Passenger flow forecasting framework based on vision transformer and inpainting: Application to a public transport system

T. Bapaume[a,b,*], E. Côme[a], J. Roos[b], M. Ameli[a] and L. Oukhellou[a]

[a] Université Gustave Eiffel, Marne La Vallée, France
thomas.bapaume@univ-eiffel.fr, etienne.come@univ-eiffel.fr, mostafa.ameli@univ-eiffel.fr,
latifa.oukhellou@univ-eiffel.fr

[b] Régie Autonome des Transports Parisiens, Paris, France
jeremy.roos@ratp.fr

[*] Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

January 11, 2022

Keywords: Short-term Forecating, Passenger Load, Deep Learning, Transformer, Inpainting

1 INTRODUCTION

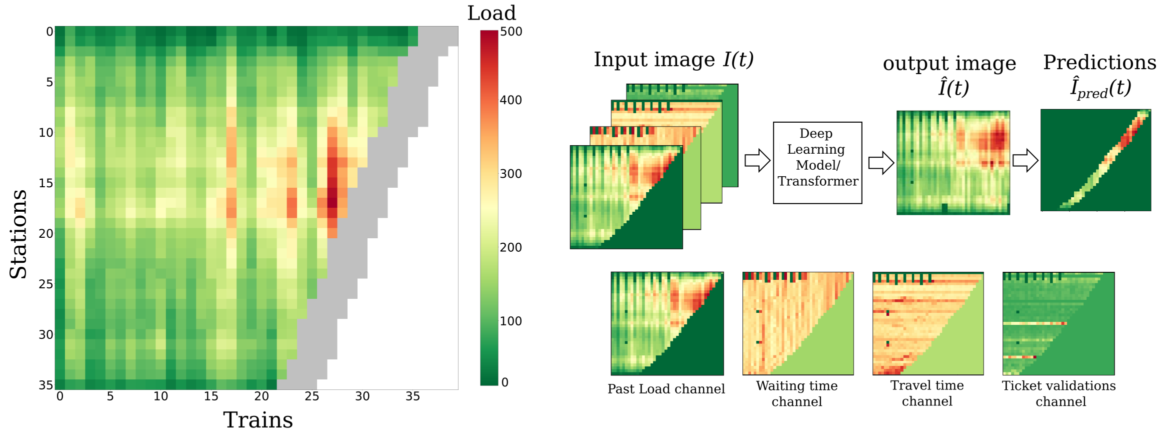
Short-term forecasting is one of the most important challenges in intelligent transport systems (ITS). The demand information is crucial for transport operators in order to anticipate and optimize their service level and for travelers to have robust information. In addition, a good predictor can contribute to system resilience by predicting disrupted situations. In most forecasting models, the data collectors are based on regular time series and single stations of the urban area. This study considers trains as data collectors, i.e., sensors. Thus, we are not limited to single location sensors. Moreover, we have to deal with irregular time series. In the literature, there are few studies that take this point into account (Pasini *et al.*, 2019).

Bapaume *et al.* (2021) proposed a U-net framework to take into account the specificity of real train data and perform forecasting over images through an inpainting task. This present study aims to extend the architecture of the deep learning process with self-attention mechanisms. These mechanisms outperform other solutions (on two popular data sets:CIFAR-100 and Imagenet) when they are used by the Transformer (Vaswani *et al.*, 2017) and Vision Transformer models (Dosovitskiy *et al.*, 2021) in Natural Language Processing (NLP) and image processing. Self-attention mechanisms show a high potential to be deployed for short-term prediction through solving image processing problems. In this study, two attention-based architectures were used: the U-transformer and Channel Vision Transformer. Then, we performed a benchmark on real load data (collected during 3 years of the Paris metro line 9) in order to compare the ability of these two architectures to forecast the loads of all the next 4 train departures of a metro line.

2 METHODOLOGY

2.1 Short-term train load prediction with metro traffic images

Figure 1(a) depicts an image that represents the train loads of all trains over the Paris metro line 9 in real time. Each pixel encodes the train load corresponding to the train identified by the column number at the station denoted by the row number. Therefore, each column represents the passage of a train along the metro line, and each row represents one station of the metro line. The images capture the train load of the metro line at time t.



(a) (b)

Figure 1 – *(a) Image of metro line 9 train loads (including past load channel) generated at time t. Colored pixels are past train loads, grey pixels are the future loads to be predicted for a short-term horizon (i.e., the pixels targeted by this study), and uncolored pixels are long-term horizon train loads. (b) Schema of the methodology with the multi-channel image as input and two-step output (inpainting and extraction of next four pixels).*

A train load image can be decomposed of multiple channels representing respectively data of the train network. For this study, 4 variables were defined: train load, remote ticket validation at the station, travel time, and waiting time as shown in Figure 1(b). Based on this image definition, we can present the methodology to forecast a certain number of unknown (future) train loads at a time t in grey in Figure 1(a).

2.2 Proposed image processing model with Transformer encoders

The forecasting task consists of two main steps: (i) the reconstruction of a missing part of an image $I(t)$ and (ii) the extraction of the forecasted loads from $\hat{I}(t)$ illustrated in Figure 1. In the first step, we applied the inpainting function f as the deep learning model to an input image. Then, to extract future train loads denoted by $\hat{I}_{pred}(t)$, we used a mask m_y in the second step to select the corresponding pixels from $\hat{I}(t)$. The following equation summarizes the forecasting task:

$$\hat{I}_{pred}(t) = \hat{I}(t) \odot m_y, \text{ where } \hat{I}(t) = f(I(t)) \text{ and } \odot \text{ is the product of Hadamard} \qquad (1)$$

The proposed architectures are built using the positional patch encoding introduced by the Vision Transformer (Dosovitskiy *et al.*, 2021) in order to deploy self-attention models to the image processing task. First, the goal is to apply this emerging framework, i.e., Transformer, to our image forecasting problem. Second, we use this methodology to directly predict all the next departures while we skip the inpainting task thanks to the Transformer's ability to enhance the relation between patches (set of pixels) inside an image. The two proposed models are detailed as follows:

U-Transformer: The idea is to replace the convolution and pooling layers with transformer encoders and decoders. The architecture is based by the Swin-Transformer proposed by Cao *et al.* (2021). Besides, the pooling task is achieved by concatenating patches together. The skip connections used by the conventional U-net are replaced by the sum of the deconvolution and convolution steps in order to preserve the spatial information during image reconstruction.

Channel Vision Transformer: This new architecture applies transformer encoders for each channel composing $I(t)$ (Figure 1(b)). The goal is to extract features independently from each variable. These features are concatenated and passed to a decoder to reconstruct the image and forecast loads.

3 RESULTS

3.1 Model configuration and numerical experiments

In order to validate our methodology, we evaluated its performance on real data collected from the Paris metro line 9, including train loads, ticket validations, and train waiting and travel times (input image presented in Figure 1(b)). The generated images have a size of 36 stations, 40 train sequences, and 4 channels (i.e., input variables). These images were generated based on the data collected every minute for the daily working hours of the line. In this study, 6 solution methods are considered as the reference methods: neural network (NN), fully convolutional network (CNN), convolutional network combined with NN (CNN+NN), U-net model, Transformer, and Vision Transformer (VIT). Furthermore, the two proposed models, U-Transformer and Channel Vision Transformer, were implemented in order to compared them with the reference solutions.

The *Softplus* function was used as the output function of all models because the load is defined positively between 0 and the maximum passenger load (800) per train. For the learning step, all models were trained with the Adam optimizer and a learning rate of 0.0001. Transformer models consisted of three encoder layers, a dimension of 64 by 128 heads, and used a patch size of 3. Finally, the learning for all models was done on 4 epochs. This number is small enough due to the high degree of similarity between the images.

For the learning and test datasets, images were generated from January 2019 to April 2021 with 1200 images per day, making a total of $900,000$ images. The test set includes 25% of the dataset and was selected so that we have at least the data of one month per year in order to illustrate different train contexts.

3.2 Forecasting results

Table 1 presents the results of the forecasting algorithms for the next four train departures from each station (i.e., grey pixels in Figure 1(a)). We measured the weighted mean absolute percentage error (WMAPE) for each scenario. The first column denotes the forecasting result for the whole *Testset*. The best result was obtained with U-Transformer (11.4%). Only two Transformer architectures (U-transformer and Channel Vision Transformer) outperform the U-net model. The other columns present the results for *Atypical* situations.

3.3 Atypical situations

In practice, it is crucial to evaluate the forecast model not only at the normal service level (as the results can be trivial), but in disrupted transport contexts which represent the uncertainty

Table 1 – *Forecasting results, expressed with weighted mean absolute percentage error (WMAPE), on the test set and on the atypical situations.*

Models (reference) / WMAPE [%] for	Testset	Nominal	Delay	High load	Strike	Lockdown
NN	18.9	16.1	25.7	20.1	40.2	19.4
CNN	13.3	11.2	17.6	13.0	29.8	15.7
CNN + NN (Ma *et al.*, 2017)	14.8	12.2	20.1	14.3	32.4	20.5
U-net (Bapaume *et al.*, 2021)	12.2	10.6	**16.0**	11.6	**28.1**	14.6
Transformer (Vaswani *et al.*, 2017)	12.7	10.1	18.4	13.1	34.8	15.6
VIT (Dosovitskiy *et al.*, 2021)	12.3	10.0	17.3	12.1	29.7	16.1
U-Transformer (This study)	**11.4**	**9.5**	**16.0**	**11.2**	28.7	**14.1**
Channel Vision Transformer **(This study)**	12.1	10.1	19.9	12.0	29.7	14.7

of the short-term forecasting in a real test case. Thus, we selected several real atypical scenarios based on two criteria extracted by (i) metric calculations and (ii) posterior knowledge of the metro traffic. The first criterion is easily computed over all metro line stations thanks to the image-oriented approach. For example, the range of total travel time of a single train allows us to identify disrupted events (e.g., longer parking times, traffic incidents). In addition, by computing the mean travel time of all trains at time t (i.e., for a single image), we can extract a *Delay* set if the difference between mean travel time and free-flow travel time is more than a given threshold (15 minutes in this test case). The results show that approximately 30% of images in the test set are identified as delayed situation. The remaining images are grouped in a *Nominal* set. In a similar way, the data set *high load* fluctuation can be extracted by recording high train load variations between two sequential stations for a single train. The second criterion can be used to identify images of a unique context or an external event. In this study, *Strike* and *Lockdown* are the two external events identified in the test case. Regarding forecasting results, U-Transformer outperforms all other models in all test cases, except for *Strike*. The U-net architecture outperforms U-Transformer for *Strike* images with a 0.06 lower WMAPE. Note that for the *Delay* set, U-net is the best model with 16.0% of error, the same as that of U-Transformer. In all image sets, Channel Vision Transformer is dominated by the best methods. However, it is among the top three methods in most of the test cases. Consequently, the results numerically prove that the proposed architecture, U-Transformer, can provide more accurate predictions for future train loads in particular for atypical situations.

4 DISCUSSION

The proposed framework forecasts all the targeted future train loads simultaneously (i.e., not recursively). The new architectures based on self-attention reformulate and solve the forecasting task as an inpainting problem. The benchmark showed the efficiency of the proposed methods over multiple test sets. Besides, we plan to include other methodologies reported in the literature, such as ensemble learning methods. We are currently extending the test case to other atypical situations (e.g., peak hours) based on posterior knowledge or metrics from the picture. The goal is to include other types of atypical situations in order to evaluate the robustness of the forecasting methods. Moreover, the explainability of the methods needs to be evaluated through attention scores.

References

Bapaume, Thomas, Côme, Etienne, Roos, Jérémy, Ameli, Mostafa, & Oukhellou, Latifa. 2021. Image Inpainting and Deep Learning to Forecast Short-Term Train Loads. *IEEE Access*, **9**, 98506–98522.

Cao, Hu, Wang, Yueyue, Chen, Joy, Jiang, Dongsheng, Zhang, Xiaopeng, Tian, Qi, & Wang, Manning. 2021. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv preprint arXiv:2105.05537*.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, & Houlsby, Neil. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.*

Ma, Xiaolei, Dai, Zhuang, He, Zhengbing, Ma, Jihui, Wang, Yong, & Wang, Yunpeng. 2017. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors*, **17**(04), 818.

Pasini, Kevin, Khouadjia, Mostepha, Samé, A., Ganansia, F., & Oukhellou, L. 2019. LSTM Encoder-Predictor for Short-Term Train Load Forecasting. *In: ECML/PKDD.*

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia. 2017. *Attention Is All You Need.*