator was trained on an easier task during the first 10 epochs for a more efficient training process. A random portion of spectra in the input sequence were exchanged by the corresponding spectra in the target sequence. This procedure reinterprets the MLM from BERT or an autoencoder application to enable faster learning of the underlying distribution. The amount of exchanged spectra was linearly decreased towards zero in epoch 10. All model architectures show decreasing training loss and error metrics while validation metrics do not show major indications for overfitting. Figure 3 shows losses and mean absolute error of the interpolator models. Additional plots can be found in the appendix. For further investigation

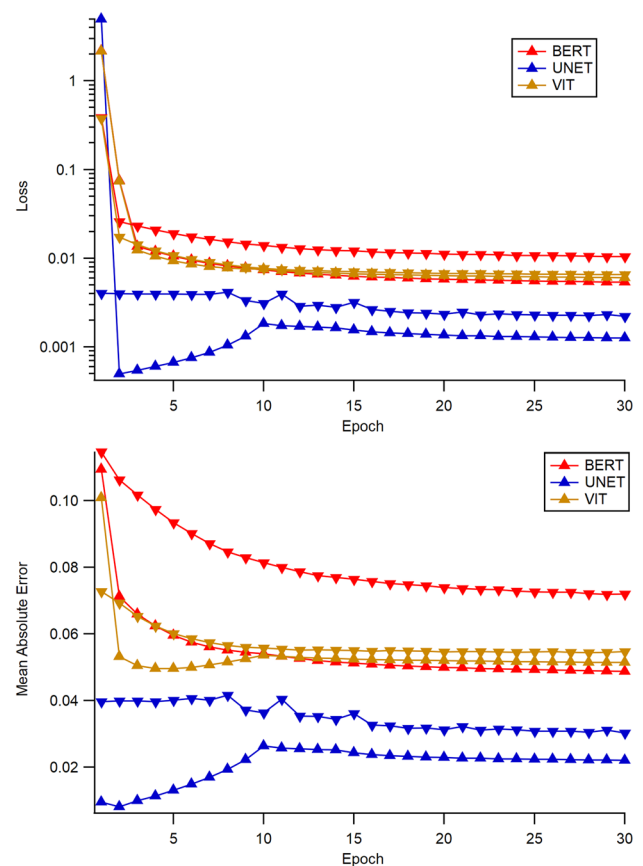


Fig. 3 Training metric (upward triangle) and validation metric (downward triangle) of interpolator models based on BERT (red), VIT (yellow) and UNET (blue) architecture: Top: Loss (mean squared error); Bottom: MAE (mean absolute error)

of the robustness, several ablation studies [29] were performed where the number of parameters of each model was reduced by scaling down the feature axes. The scaled down models achieved faster convergence speeds while resulting in similar final metrics. Example plots can also be found in the appendix.

4.2 Validation set performance

Although different performances during training might suggest the most suitable model choice for each application, these metrics only provide guidance during the training process. Model performance should always be derived from experiments that more closely reassemble real applications. Thus, the models are evaluated in more detail using the validation sets.

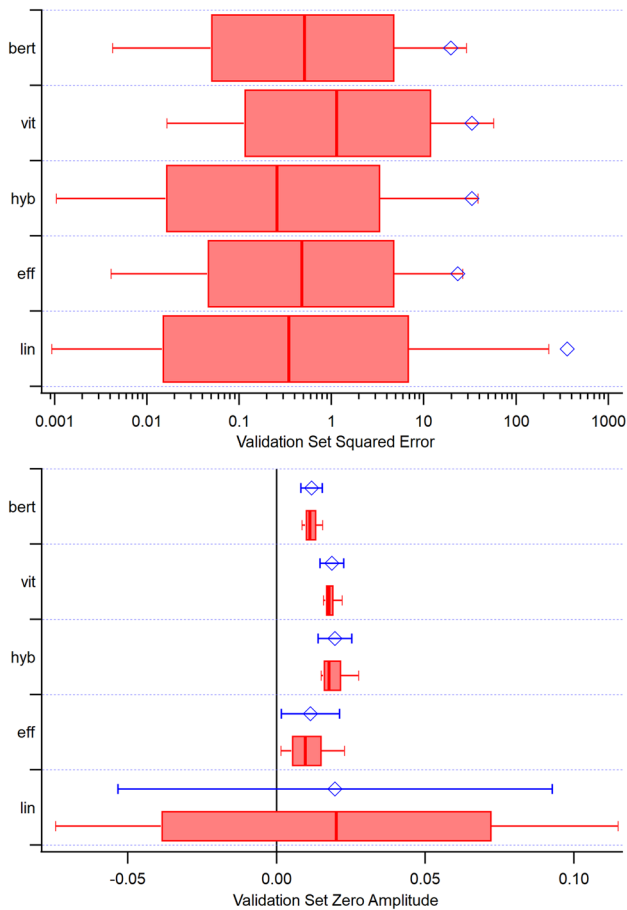


Fig. 4 Evaluation of fitter models trained on low SNR range and comparison to linear fit. Red boxes indicate 25%-quantile, median and 75%-quantile; red whiskers indicate 10% and 90% quantile. Blue diamonds show the mean, blue whiskers (if provided) show one standard deviation. Top: absolute squared error between prediction and true value. Maxima are in the order 10^4 . Bottom: predicted amplitude of pure background spectra

The fitter model is evaluated by calculating the absolute squared error between model prediction and true signal amplitude of each spectrum in the validation set. For the investigation of small signal behavior the pure background spectra are also input to the model to obtain predicted zero amplitudes.

Fig. 4 shows an overview of the evaluation results for the fitters trained on a low SNR range. The BERT-styled variant shows the least combination of bias and variance for the zero amplitude while performing similar to the EffV2 and Hybrid variant overall. The mean squared error over the validation set is reduced by 99.5% and the mean squared error on pure background spectra by 97.4%. The absolute squared error can be further investigated depending on the input SNR.

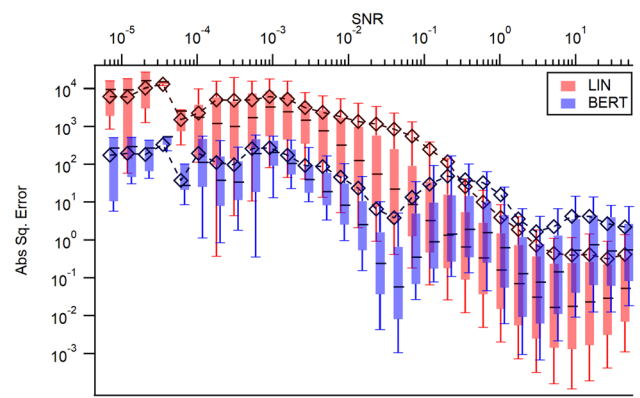


Fig. 5 Fitter squared error as a function of input SNR. Linear fit in red, low SNR trained BERT variant prediction in blue. Boxes indicate 25%-quantile, median and 75%-quantile; whiskers indicate 10% and 90% quantile. Diamonds and lines show the mean. For better visualization the BERT plot is slightly offset along the x-axis

This is depicted in Fig. 5. The BERT variant outperforms linear fitting in the low SNR regime by 1–2 orders of magnitude but falls short in the high SNR range.

The fitters trained on a high SNR range fail to outperform linear fitting and perform poorly for SNR values below their training data. A large bias is introduced when fitting pure background spectra due to the lack of low signal examples in their training distribution. Detailed plots similar to the ones shown in the low SNR example can be found in the appendix.

The interpolator models are evaluated by calculating point-wise squared errors between predicted background spectra and measured background. The evaluation result

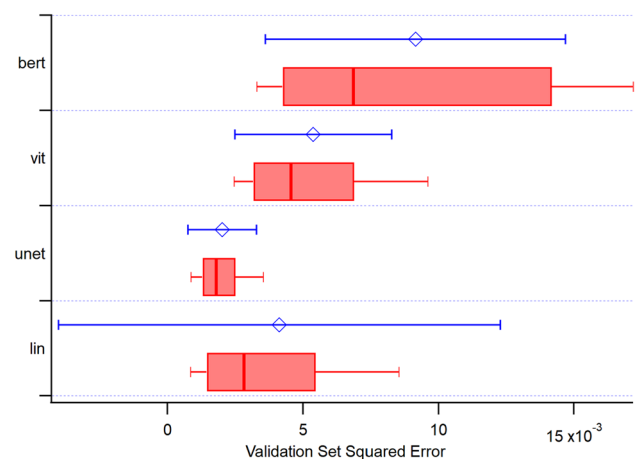


Fig. 6 Point-wise squared error between prediction and true spectra of interpolator models and comparison to linear interpolation. Red boxes indicate 25%-quantile, median and 75%-quantile; red whiskers indicate 10% and 90% quantile. Blue diamonds show the mean, blue whiskers show one standard deviation

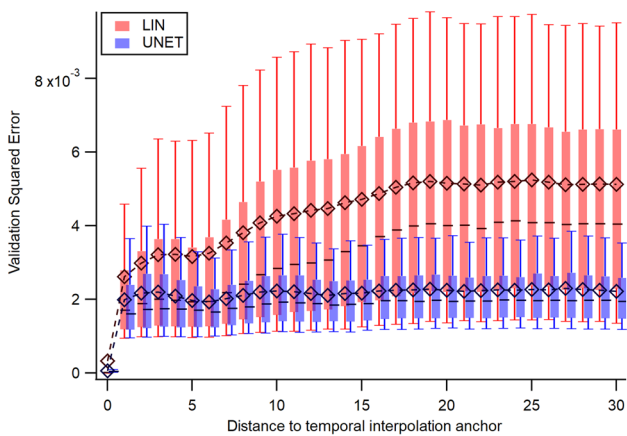


Fig. 7 Point-wise squared error as a function of distance to nearest interpolation anchor. Linear interpolation in red, U-Net variant prediction in blue. Boxes indicate 25%-quantile, median and 75%-quantile; whiskers indicate 10% and 90% quantile. Diamonds and lines show the mean. For better visualization the U-Net plot is slightly offset along the x -axis

is shown in Fig. 6. Here the U-Net variant clearly shows the best performance. The mean squared error is reduced by 93.2%. This does not necessarily indicate an advantage of this exact model architecture compared to a transformer type, since further hyper-parameter optimizations can result in increased training efficiency and model performance. The behavior of the U-Net variant is analyzed in more detail by calculating the point-wise squared errors in dependence of the distance of the spectrum from the nearest interpolation anchor. This relation can be found in Fig. 7. In the case of a stable, slowly changing background, the linear interpolation would give best estimations near the anchors, which is the case in this dataset. The model prediction shows no dependency on the distance to the nearest anchor and reduces the error evenly over the sequence.

4.3 Denoising behavior

Despite the extraordinary mean squared error reduction from application of the fitter model, the denoising behavior shows undesirable properties that introduces disadvantages compared to classical linear fitting. In the low SNR limit the variance of the linear method is very large compared to the signal amplitude. It is shown in Fig. 8 that the fitter model introduces a small bias while substantially reducing the variance. This leads to an effective reduction of the MSE. However, the bias is comparable to the signal amplitude which renders the method almost useless in applications.

A possible explanation of this behavior is a poorly chosen loss function. While the loss is reduced, the actual objective is not reached. Alternative explanations are non-optimal model and optimizing schemes or an impossible task.

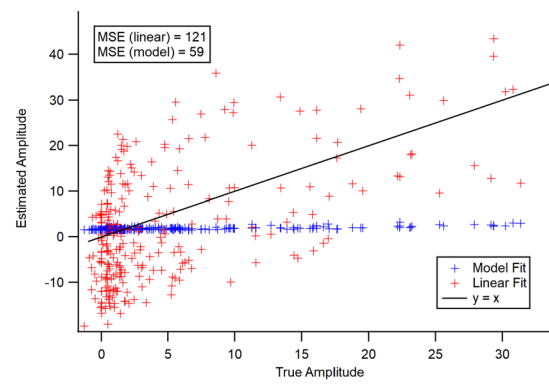


Fig. 8 Denoising behavior of the fitter model. Estimated amplitude against true value for the linear fit (red) and the low-SNR fitter model output using the BERT-styled variant (blue) for a low SNR example batch with 256 spectra. The black solid line refers to a perfect fit $y = x$. The MSE values for this example batch are 121 and 59 for the linear fit and the fitter model, respectively. Despite the improvement in MSE, the fitter model produces a large bias which makes the estimation less meaningful

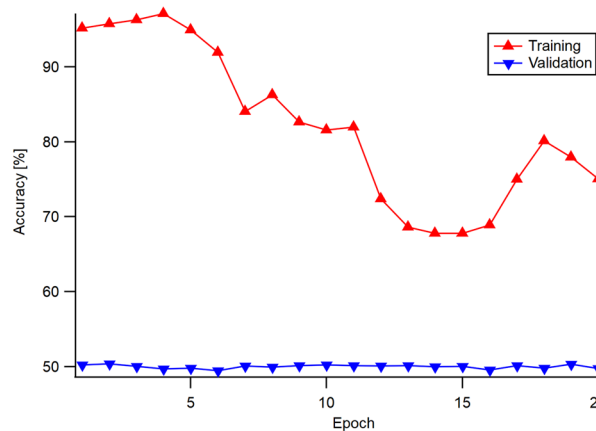
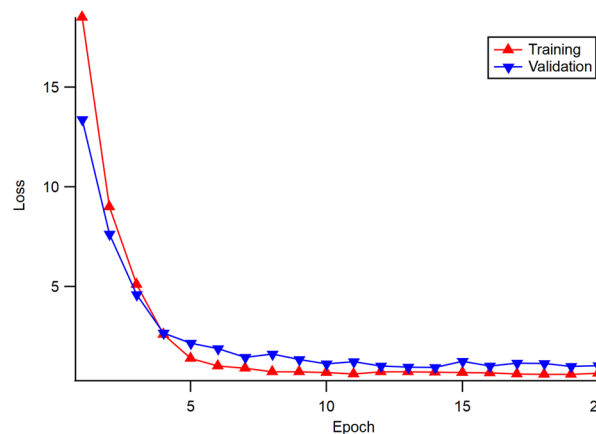


Fig. 9 Training metric (red upward triangle) and validation metric (blue downward triangle) of classifier model based on EffV2 architecture. Top: Loss (binary crossentropy); Bottom: accuracy

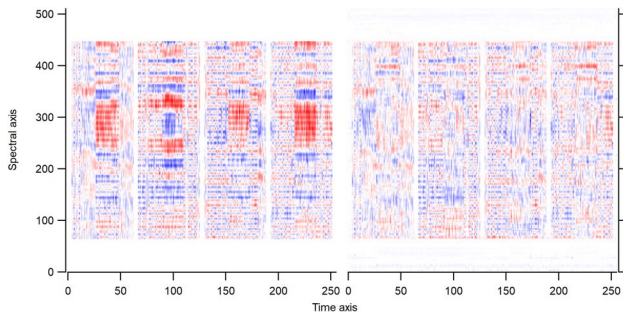


Fig. 10 Denoising behavior of the interpolator model. Difference between the example of the temporal background spectrum sequence shown in Fig. 2 and the linear interpolation along the time axis (left) and the interpolator model output using the U-Net variant (right)

For the investigation of the loss function, a variation of the experiment is conducted where instead of estimating the absolute concentration, the final layer is activated using the sigmoid function and the model is trained to distinguish pure background spectra from absorption spectra with an SNR of 0.01. Figure 9 shows the training performance of this model. The accuracy of the validation set shows no significant performance increase compared to random guessing. Thus, either the combination of model architecture and optimizing scheme is not optimal or the distinction between background and low SNR signal is not possible. The latter is the case if the data does not match the requirements given in Sect. 2.3. More specifically, the noise structure might not be locally stable or not sparse enough for the fitter task.

The interpolator shows desirable properties in the denoising behavior. In Fig. 10 the deviation from the true value of both approaches is depicted. The example from Sect. 2 is used. While coarse structures are visible in the difference between linear interpolation and original spectra, the residual structure of the model interpolation is much less correlated. This indicates effective learning of the underlying distribution and successful reconstruction.

For the rest of this chapter, only the interpolator will be considered.

4.4 Test set performance

The test set is evaluated using experimental performance metrics instead of model loss. This allows further quantification of the performance and generalization ability of each method. In absorption spectroscopy experiments, key parameters that limit the instrument performance are the

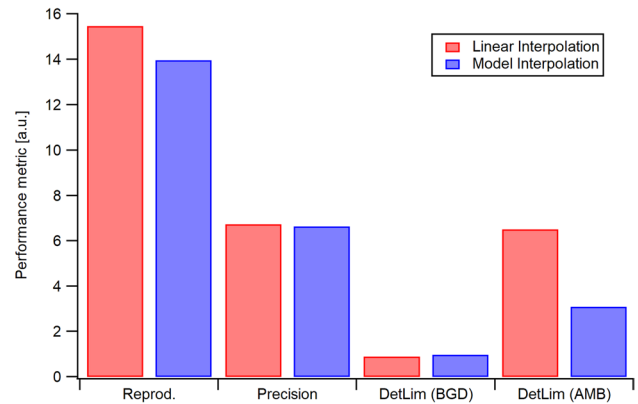


Fig. 11 Experimental performance metrics for the different interpolation techniques. Reprod: reproducibility of calibrations; DetLim (BGD): detection limit (2σ) evaluated at spectra marked background; DetLim (AMB): detection limit (2σ) evaluated at spectra marked ambient

reproducibility of calibrations due to temporal drifts, the relative precision, and the detection limit. The reproducibility of calibrations is retrieved by first averaging individual calibration intervals and calculating the standard deviation of these averages. It is a measure for long-term drifts and stability. The precision is the relative standard deviation of all calibration amplitudes after correction of the long-term drifts. The detection limit is two times the standard deviation of background amplitudes. Since only absorption-free gas was measured during the experiment, the detection limit can be evaluated using spectra marked as background as well as spectra marked as ambient.

Figure 11 gives an overview of the resulting experimental performance metrics. Similar to the validation set results the model-based approach achieve a better detection limit than the linear approach. Applying the interpolator model and a linear fit achieves a very robust reduction of detection limit as both the absolute and the relative errors are reduced.

Example time series of background and calibration amplitudes are shown in Fig. 12. Again, very robust behavior can be observed for the combination of interpolator model and linear fit. This indicates evidence that the proposed interpolator architecture provides robust noise reduction.

4.5 Transferability

An important question left unanswered is the transferability of the networks. What kind of output is created if a different spectrometer is used? Is a similar performance increase

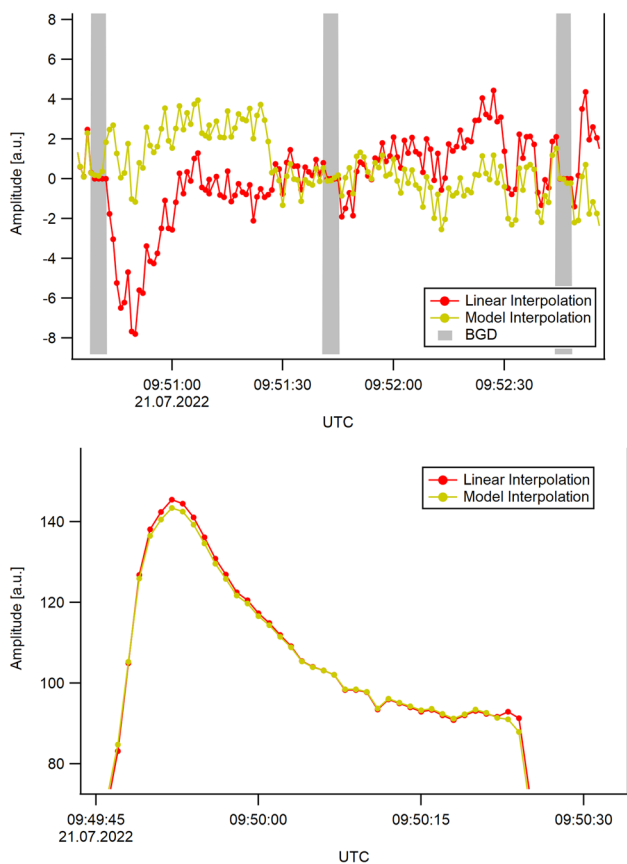


Fig. 12 Examples of amplitude results for different interpolation techniques. Top: Example result during zero gas measurement. Grey regions indicate declaration of the spectra as background. Bottom: Example result during calibration gas injection

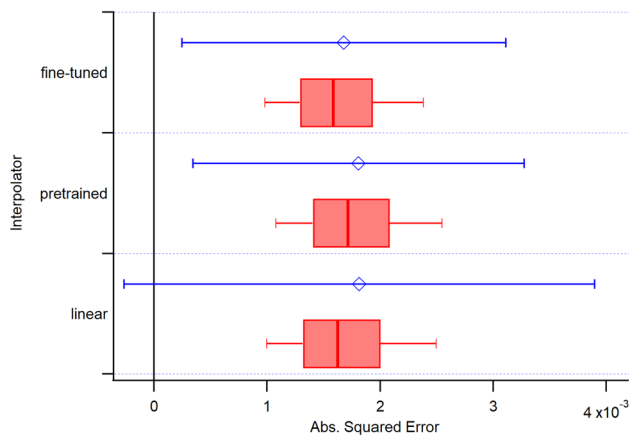


Fig. 13 Performance comparison of linear approaches to interpolator model transferred to the alternative spectrometer. Point-wise squared error between background spectra and interpolation reconstruction with linear interpolation, pre-trained variant and fine-tuned variant. Red boxes indicate 25%-quantile, median and 75%-quantile; red whiskers indicate 10% and 90% quantile. Blue diamonds show the mean, blue whiskers show one standard deviation

possible? Will the result contain chaotic artifacts due to out-of-distribution problems?

To answer these questions, the trained model is applied to a different QCL absorption spectrometer which utilizes a 20 cm long Herriott Cell configuration with 182 passes [42, 43] and is operated at the carbon monoxide (CO) transition at 2190.02 cm^{-1} [44]. Data acquisition and detectors are similar, but the experiment is driven at a different modulation and sweeping frequency.

Evaluation data for the interpolator was gathered by flooding the absorption cell with nitrogen. Zero gas was measured for 70 h. Application of the pre-trained version directly resulted in similar performance than linear interpolation. This indicates that the transferred model may require additional fine-tuning to achieve a good performance, while it does not result in chaotic artifacts. For fine-tuning, the first 7 h of nitrogen measurement were taken as new training and validation data and the pre-trained interpolator model was trained further for 10 epochs, which results in 0.8% of fine-tuning iterations compared to pre-training iterations. A comparison of the evaluation is shown in Fig. 13.

5 Summary and conclusion

In this study several possible applications of neural networks to absorption spectroscopy experiments were tested by interpreting the data structure in a way that several state-of-the-art neural network architectures can achieve good performance. These architectures were chosen from image classification and natural language processing tasks. A model for interpolation of background spectra and a model for gas concentration fitting of absorption spectra were created. Each neural network was trained on measured data and showed good generalization performance. The best performing instance of each task was further evaluated using test data and data from a different type of absorption spectrometer.

Fitters trained on a high SNR range did not outperform linear fitting. The best performing fitter trained on a low SNR range was of the BERT-type. It decreased the mean squared error on the validation set by 99.5% and the mean squared error on pure background spectra by 97.4%. However, undesirable denoising behavior was observed that rendered the method unusable. Training a classifier with the same architecture showed that this behavior was not caused by a poor choice of loss function but is caused either by insufficient architecture and optimizing scheme or an impossible objective. Considering the human-level performance in image recognition and natural language processing tasks of the chosen architectures, the objective might not be possible due to the strong interference of signal and background. However, due to the large search

space and choice of hyper-parameters this can only be speculated.

The best performing interpolator was of the U-Net type and reduced the mean squared error of the validation set by 93.2%. It showed less dependence on the distance from the nearest interpolation anchor than the linear interpolation. The combination of model interpolation and linear fitting showed very robust behavior and decreased the relative error by 8.2% and the detection limit by 52.4% on the test set.

It was shown that the interpolator model can be transferred to a different spectrometer without chaotic out-of-distribution effects. However, the performance of the pre-trained model on the different setup does not match the performance on the original spectrometer setup and may become worse than conventional approaches. The performance can be enhanced via fine-tuning on new data. Using just 0.8% fine-tuning iterations in relation to initial training iterations, the interpolator mean squared error was reduced by 36.3% compared to the conventional approach.

In conclusion, using state-of-the-art architectures is no guarantee to obtain a well performing neural network if the task is not appropriate. But, interpreting the task in multiple ways to include many state-of-the-art architectures can make the application less sensitive to specific properties of a chosen network and speed up the architecture design significantly.

In this study only $2f$ -wavelength modulation spectroscopy was considered, but the concept should also work for other absorption spectroscopy data acquisition schemes due to the similarities of the spectral features and the dominant noise sources. More fine-tuning may be required in this case.

Appendix A Additional figures

See Figs 14, 15, 16, 17, 18, 19, 20

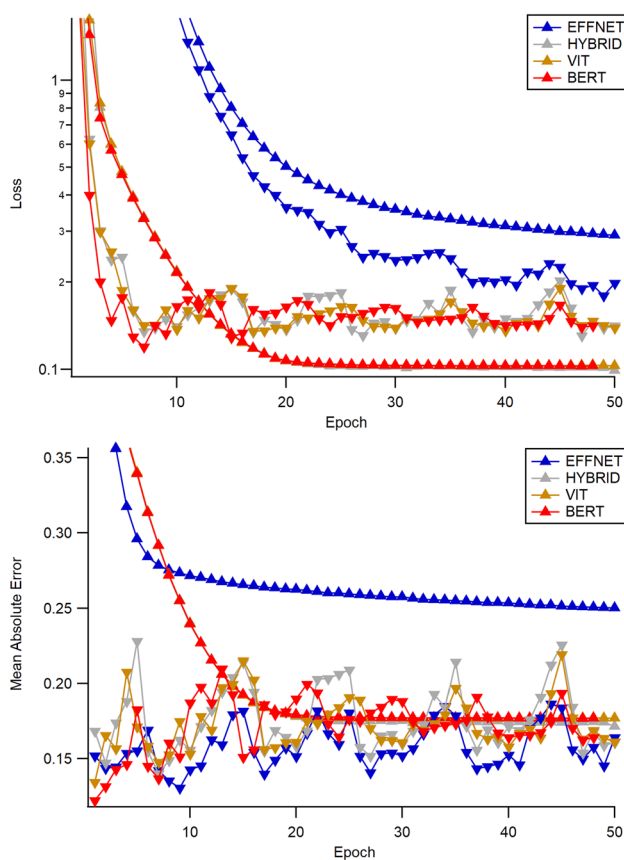


Fig. 14 Training metric (upward triangle) and validation metric (downward triangle) of high-SNR fitter models based on BERT (red), VIT (yellow), Hybrid (grey) and EffV2 (blue) architecture: Top: Loss (mean squared error); Bottom: MAE (mean absolute error)

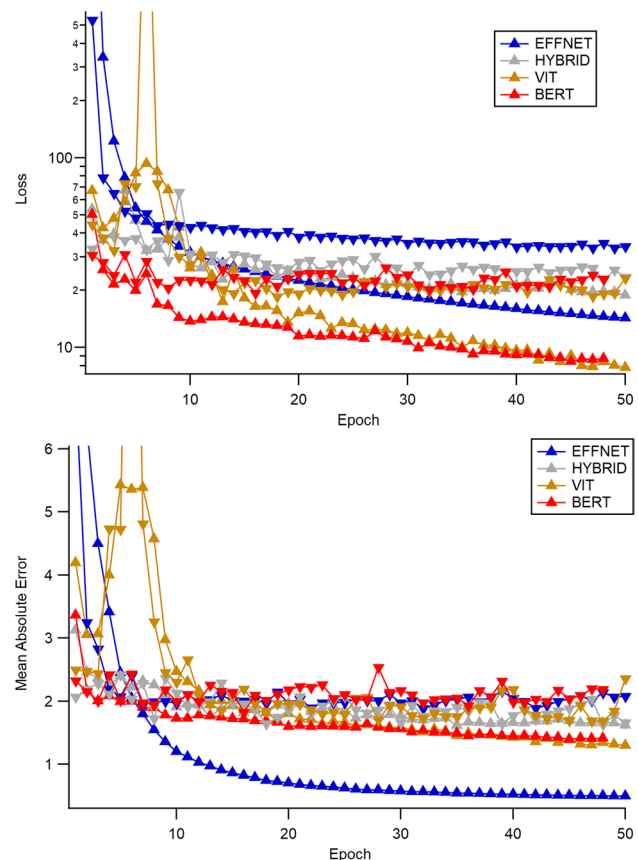


Fig. 15 Training metric (upward triangle) and validation metric (downward triangle) of low-SNR fitter models based on BERT (red), VIT (yellow), Hybrid (grey) and ffV2 (blue) architecture: Top: Loss (mean squared error); Bottom: MAE (mean absolute error)

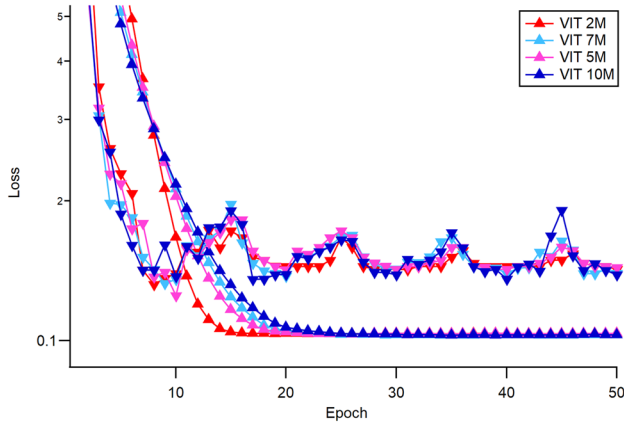


Fig. 16 Ablation study. Training loss (upward triangle) and validation loss (downward triangle) of high-SNR fitter models based on VIT. Original model size performance in blue and reduced parameter versions in light blue, pink, red in the order of smaller model sizes

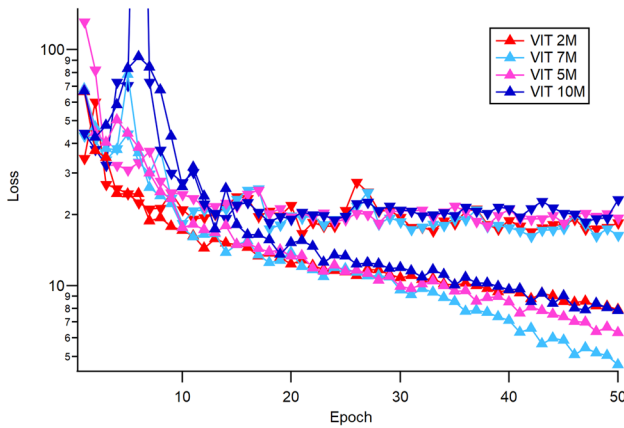


Fig. 17 Ablation study. Training loss (upward triangle) and validation loss (downward triangle) of low-SNR fitter models based on VIT. Original model size performance in blue and reduced parameter versions in light blue, pink and red in the order of smaller model sizes

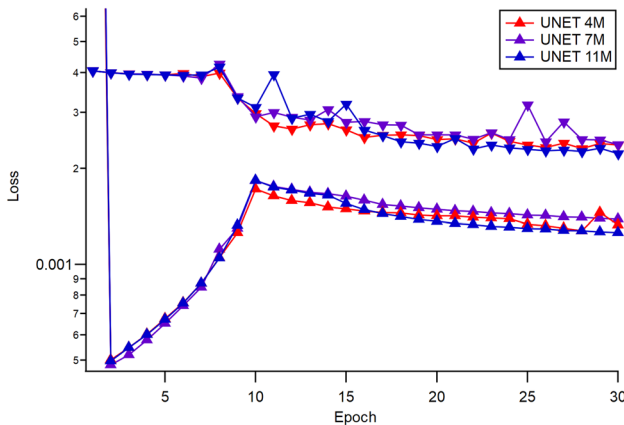


Fig. 18 Ablation study. Training loss (upward triangle) and validation loss (downward triangle) of interpolator models based on UNET. Original model size performance in blue and reduced parameter versions in purple and red in the order of smaller model sizes

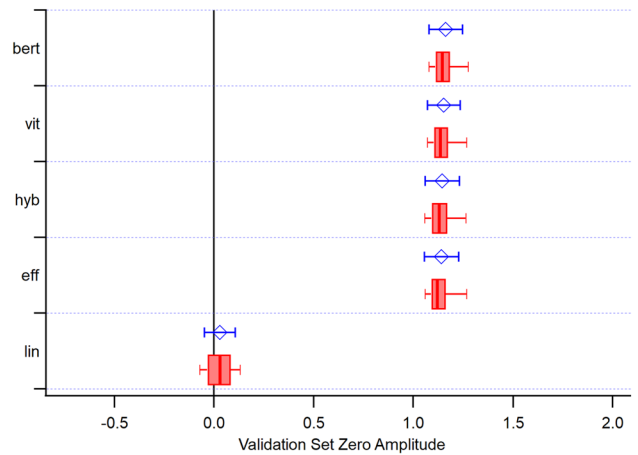
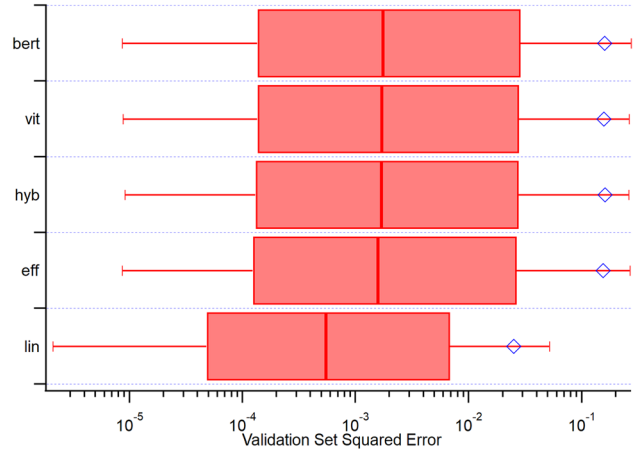


Fig. 19 Evaluation of fitter models trained on high SNR range and comparison to linear fit. Red boxes indicate 25%-quantile, median and 75%-quantile; red whiskers indicate 10% and 90% quantile. Blue diamonds show the mean, blue whiskers (if provided) show one standard deviation. Top: absolute squared error between prediction and true value. Maxima are in the order 10^1 . Bottom: predicted amplitude of pure background spectra

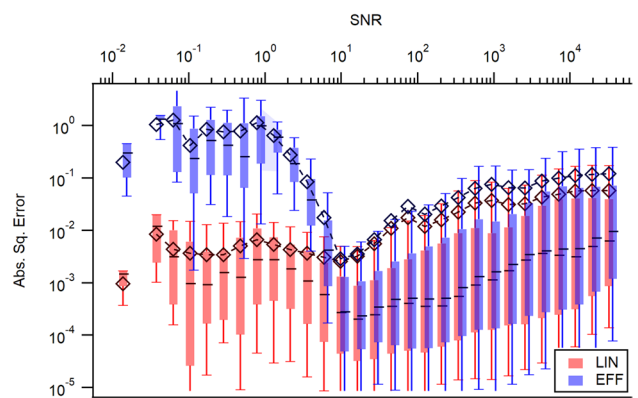


Fig. 20 Fitter squared error as a function of input SNR. Linear fit in red, high SNR trained EffV2 variant prediction in blue. Boxes indicate 25%-quantile, median and 75%-quantile; whiskers indicate 10% and 90% quantile. Diamonds and lines show the mean. For better visualization the EffV2 plot is slightly offset along the x -axis

Acknowledgements I would like to thank Horst Fischer and Michael Wand for the helpful discussions, Linda Ort for providing the measurement data, and the Max Planck Computing and Data Facility (MPCDF) for providing and maintaining the HPC clusters used for computing tasks during this study.

Author contributions L.R. collected the data, adjusted and trained the models and wrote the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Max Planck Graduate Center with the Johannes Gutenberg-Universität Mainz (MPGC).

Declarations

Conflict of interest The authors declare no Conflict of interest.

Code and data availability Python code for model training, config files, pre-trained model snapshots and raw data are publicly available at <https://keeper.mpd.mpg.de/d/24ecd2e114c94725961b/>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**(7553), 436–444 (2015)
2. I. Goodfellow, Y. Bengio, A. Courville, : Deep Learning. MIT Press, Cambridge (MA) (2016). <http://www.deeplearningbook.org>
3. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer
4. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020)
5. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
6. I. Cortés-Ciriano, A. Bender, Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.* **59**(3), 1269–1281 (2018)
7. I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv preprint [arXiv:1412.6572](https://arxiv.org/abs/1412.6572) (2014)
8. J. Acquarelli, T. Laarhoven, J. Gerretzen, T.N. Tran, L.M. Buydens, E. Marchiori, Convolutional neural networks for vibrational spectroscopic data analysis. *Anal. Chim. Acta* **954**, 22–31 (2017)
9. J. Xia, Y. Huang, Q. Li, Y. Xiong, S. Min, Convolutional neural network with near-infrared spectroscopy for plastic discrimination. *Environ. Chem. Lett.* **19**(5), 3547–3555 (2021)
10. J. Huang, H. Liu, J. Dai, W. Cai, Reconstruction for limited-data nonlinear tomographic absorption spectroscopy via deep learning. *J. Quant. Spectrosc. Radiat. Transfer* **218**, 187–193 (2018)
11. Y. Fu, R. Zhang, G. Enemali, A. Upadhyay, M. Lengden, C. Liu, Convolutional neural network aided chemical species tomography for dynamic temperature imaging. In: 2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–5 (2022). IEEE
12. C.D. Rankine, M.M. Madkhali, T.J. Penfold, A deep neural network for the rapid prediction of x-ray absorption spectra. *J. Phys. Chem. A* **124**(21), 4263–4270 (2020)
13. J. Nicely, T. Hanco, H. Riris, Applicability of neural networks to etalon fringe filtering in laser spectrometers. *J. Quant. Spectrosc. Radiat. Transfer* **211**, 115–122 (2018)
14. J. Pyo, S.M. Hong, Y.S. Kwon, M.S. Kim, K.H. Cho, Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. *Sci. Total Environ.* **741**, 140162 (2020)
15. L. Tian, J. Sun, J. Chang, J. Xia, Z. Zhang, A.A. Kolomenskii, H.A. Schuessler, S. Zhang, Retrieval of gas concentrations in optical spectroscopy with deep learning. *Measurement* **182**, 109739 (2021)
16. L.L. Röder, H. Fischer, Theoretical investigation of applicability and limitations of advanced noise reduction methods for wavelength modulation spectroscopy. *Appl. Phys. B* **128**(1), 1–10 (2022)
17. P. Werle, R. Mücke, F. Slemr, The limits of signal averaging in atmospheric trace-gas monitoring by tunable diode-laser absorption spectroscopy (tdlas). *Appl. Phys. B* **57**(2), 131–139 (1993)
18. Z. Wang, P. Fu, X. Chao, Laser absorption sensing systems: challenges, modeling, and design optimization. *Appl. Sci.* **9**(13), 2723 (2019)
19. K.-I. Funahashi, On the approximate realization of continuous mappings by neural networks. *Neural Netw.* **2**(3), 183–192 (1989)
20. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**(5), 359–366 (1989)
21. G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**(4), 303–314 (1989)
22. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
23. J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
24. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
25. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
26. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017)
28. N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer. In: International Conference on Machine Learning, pp. 4055–4064 (2018). PMLR

29. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
30. C. Schiller, H. Bozem, C. Gurk, U. Parchatka, R. Königstedt, G. Harris, J. Lelieveld, H. Fischer, Applications of quantum cascade lasers for sensitive trace gas measurements of co, ch₄, n₂o and hcho. *Appl. Phys. B* **92**(3), 419–430 (2008)
31. D. Richter, P. Weibring, J.G. Walega, A. Fried, S.M. Spuler, M.S. Taubman, Compact highly sensitive multi-species airborne mid-ir spectrometer. *Appl. Phys. B* **119**(1), 119–131 (2015)
32. J.B. McManus, C. Dyroff, Spectroscopic Measurement Response to Interference Fringes: Fundamental and Aliased Fringes. FLAIR Conference (2022)
33. C. Torrence, G.P. Compo, A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* **79**(1), 61–78 (1998). [https://doi.org/10.1175/1520-0477\(1998\)079%3C0061:APGTWA%3E2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079%3C0061:APGTWA%3E2.0.CO;2)
34. A.K. Leung, F. Chau, J. Gao, A review on applications of wavelet transform techniques in chemical analysis: 1989–1997. *Chemom. Intell. Lab. Syst.* **43**(1–2), 165–184 (1998)
35. M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106 (2021). PMLR
36. X.-J. Mao, C. Shen, Y.-B. Yang, Image restoration using convolutional auto-encoders with symmetric skip connections. arXiv preprint [arXiv:1606.08921](https://arxiv.org/abs/1606.08921) (2016)
37. F. Wienhold, H. Fischer, P. Hoor, V. Wagner, R. Königstedt, G. Harris, J. Anders, R. Grisar, M. Knothe, W. Riedel, Tristar-a tracer in situ tdlas for atmospheric research. *Appl. Phys. B* **67**(4), 411–417 (1998)
38. J. Faist, F. Capasso, D.L. Sivco, C. Sirtori, A.L. Hutchinson, A.Y. Cho, Quantum cascade laser. *Science* **264**(5158), 553–556 (1994). <https://doi.org/10.1126/science.264.5158.553>
39. A. Perrin, D. Jacquemart, F.K. Tchana, N. Lacome, Absolute line intensities measurements and calculations for the 5.7 and 3.6 μm bands of formaldehyde. *J. Quant. Spectrosc. Radiat. Transfer* **110**(9–10), 700–716 (2009)
40. J.U. White, Optical system providing a long optical path. 2779230, January 1957
41. A. Sergeev, M.D. Balso, Horovod: fast and easy distributed deep learning in TensorFlow. arXiv preprint [arXiv:1802.05799](https://arxiv.org/abs/1802.05799) (2018)
42. D. Herriott, H. Kogelnik, R. Kompfner, Off-axis paths in spherical mirror interferometers. *Appl. Opt.* **3**(4), 523–526 (1964)
43. J.B. McManus, P.L. Keabian, M. Zahniser, Astigmatic mirror multipass absorption cells for long-path-length spectroscopy. *Appl. Opt.* **34**(18), 3336–3348 (1995)
44. G. Li, I.E. Gordon, L.S. Rothman, Y. Tan, S.-M. Hu, S. Kass, A. Campargue, E.S. Medvedev, Rovibrational line lists for nine isotopologues of the co molecule in the x1σ⁺ ground electronic state. *Astrophys. J. Suppl. Ser.* **216**(1), 15 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.