# Germany Renewable Energy Supply Forecasting with the Transformer Architecture

**Long Hoang Nguyen**
M.Sc. Machine Learning

## Abstract

This project explores the feasibility of using Transformer models to predict solar and wind energy outputs. Applied to energy supply data of Germany from 2019 to 2022, Transformer models show promising accuracy in capturing complex temporal dependencies. The code for this project is available at https://github.com/LongHoangNguyen06/ML-Energy-System

## Introduction

The dataset used in this project includes

- **Energy Supply**: Information on the amount of energy generated in Germany from renewable and non-renewable sources in Megawatt (MW). The data has a resolution of 15 minutes.
- **Energy Demand**: Records of the realisied energy demand of Germany in MW, with a resolution of 15 minutes.
- **Weather Forecasts**: Forecasts of 15 meteorological variables from 80 locations in Germany that influence the production of solar and wind energy. These forecasts are generated daily at midnight for the following 24 hours, with a resolution of one hour.
- **Energy Capacity**: Data on the capacity in MW of installed renewable energy infrastructure, with a resolution of one year.
- **Energy Prices**: Market prices for energy across Europe in €/MWh, with a resolution of one hour.

The task of this project is to forecast energy supply for one-hour and 24-hour horizons for solar energy, offshore wind energy, and onshore wind energy, resulting in a total of six target quantities.

The modelling pipeline involves several key steps. Initially, raw data is loaded and inspected for missing values and data points are synced to a common resolution. The next steps handle missing values, convert different data sources to a common time-zone.

Once the data is prepared, feature selection is performed to identify the most relevant variables influencing energy supply. Techniques such as pair-wise feature correlation, feature-target correlation, Lasso regression, and decision tree analysis are employed.

Finally for training purpose, the data is normalized and put into a suitable format for the autoregressive encoder-decoder architecture of Transformer.

Special attention was given to the hyperparameter optimization of the transformer model, utilizing Bayes optimization. The model is trained on the training data from 2019 to 2021, and its performance is evaluated on the test set of 2022.

Seminar Machine Learning for Renewable Energy (Summer Term 2024), Tübingen.

## Related work

The transformer model, introduced by Vaswani et al. Vaswani et al., 2023, at its core employs a self- and cross-attention mechanism for neural machine translation tasks. The encoder-decoder structure fuses all source language tokens together in the encoding part and produces target language tokens in an auto-progressive fashion in the decoding part. Leveraging the self- and cross-attention's nature of data fusing, this architecture can also be transferred naturally to time-series forecasting tasks which requires temporal dependencies modelling.

## Methodology

### Network Architecture

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be sequence consist of $n$ time point with $d$ features, then the encoder produces a sequence of attentions $A \in \mathbb{R}^{n \times d}$ as

$$A = \textbf{Encoder}(x_1, \ldots, x_n)$$

Let $w_1, \ldots, w_m \in \mathbb{R}^k$ be the weather forecasts for the next $m$ hours, the decoder produces the energy supply forecast $E \in \mathbb{R}$ as

$$E = \textbf{Decoder}(w_1, \ldots, w_m, A)$$

See figure 1 for a visual representation of the model architecture.

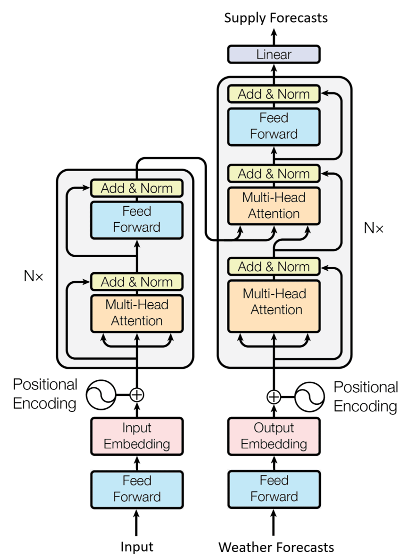


Figure 1: Model architecture for a single supply target prediction

Given the need to forecast energy outputs for 1-hour and 24-hour horizons for solar energy, offshore wind energy, and onshore wind energy, we have six targets in total. To handle these targets effectively, we experimented with four different network architectures. **Complete Multitasking**: A single network produces all six targets. **Horizon Multitasking**: Two separate networks, each responsible for one of the two forecasting horizons. **Target Multitasking**: Three separate networks, each dedicated to one of the energy sources (solar, offshore wind, onshore wind) across both horizons. **No Multitasking**: Six separate networks, one for each combination of horizon and energy source.

The final architecture is chosen through hyperparameter optimization. As metric for optimization, we use RMSE.

### Features and Inputs

At the most complete setting, the input of the encoder consists of the installed capacity, the energy demand, the energy supply and weather forecasts of the forecasting origin; together they form a token of dimension $d$. To modelling temporal dependencies, we include the last $n$ lags; this gives us an input and output of common dimension $n \times d$.

For the decoder, the input consists of the weather forecasts of dimension $k$ for the next $m$ hours, where $m$ is the forecasting horizon; this gives us an input of dimension $m \times k$.

Note that the self- and cross-attention operator requires every input to have the same dimension, so a linear function is applied to the inputs to ensure this.

### Data Preprocessing

The raw data format while consists sufficient information, but is not the optimal format for the training. The data preprocessing therefore includes the following steps:

- **Handling Missing Values**: the few columns with very high missing values in the prices dataset are dropped.
- **Aggregating Weather Data**: weather data is aggregated to a 1-hour resolution, we compute for whole Germany the min, max and mean value of each forecast column.
- **Daylight Saving Adjustment**: Adjustments are made for daylight saving time.
- **Time Conversion**: All timestamps are converted to Greenwich Mean Time.
- **Price Differencing**: Differencing is applied to the prices to make the series mean-stationary.
- **Normalizing Data**: The normalization parameters are learned only from the training set.
- **Resolution Alignment**: The time resolution of different datasets is aligned to cover the same date range. This includes decreasing the time resolution of demand and supply data and increasing the time resolution of installed capacity data.

On each token of both encoder and decoder we furthermore insert time information by adding the hour of the day, the day of the week and day of year as additional features.

### Feature Selection

We employ the following techniques to identify the most important features:

- **Pair-wise Feature Correlation**: analyzes linear relationship between each pair of feature and possibility to disgard redundant feature.
- **Target Correlation**: assess the correlation of each feature with the target variables to gauge their potential predictive power.
- **Lasso Regression**: Lasso (Least Absolute Shrinkage and Selection Operator) regression is used to perform feature selection by choosing feature with coeffients higher than $0.01$.
- **Decision Tree Analysis**: Decision tree models are used to assess feature importance.

### Hyperparameter Optimization

We use Bayesian optimization with the best validation loss of a training session as the metric for optimizing hyperparameters. The final model is then trained on the entire training set and validation with the best hyperparameters. The hyperparameter optimization process was executed on four RTX-2080ti GPUs over the course of 13 hours.

To summarize, we apply AdaGrad optimizer with an initial learning rate of $1 \cdot 10^{-3}$, a cosine scheduler with a final learning rate of $1 \cdot 10^{-7}$, and do gradient clipping at $1.0$. The model consists of one Transformer layer with one attention head, and the latent space dimension is 11 times higher than the input dimension. We use only 1 hour of past data and have three independent networks for wind off., wind on., and photovoltaic. Each network has two outputs for 1h and 24h predictions.

# Evaluation

## Qualitative evaluation

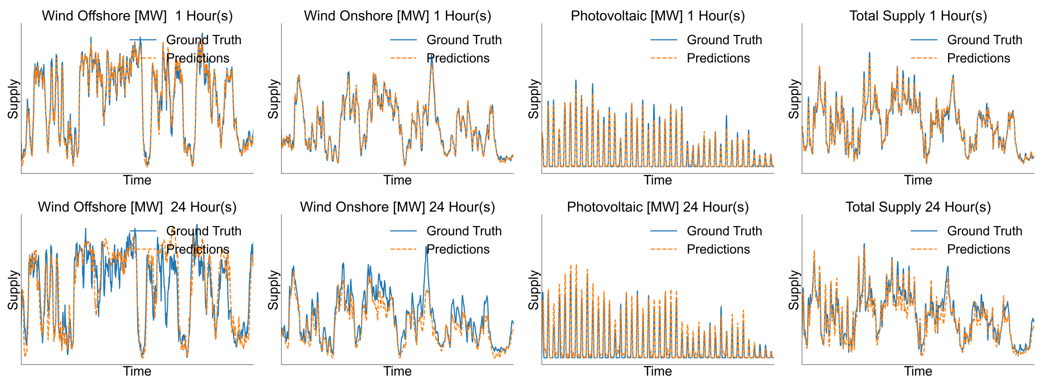


Figure 2: Comparing groundtruth and prediction for each source of energy supply and total supply.

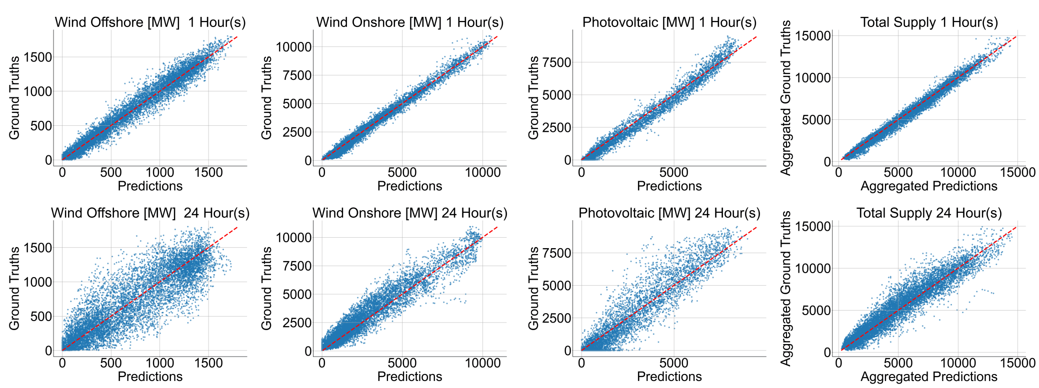


Figure 3: Alignment of groundtruth and prediction for each source of energy supply and total supply.

## Quantitative evaluation

| | **Wind Off.** | **Wind On.** | **Photo.** | Total |
|---|---|---|---|---|
| **1 Hour** | 91.40 | 290.13 | 315.70 | 232.41 |
| **24 Hours** | 235.66 | 719.18 | 737.53 | 564.12 |

Table 1: RMSE of the model for each source of energy supply and total supply.

| | **Wind Off.** | **Wind On.** | **Photo.** | Total |
|---|---|---|---|---|
| **1 Hour** | 0.99 | 0.98 | 0.99 | 0.98 |
| **24 Hours** | 0.95 | 0.87 | 0.95 | 0.92 |

Table 2: R value of the model for each source of energy supply and total supply.

## Discussion

The RMSE values show that forecasting errors increase significantly over a 24-hour horizon compared to a 1-hour horizon, reflecting the inherent difficulty of predicting longer-term future values accurately. Wind Onshore and Photovoltaic energy have higher RMSE values. To address the weakness of Wind On. and Photo., we suggest improving the handling of weather data by reducing the level of aggregation. More granular weather forecasts could capture the nuances of local conditions that significantly influence Wind Onshore and Photovoltaic energy production.

## Conclusion

Transformers, although highly effective in language processing tasks, encounter significant challenges in time series forecasting due to the inherent nature of time series data, where similar past values can lead to drastically different future values. The increased computational costs and the need for hyperoptimization may not offer a significant advantage over traditional forecasting methods, which can sometimes provide comparable or better results with lower complexity.

## References

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (Aug. 2023). *Attention Is All You Need*. arXiv:1706.03762 [cs]. DOI: 10.48550/arXiv.1706.03762.