Exploring Indoor Localization with Transformer-Based Models: A CNN-Transformer Hybrid Approach for WiFi Fingerprinting

Nicu Savin
n.savin@student.utwente.nl
University of Twente
The Netherlands

ABSTRACT

Indoor localization has become a target for many researchers due to its vast range of applications. Due to signal fading and scattering, conventional GPS-based techniques are impractical for indoor localization. However, state-of-the-art deep learning models have shown promising results in this field.

The method for indoor localization presented in this research makes use of a transformer-based model and Received Signal Strength (RSS) measurements. The proposed model will be assessed in both regression tasks: predicting X and Y coordinates, and classification tasks: floor classification.

The results of this research aim to contribute to the advancement of indoor localization systems by providing evidence that transformer-based models might be a good direction to follow for enhancing localization accuracy.

KEYWORDS

WiFi fingerprinting, Deep learning, Indoor localization, Transformer

1 INTRODUCTION

The demand for accurate and efficient indoor localization systems has grown due to their importance in numerous applications. Indoor localization remains a significant challenge, due to the attenuation and scattering of Global Positioning System (GPS) signals by roofs, walls, and other obstructions. To address this issue, WiFi fingerprinting has emerged as a popular approach, as most smartphones and IoT devices are equipped with WiFi interfaces, which leverage the Received Signal Strength Indicator (RSSI) values from multiple Access Points (APs) to determine the location of devices within indoor environments [6].

Recently, deep learning-based methods have demonstrated promising advancements in indoor localization tasks using WiFi fingerprinting compared to traditional techniques [1, 5, 8, 11]. These models have demonstrated their effectiveness in capturing spatial information and learning the underlying structure of the data, resulting in improved localization accuracy compared to traditional approaches.

Transformers, first introduced in [10], have shown excellent results in Natural Language Processing (NLP) tasks and have started to be applied in a variety of other domains, including Computer Vision, due to their effectiveness at learning spatial relationships and capturing long-range dependencies. Thus, transformers have the potential for improving performance in a wide range of tasks, including indoor localization.

In this research, we propose a novel indoor localization approach using a CNN-Transformer hybrid model. The main aim of this research is to combine the strengths of both CNNs and Transformer models, to harness the local feature extraction capabilities of CNNs and the global correlation understanding of the Transformer model, and apply this powerful combination to the task of indoor localization. We show the effectiveness of our suggested approach in predicting the floor number in a multi-story building as well as the (X, Y) coordinates of a device within an indoor environment. We compare our model's performance with state-of-the-art models, aiming to show the potential of transformer-based models in this domain.

This research aims to contribute to the advancement of indoor localization systems by providing evidence that transformer-based models might be a good direction to follow for enhancing localization accuracy. We also aim to stimulate further research and exploration of other transformer-based approaches in the context of indoor localization.

1.1 Paper Structure

The remainder of the paper is structured as follows:

- **Section 2 - Problem Statement:** This section defines the challenges associated with indoor localization. It introduces the paper's goal of proposing a transformer-based model, and the research questions that will be answered.
- **Section 3 - Related Work:** This section provides a comprehensive review of the literature on indoor localization using WiFi fingerprinting and the deep learning techniques that have been applied to improve performance. It discusses the potential of transformer-based models for indoor localization tasks.
- **Section 4 - Methodology:** This section describes the system architecture of the proposed model and the idea behind it. It introduces the dataset that will be used and the data preparation process.
- **Section 5 - Training Phase:** This section provides a detailed analysis of the model training process. It elaborates on the model's parameters, training parameters and techniques used for training.
- **Section 6 - Results:** This chapter presents the research findings. It compares the performance of the proposed model with other state-of-the-art models, both in terms of coordinate prediction and floor accuracy.

39th Twente Student Conference on IT , July 7, 2023,

- **Section 7 - Conclusion:** This section wraps up the findings of the study. It also elaborates on a possible enhancement that could be made to the model for a potential better performance.

2 PROBLEM STATEMENT

Even with major improvements in indoor localization, such as WiFi fingerprinting and the use of various deep learning approaches, it is still difficult to accurately locate devices indoors. Despite the good progress in indoor localization tasks using deep learning-based approaches, there is still a constant need for more reliable, and accurate models.

Therefore, this research's primary goal is to propose a local feature transformer-based (LF-Transformer) model that competes in terms of accuracy with existing state-of-the-art deep learning models. Furthermore, this research aims to stimulate interest in exploring other transformer-based approaches for indoor localization using WiFi fingerprinting.

The performance of the proposed model will be tested based on two main tasks: a regression task, which will assess the model capabilities of correctly predicting the (X, Y) coordinates, and a classification task, which will assess the model capabilities of correctly predicting the floor number. For the regression task, we will assess the model based on the average Euclidean distance between the predicted ($\hat{X}$, $\hat{Y}$) and actual (X,Y) coordinates. Additionally, the 75th percentile and 95th percentile will be used to understand the distribution of errors. For the classification task, we will assess the model based on the accuracy of correctly predicting the floor number.

2.1 Research Questions

This paper, introducing a novel LF-Transformer model, aims at answering the following research questions:

(1) How can transformer based model be exploited to the field of indoor localization using WiFi fingerprinting?
(2) How does the proposed transformer based model perform in the task of predicting the X, Y coordinates and floor number in multi-story indoor environments, and how does this new approach compare to established deep learning methods?

3 RELATED WORK

Extensive research has been conducted in the field of indoor localization using WiFi fingerprinting, with various deep-learning techniques being employed to improve performance. [1] employed local features through deep LSTM, leveraging the temporal patterns contained inside WiFi fingerprints, and successfully addressing the complexities of indoor situations. [5, 8] demonstrated the efficacy of Convolutional Neural Networks (CNNs) for indoor localization. They applied CNNs for indoor localization, highlighting the capability of CNNs to extract spatial features and enhance localization accuracy.

Recent studies have also reviewed the state of the art in machine learning-based indoor localization using WiFi RSSI fingerprints, outlining various techniques, challenges, and opportunities in the field [6]. [3] proposed a fingerprinting indoor localization algorithm based on deep learning, showing the importance and efficacy of deep learning models compared to state-of-the-art approaches.

However, despite these remarkable developments, there remains vast uncharted territory in the application of transformer architectures for indoor localization tasks. These architectures, introduced by Vaswani et al. [10], have revolutionized the field of Natural Language Processing and recently shown promise in diverse domains such as Computer Vision [2]. In [4], the authors were the first who developed a novel transformer-based approach to the task of indoor localization, greatly surpassing, in terms of accuracy, other state-of-the-art methodologies.

This paper proposes another novel LF-Transformer model. This model combines the benefits of CNNs and Transformer-based models, efficiently comprehending both local features and global correlations in data sequences, and thereby addressing the unique challenges of indoor localization tasks. To the best of our knowledge, this is the first study that investigates the combination of CNNs and Transformers for WiFi-based indoor localization tasks.

4 METHODOLOGY

4.1 System design

4.1.1 Transformer Overview. The general transformer model uses an encoder-decoder architecture. Each has multiple identical layers. The encoder maps an input sequence to a continuous representation that holds the entire input information. The decoder then generates an output sequence from this representation. However, the original transformer was designed for sequence-to-sequence tasks, making it unsuitable for our requirements. Given that we aim to predict the (X,Y) coordinates or floor number from a sequence of WiFi signal strengths, we do not need a sequence output. Therefore, the decoder portion of the original transformer model is excluded from our proposed model.

In this research, we adopt a sequential approach to the problem of indoor localization. Each sample in our dataset is a sequence of 520 Received Signal Strength (RSS) values, with each value corresponding to a signal reading from a distinct Access Point (AP). Despite the APs order not reflecting the spatial location, each AP holds a specific and consistent position in the sequence across all samples. This essentially means that, for example, the first value in the sequence always represents the signal strength reading from the same AP. This consistency in positioning is maintained for all APs and across all samples, providing a basis for our model to learn the relationships between individual APs' RSS values and the corresponding physical coordinates or floor numbers. The assumption is that nearby APs might influence each other's signal strength due to interference or shared environmental factors (walls, floors), hence the local dependency. On the other hand, long-range dependencies could arise due to factors like signal propagation patterns or large-scale environmental structures.

4.1.2 Transformer Block. The core of the proposed model consists of a sequence of transformer blocks, each transformer block will follow the general encoder principle of transformers:

(1) The first sublayer of the transformer block implements a multi-head self-attention mechanism. The sublayer is also

followed by a normalization and a dropout layer with a dropout rate of 0.1.

(2) The second sublayer of the transformer block is a fully-connected feed-forward networks. It is consisted of two successive linear transformations. The Rectified Linear Unit (ReLU) activation function is used in between these two transformations.

Layer (type)	Output Shape	Param #
input_18 (InputLayer)	[(None, 520, 1)]	0
conv1d_17 (Conv1D)	(None, 520, 64)	192
re_lu_14 (ReLU)	(None, 520, 64)	0
batch_normalization_17 (BatchNormalization)	(None, 520, 64)	256
positional_encoding_28 (PositionalEncoding)	(None, 520, 64)	0
transformer_regressor_7 (TransformerRegressor)	(None, 520, 64)	130688
global_average_pooling1d_19 (GlobalAveragePooling1D)	(None, 64)	0
dense_602 (Dense)	(None, 2)	130

Total params: 131,266
Trainable params: 131,138
Non-trainable params: 128

Figure 1: Model Summary

4.2 Proposed LF-Transformer Model

The proposed LF-Transformer model is a hybrid model combining Convolutional Neural Networks (CNN) and the Transformer model, with a primary focus on the Transformer. The core of the proposed model adapts the main transformer architecture with a few modifications to make it suitable for regression and classification tasks. The model is composed of a convolutional layer before the transformer regressor (or classifier in case of predicting the floor), each consisting of a sequence of transformer blocks. The CNN layer of our model is proficient in identifying local features in the data, specifically patterns across the RSS values that might indicate a particular location. Following the ReLU activation function then introduces non-linearity into the feature map, followed by a batch normalization layer. The output is then passed through a positional encoding function before being processed by a sequence of transformer blocks. After processing the data through these Transformer Blocks, we extract the output for the regression and classification tasks. For regression, the output from the last Transformer block is passed to a Dense Layer, which generates the predicted X, Y coordinates. For the classification task, the output from the last Transformer block is passed to a Dense Layer , with softmax activation, to predict the floor number. The overview of the model can be observed in Figure 1.

4.2.1 Local and Long-Range Dependencies. The idea of the model is the following:

(1) The convolutional layer is focusing on identifying local dependencies in the data. Local dependencies refer to the patterns or features that can be detected within close proximity, such as the strength of signals from a group of Access Points (APs). This local information is important as it provides a base understanding of the environment.

(2) The transformer layer specializes in identifying long-range dependencies. Long-range dependencies refer to relationships between elements that are not immediately adjacent or close to each other. The self-attention mechanism in the transformer layer helps to understand these relationships by weighting the importance of different parts of the input. This enables the model to highlight relevant patterns across the entire sequence, combining them with the local patterns learned by the convolutional layer to understand both nearby and distant features.

This is a powerful and promising combination that allows the model to understand both local and long-range dependencies in the data.

4.3 Dataset and preprocessing

Throughout this research, UjiIndoorLoc, the biggest open-access indoor localization database, will be used for testing and evaluation. The database contains a total of 21048 fingerprint samples collected from 520 Access Points (APs) covering 3 buildings and several floors. The dataset will be divided into training, validation, and testing sets, with the testing set reserved for model evaluation, ensuring that the model's performance is evaluated on unseen data.

4.3.1 Data normalization. Before the training process, the dataset is normalized. In this dataset, the RSS fingerprints from APs range from -104 dBm to 0 dBm. Also, for any AP that is not detected, it will have a RSS value of 100. This dataset will be normalized according to the formula below (Formula 1), we will convert the RSS values to be in range (0,1).In [9], it was shown that this type of normalization tends to represent RSS values with the best performance, thus it has been also used in this paper.

$$Powed = \begin{cases} 0 & , RSS_i = 100 \\ \left(\frac{RSS_i - \min}{-\min}\right)^e & ,\text{otherwise} \end{cases} \quad \text{(Formula 1)}$$

where RSS_i represents the RSS value of the i-th AP, min represents the lowest RSS value registered, and e represents Euler's number.

5 TRAINING PHASE

The training of our model involved numerous processes to ensure optimal performance. We incorporated K-Fold cross-validation with a fold count of five, which served to reduce overfitting and enhance model generalizability. For the initial convolution layer, we set the filter count to 64 and kernel size to 2. The Transformer model parameters were selected as follows: 6 Transformer layers, an embedding dimension of 64 (equal to the filter count), 4 attention heads, and a feed-forward dimension of 32. We also incorporated a dropout rate of 0.1 after each layer in the feed-forward networks

Parameters		Value
Convolutional	Filters	64
	Kernel size	2
Transformer	Number of transformer layers	6
	Embedding dimension	64
	Number of heads	4
	Feed-forward dimension	32
	Dropout rate	0.1
Training	Learning rate	1e-4
	Epochs	200
	Early stopping patience	20

Table 1: Overview of all the parameters

to prevent overfitting. We trained our model for a maximum of 200 epochs with an early stopping mechanism to prevent overfitting. The early stopping was configured to monitor the validation loss, with a patience of 20 epochs. After the training, the model with the lowest validation loss is restored. An overview of all the training parameters can be seen in Table 1. Figure 2 shows the training loss and the validation loss of the positioning model. Figure 3 shows the training and validation accuracy for the classification task.

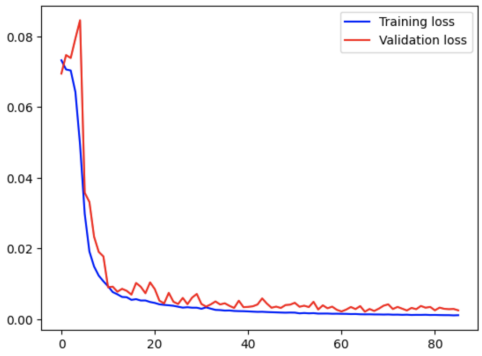

Figure 2: Training overview (X, Y) prediction (Loss/Epochs)

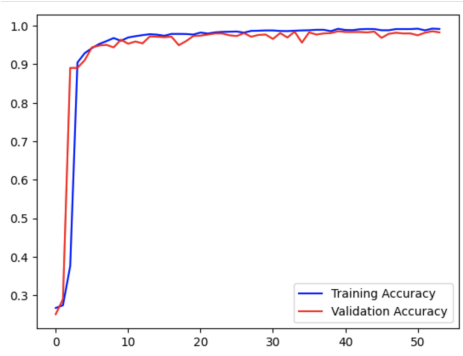

Figure 3: Training overview floor prediction (Accuracy/Epochs)

6 RESULTS

The performance of the proposed LF-Transformer was compared to other state-of-the-art deep learning models, specifically: CNNLoc [8], BayesCNN [7] , DNN [3], bAaT [4], and eAaT [4].

As displayed in Table 3, the LF-Transformer achieved a Mean Absolute Error (MAE) of 9.44 meters. This result was much better compared to the CNNLoc, BayesCNN, and DNN models, which achieved MAEs of 11.78, 41.79, and 133.40 meters, respectively. However, it was surpassed by the bAaT and eAaT models, which achieved MAEs of 8.45 and 8.40 meters.

The model also demonstrated its proficiency in predicting the floor number (Table 2). It achieved an accuracy of 95.90%, slightly lower than the accuracy of the CNNLoc model, which achieved the highest floor prediction accuracy of 95.92%. However, the LF-Transformer outperformed all other models, including BayesCNN, DNN, bAaT, and eAaT, which achieved accuracies of 90.64%, 41.58%, 94.42%, and 94.69%, respectively.

We can observe that transformer-based models, such as our LF-Transformer, eAaT, and bAaT, outperform the non-transformer model such as CNN and DNN, highlighting the effectiveness of transformer models. While CNN models are able to detect local patterns, they have difficulty identifying long-range dependencies. On the other hand, DNNs are ineffective of detecting sequential dependencies in the data. This outperformance can be ascribed to the unique aspect of the transformer model, specifically self-attention mechanisms. This self-attention capability allows the model to assign different levels of significance to various parts of the input sequence when making predictions. This is useful where certain local patterns or signal strengths (RSS values) can be more indicative of a particular location than others. Additionally, the position at a particular location might be influenced by the signal strengths at several other locations, this is captured by the transformers models due to their capacity for handling long-range data dependencies.

Model	Floor Accuracy
LF-Transformer	95.90%
CNNLoc [8]	**95.92%**
BayesCNN [7]	90.64%
DNN [3]	41.58%
bAaT [4]	94.42%
eAaT [4]	94.69%

Table 2: Results for predicting coordinates compared to other approaches

7 CONCLUSION

Despite the promising results, there remains room for further improvement. The eAaT and bAaT models still surpass our model in average distance error, suggesting that more research can be done to improve the performance of our model further. Future work may also involve exploring other parameter configurations, adding more data for training, or investigating other transformer-based models. Despite these possibilities for future exploration, our work successfully demonstrates the potential of using a transformer-based model

Model	Mean Absolute Error (MAE)	75th Percentile (m)	95th Percentile (m)
LF-Transformer	9.44	14.63	22.31
CNNLoc [8]	11.78	-	-
BayesCNN [7]	41.79	49.28	75.25
DNN [3]	133.40	170.85	213.10
bAaT [4]	8.45	**10.64**	20.41
eAaT [4]	**8.40**	10.66	**20.33**

Table 3: Results for predicting coordinates compared to other approaches

for coordinate prediction and floor accuracy for indoor localization tasks. Looking also at Table 3, we can see that transformer-based models (LF-Transformer, eAaT, bAaT) dominate in terms of accuracy compared to other deep learning models. The findings of this research suggest that transformer models, due to their unique properties and capabilities, present a promising avenue for future research in indoor localization tasks.

7.1 Potential Improvements

A potential improvement of the model would be to arrange the APs in the database in a meaningful way. As the order of the APs in the columns' database does not represent their physical locations, the model might not capture the spatial correlation between nearby APs accurately. However, if the APs were arranged according to their physical locations (for example, in increasing order of x and y coordinates), it could introduce a form of spatial relevance into the data. The model could then potentially discover spatially relevant features, which could result in enhanced performance. But because of the dynamic nature of indoor environments (for example, APs being added, removed, or relocated), this would require a consistent configuration of APs across all samples, which could prove to be a challenging task.

REFERENCES

[1] Zhenghua Chen, Han Zou, Jianfei Yang, Hao Jiang, Hao Jiang, and Lihua Xie. 2019. WiFi Fingerprinting Indoor Localization Using Local Feature-Based Deep LSTM. *IEEE Systems Journal* (2019). https://doi.org/10.1109/jsyst.2019.2918678

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (2020). https://doi.org/null

[3] Gibran Felix, Mario Siller, and Ernesto Navarro Alvarez. 2016. A fingerprinting indoor localization algorithm based deep learning. *International Conference on Ubiquitous and Future Networks* (2016). https://doi.org/10.1109/icufn.2016.7536949

[4] Son Nguyen, Duc Viet Le, and P. Havinga. 2023. Learning the world from its words: Anchor-agnostic Transformers for Fingerprint-based Indoor Localization. *Annual IEEE International Conference on Pervasive Computing and Communications* (2023). https://doi.org/10.1109/percom56429.2023.10099376

[5] Wafa Njima, Iness Ahriz, Rafik Zayani, Michel Terre, Ridha Bouallegue, Ridha Bouallegue, Ridha Bouallegue, and Ridha Bouallegue. 2019. Deep CNN for Indoor Localization in IoT-Sensor Systems. *Sensors* (2019). https://doi.org/10.3390/s19143127

[6] Navneet Singh, Sangho Choe, and Rajiv Punmiya. 2021. Machine Learning Based Indoor Localization Using Wi-Fi RSSI Fingerprints: An Overview. *IEEE Access* (2021). https://doi.org/10.1109/access.2021.3111083

[7] Shreya Sinha and Duc V. Le. 2021. Completely Automated CNN Architecture Design Based on VGG Blocks for Fingerprinting Localisation. *International Conference on Indoor Positioning and Indoor Navigation* (2021). https://doi.org/10.1109/ipin51156.2021.9662642

[8] Xudong Song, Xiaochen Fan, Xiangjian He, Chaocan Xiang, Qianwen Ye, Xiang Huang, Gengfa Fang, Liming Luke Chen, Jing Qin, and Zumin Wang. 2019. CNNLoc: Deep-Learning Based Indoor Localization with WiFi Fingerprinting. *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (2019). https://doi.org/10.1109/smartworld-uic-atc-scalcom-iop-sci.2019.00139

[9] Joaquín Torres-Sospedra, Raúl Montoliu, Sergio Trilles, Oscar Belmonte, and Joaquín Huerta. 2015. Comprehensive analysis of distance and similarity measures for Wi-Fi fingerprinting indoor positioning systems. *Expert Systems With Applications* (2015). https://doi.org/10.1016/j.eswa.2015.08.013

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[11] Wei Zhang, Kan Liu, Weidong Zhang, Youmei Zhang, and Jason Gu. 2016. Deep Neural Networks for wireless localization in indoor and outdoor environments. *Neurocomputing* (2016). https://doi.org/10.1016/j.neucom.2016.02.055