Research Article

On the Vision-Beam Aided Tracking for Wireless 5G-Beyond Networks Using Long Short-Term Memory with Soft Attention Mechanism

Nasir Sinani[1*], Ferkan Yilmaz[2*]

[1] Computer Engineering, Yildiz Technical University, (ORCID: 0000-0001-6824-7323), nasir.sinani@std.yildiz.edu.tr
[2] Electronics and Communications Engineering, Yildiz Technical University, (ORCID: 0000-0001-6502-8280), ferkan@yildiz.edu.tr
[*] Correspondence: nasir.sinani@std.yildiz.edu.tr

(First received April 02, 2022 and in final form May 24, 2022)

Reference: Sinani, N., & Yilmaz, F. On the Vision-Beam Aided Tracking for Wireless 5G-Beyond Networks Using Long Short-Term Memory with Soft Attention Mechanism. *The European Journal of Research and Development*, *2*(2), 505–520.

Abstract

The growth of 5G technology and the continuous success of deep learning for various computer vision tasks in healthcare, self-driving cars, visual recognition, and many other areas, brought new challenges in the field of wireless communication. Moreover, 5G-Beyond networks primarily rely on how to maintain line-of-sight (LOS) links between base stations and mobile users. As such, one of the main challenges in 5G-Beyond networks is how to proactively maintain the hand-over mechanism for mobile users before blockages prevent mobile users from communicating, so as to avoid the latency of searching the best beamforming for the best performance. Accordingly, vision-aided millimeter-wave (mmWave) beam and blockage prediction has opened the door for new research for proactive hand-off and resource allocation. The purpose of this paper is to study wireless beam tracking on mmWave bands using deep learning approach evaluated on the Vision-Wireless ViWi-BT dataset [1]. We present how to predict future beam sequences from previously observed beam sequences and images using a long short-term memory (LSTM) network as a base predictive method. As such, we utilize the soft attention mechanism to intelligently choose the most important features and thus we suggest replacing the softmax attention function with different periodic attention functions to eliminate the gradient vanishing problem.

Keywords: Deep Learning, computer vision, wireless communication, beam prediction, long short-term memory, beam tracking

1. Introduction

Wireless data traffic has been increasing at a rate of over 50% per year per subscriber, and this trend is expected to accelerate over the next decade with the continual use of video and the rise of the Internet-of-Things (IoT) [3]. With the explosive growth of mobile data demand, the fifth generation (5G) mobile network would exploit the enormous amount of spectrum in the millimeter-wave (mmWave) [4]. Thus, is suggested by recent studies that mmWave frequencies could be used to augment the currently saturated 700 MHz to 28 GHz and 38 GHz radio spectrum frequencies for wireless transmission [5]. Even the fact that there exists huge available bandwdith at mmWave frequency bands makes it seem achieving ultra-high data rate communication, the features and characteristics of propagation of mmWave signals introduce new challenges at both the physical and network layers. One of the key challenges is how to reduce the latency for searching the best beamforming for the best performance for hand-over of mobile users. The other one is how to proactively hand over mobile users before blockages prevent mobile users from communicating, so as to avoid the latency of searching the best beamforming for the best performance.

In recent years, artificial intelligence (AI) manifestations of machine learning and deep learning have become an important area of research in video and image processing for camera surveillance (CS) applications and technologies that have become very common tools not only for public but also for private security purposes, such as traffic monitoring, manufacturing monitoring, reducing false liability claims, protection of building and assets, etc. In course of time, number of intelligent surveillance (IS) of cameras (ISC) is increasing, and this fact renders promising applications for various fields. With the rapid development of the usage of deep learning-based approaches, many researchers have proposed methods based on the learning characteristics of models about hidden features in scenes and segmented moving objects in video sequences. Based on the same notion behind ICS, with the aid of AI techniques exploiting images and video streams, several research directions are drawn attention by Alrabeiah et al. in [1] for solving main wireless communication problems in 5G-beyond networks, increasing the reliability of wireless networks using visual data.

To address the most promising challenges of applying deep learning for solving 5G problems several surveys are presented in the literature. Using deep learning and machine learning, the authors of [6] and [7] present a literature survey for searching the best beamforming in mmWave networks and provide the application list of comprehensive different deep learning models such as CNN, RNN, DRL, and LSTM. In more detail, [7] focuses on machine learning applications for handover management.

To exploit visual data either from CS or ICS to advance the research in a wider range of vision-wireless applications (machine learning tasks), a vision-aided wireless-beam tracking (ViWi-BT) dataset is introduced in [1]. This incorporates multiple base stations and mobile users or objects (cars, buses, and pedestrians) featuring rich dynamics. Further, presented in [1] is the ViWi-BT competition in which the main task is to predict the future mmWave beams of a user using the previously observed beams and RGB images. The research interest in vision-aided beam tracking became very attractive in the last years. In the ViWi-BT competition attended nine teams where the approaches of the participants are not publicly available, but the results achieved can be found on the official website [2]. The approaches that are being public but not as part of the competition can be summarized as follows. Firstly, in [1] is proposed the baseline solution for beam tracking and prediction using the GRU model. The main question in [1] was whether or not the visual data can improve the prediction results. In [8], Alrabeiah and Alkhateeb, investigated the problem of mmWave link blockage prediction and beam prediction. They proved that using neural network models that are deep enough can predict mmWave beams and blockages with a success probability close to one. The authors in [9], proposed a beam selection method using 3D reconstruction images using one camera to capture images that relies only on the environmental data. The implementation of a new framework using CNNs and RNN-based recurrent prediction network for dynamic link blockage prediction using beamforming and images provided by the ViWi-BT dataset was provided by Charan et al. in [10]. In addition, the authors in [11], presented new deep learning architecture which is composed from two components of CNN and GRU to solve the blockage-prediction problem streaming the predictions of blockage to the central unit to determine if the communication should be handed off to different camera for a particular user. They extended the ViWi-BT challenge dataset into two datasets presenting the blockage-prediction and object-detection datasets. Reus-Muns et al. utilized in [12] spatial information of mobile users such as location, speed, and surrounding scene images and proposed a method called channel covariance matrix to estimate the moving region containing all possible user locations at any given time and later they introduced deep learning based denoising method to reduce the error of the user locations not estimated accurately. In [13], Roy et al. presented a survey related to deep learning-based fusion framework leveraging the GPS location information with the combination of the visual data. Further, Salehi et al presented in [14] beam selection for mmWave links in a vehicular scenario by leveraging the data collected from sensors like LiDAR, camera images, and GPS. The key contribution is presenting the beam prediction framework using Fusion-based deep learning which can be executed locally or at mobile edge computing center (MEC) and presenting the beam sweeping running at mmWave band to select the best pair beam for establishing the link based on the suggested proposed optimization method. In [15], Tian et al. presented beam tracking approach based on different beam embedding, image recognition, image feature embedding, and sequential models. The key contribution is also the reformulated ViWi dataset where the images in training and validation are mutually exclusive. Hu and Han in [16], presented new algorithm to solve the ViWi beam tracking problem consisting of two base models of learning and fused learning model by shrinking the size of the images through bilinear scaling to ease the computation during the training.

In this paper, we are investigating a method of solving the ViWi-BT challenge by utilizing the soft attention mechanism. To the best of our knowledge, soft attention mechanism is not explicitly exploited in the literature so far for solving the beam tracking prediction task. Moreover, the paper could be summarized in the following points:

- Feature extraction module: The focus is to extract the features using ResNet-50 and 3D Resnext-101 CNN architectures from the visual data. The ResNet-50 is considered to be utilized for extracting the 2D features due to showing great success on feature extraction tasks.
- Capture the global information module: The aim is to leverage the extracted features and intelligently select the most important features extracted from the feature extraction module where the soft attention mechanism is playing an important role. Alternatively, we suggest replacing the softmax attention function with alternative attention functions, such as Taylor softmax, soft-margin softmax, and so on (See Section 2.2.3), to eliminate the gradient vanishing problem.
- Sequence generator module: The role of sequence generator module is to predict the future beam sequences by utilizing the merged features that were intelligently selected by soft attention mechanism and beam indices.
- We evaluate the results of our model with the baseline solution where we show an increment of accuracy by ~10%.

The remainder of this papers is organized as follows. In Section 2.1, the problem of predicting the future mmWave beams for ViWi-BT challenge is revisited and presented. By means of presenting the problem, in Section 2.2 we introduce the modules of the leveraged method as a solution to the presented challenge. Thereafter, in Section 2.3, we outline the utilized soft attention mechanism and suggest alternative attention functions to replace the softmax attention. Furthermore, in the Section 2.3, we provide detailed information of training and testing the leveraged method. Lastly, in Section 3, we obtain the results, and we provide primitive comparison with the baseline solution. Finally, we draw our conclusions and address further research directions.

2. Materials and Methods

The problem of vision-aided beam tracking and the challenge is briefly explained in [1]. In the following section, we will shortly revisit the problem of the challenge and in a later section, we investigate the methods for the vision-aided beam tracking problem with leveraging visual data.

2.1. Problem definition

The core challenge for ViWi-BT is to proactively predict the future beam sequences based on previously observed beam sequences and RGB images. Authors of the ViWi-BT dataset presented different scenarios on their official website, but the scenario of our interest is with two small-cell mmWave base-stations. The two base stations are assumed to be equipped with a mmWave uniform linear array antenna operating in 28G Hz band and three RGB cameras each [2]. These equipped cameras are considered to record the movement of the objects in the environment containing the eight pairs of observed images and corresponding beam indices. The problem can be defined as in Eq. (1):

$$S_u(t) = \left[\left(f_u(t-\tau+1), X_u(t-\tau+1)\right), \dots, \left(f_u(t), X_u(t)\right)\right] \quad (1)$$

where the f_u is the beamforming vector in the codebook $\mathcal{F}$ used to present the user μ at time instance m. The X_u is a 3D tensor representing the RGB images captured by the camera. Given the sequence of beams $S(t)$, the aim is to predict the best beams in the next m time instances, denoted by $\hat{f}(t') = t+1 \dots, t+m$, where m in the challenge defines a three-tier prediction task 1, 3, and 5. Each record in the dataset consists of an observed sequence and five ground truth future beam indices with the corresponding RGB image under the sense of maximizing the received signal-to-noise-ratio (SNR) at the user. The function $\hat{f}(t')$ it is not expected to be customized for a single user, but it should be capable to predict m set of any user in the given wireless environment, whether it is a line of sight (LOS) or non-line of sight (NLOS) user since no localization information is provided.

2.2. Deep learning framework

The leveraged framework is presented in Figure 1, composed by feature extraction module consisting of ResNet-50 [17] and 3D ResNext-101 [18], capture the global information module consisting cross gating strategy [19] and soft attention mechanism [20], and sequence generator module [21]. The modules will be introduced individually in the following subsections.

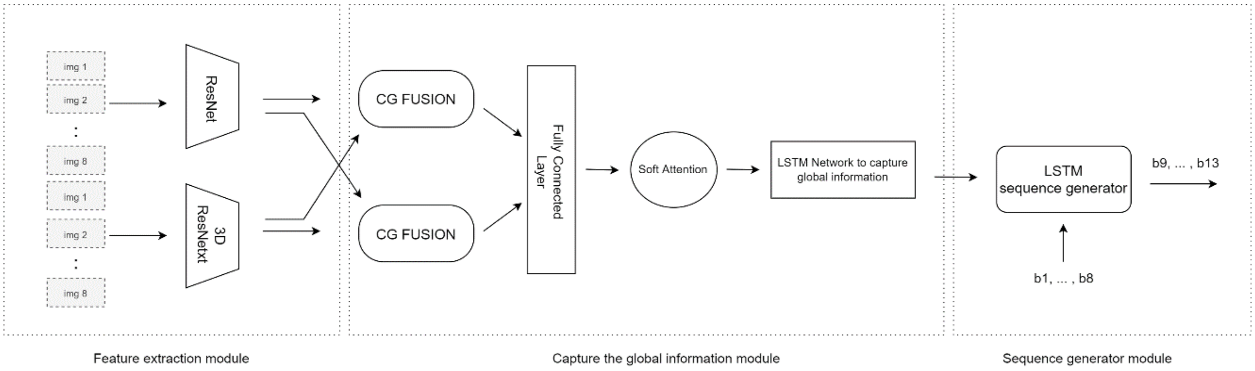


Figure 1 Architecture of the leveraged deep learning model [19]

2.2.1. Feature extraction module

Recently, the research has proved that extracting multiple image features and feeding the extracted features to a predictive model or text generation model can play an important role to improve the accuracy [22, 23, 24]. Taking this into account, in the framework the feature extraction module is used to extract the features of the images before feeding the wireless data to the model. Every image contains the information of moving user and the information of environment such as various objects, buildings, trees, cars, and trucks. Even though it is not known which user is of interest in the dataset of the challenge, the feature extraction of the image is proved be helpful for the predictive model.

ResNet-50 and 3D ResNext-101 are shown to be powerful feature-extractors models due to having large number of layers. Training these models from scratch requires long time and a lot of resources would be occupied during the training process. In the framework, we used ResNet-50 and 3D ResNext-101 model pretrained on the Kinetics dataset to extract spatiotemporal features from the images [25]. The architecture of the ResNet-50 and 3D ResNext-101 is provided in the Figure 2.

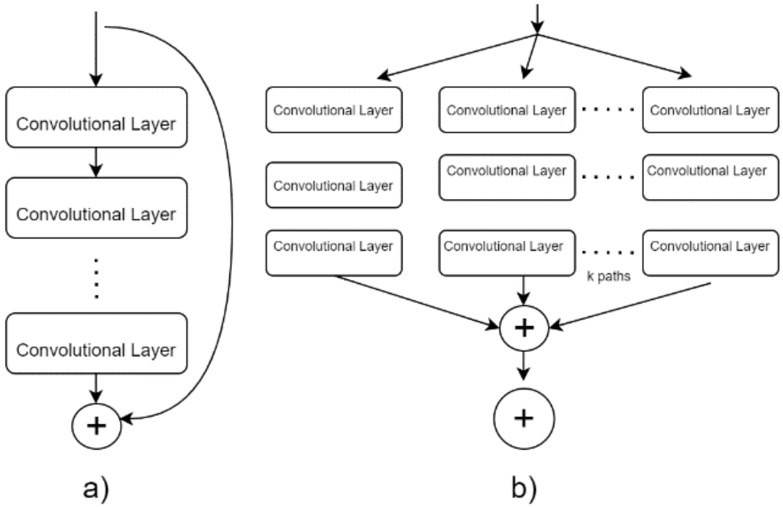


Figure 2 ***a)*** *A block representing the architecture of ResNet-50 [17].* ***b)*** *A block of 3D ResNeXt-101 with cardinality = k [18]*

ResNet-50 is proved to be the most successful architecture in image classifications [18]. ResNet-50 provides shortcut connections as presented in Figure 2-a, which allows a signal to bypass one layer and move to the next layer in sequence. Since these connections are passed through the networks, this results in the ability to train much deeper networks. 3D ResNext-101 is inherited from ResNet-50, which introduces cardinality where cardinality presents the size of the set of transformations. Cardinality k in Figure 2-b presents different dimensions from deeper and wider architectures which introduces group of convolutions dividing the feature maps into small groups. 3D ResNext-101 architecture helps to capture spatiotemporal 3D features from the images. The parameters of the architecture of ResNet-50 and 3D ResNext-101 used to extract features is shown in Table 1.

2.2.2. Capture the global information module

Capture the global information module is used to fuse the extracted features from ResNet-50 and 3D ResNext-101 models. The gate fusion network is performed on two stages as it is provided in [19]. The first stage is to aggregate the extract features by using LSTM networks whereas the second stage is to capture the relationships of the fused frames since fusing the frames from the first stage will not capture the relationships. The illustration of cross gating strategy is provided in Figure 3.

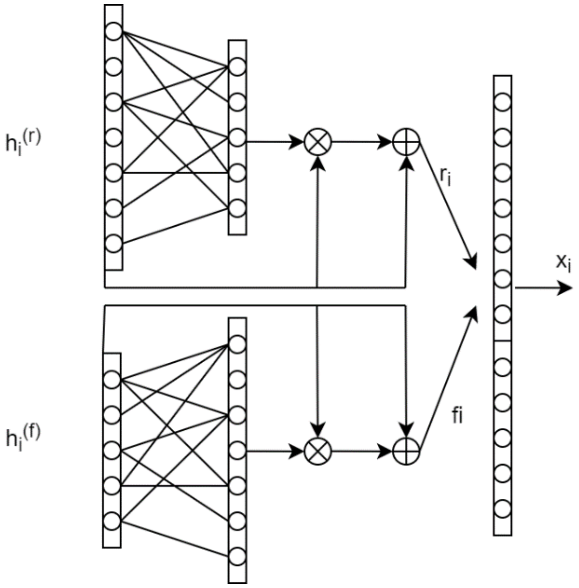


Figure 3 Illustration of the cross-gating network proposed in the gated fusion network. The cross gating strategy strengthens the information that is related to each other and fuses them together [19]

Table 1 Network architectures used to extract visual data features. Each convolutional layer is followed by batch normalization [26] and ReLU [27]. F is the number of feature channels and N is the number of blocks in each layer. conv1 represents the operation of spatially down-sampling the inputs and conv2_x, conv3_x, conv4_x, conv5_x represents the convolutional layer

Model	conv1	*conv2_x*		*conv3_x*		*conv4_x*		*conv5_x*	
		F	*N*	*F*	*N*	*F*	*N*	*F*	*N*
ResNet-50	Conv, 7 x 7 x 7, 64, Temporal stride 1, Spatial stride 2	64	3	128	{4, 4, 4, 24}	256	{6, 23, 36, 35}	512	3
ResNeXt-101		128	3	256	24	512	36	{512,896}	{16,32}

The LSTM networks used to aggregate the feature representations are shown by $h_i^{(r)}$ and $h_i^{(f)}$ which represents the hidden states and memory cells.

The cross-gating strategy can make sure that the obtained ResNet-50 and 3D ResNext-101 features are used efficiently for the corresponding semantic information by multiplication and summation operations. The r_i and f_i are the gated results for Resnet-50 and 3D ResNext-101 extracted features.

$$\begin{aligned} r_i &= \mathrm{Gating}_r^{(E)}\left(h_i^{(f)}, h_i^{(r)}\right), \\ f_i &= \mathrm{Gating}_r^{(E)}\left(h_i^{(f)}, h_i^{(r)}\right), \end{aligned} \tag{2}$$

where the Gating function is represented as follows:

$$\mathrm{Gating}(x, y) = \sigma(wx + b)y + y, \tag{3}$$

where y denotes the target feature, which is updated by the features of x. The w and b represents the learnable parameters and σ represents the ReLU activation function. After strengthening the content information by the LSTM network and cross gating strategy, the gated representations of ResNet-50 and 3D ResNext-101 are fused together by a fully connected layer presented as follow:

$$x_i = w^{(E)}\left([r_i, f_i] + b^{(E)}\right), \tag{4}$$

where x_i denotes the fused features from CNN networks, $w^{(E)}$ *and* $b^{(E)}$ represents the learnable parameters.

The LSTM network to capture global information is presented in Figure 4.

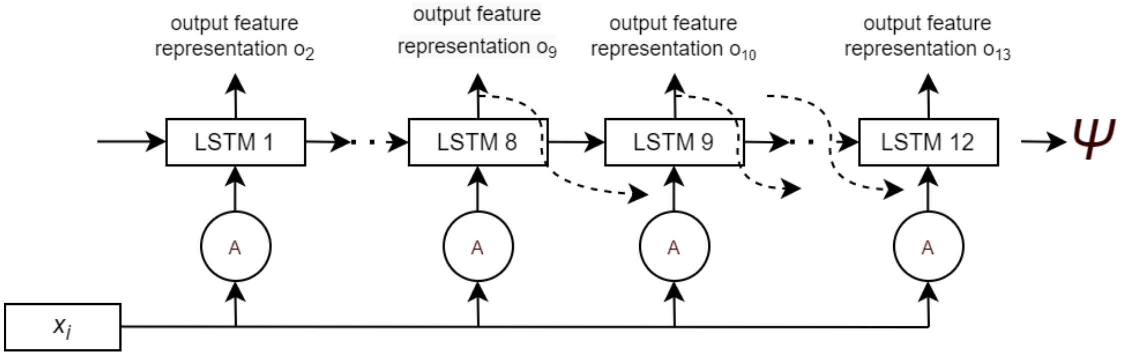

output feature representation o$_2$
output feature representation o$_9$
output feature representation o$_{10}$
output feature representation o$_{13}$
LSTM 1
LSTM 8
LSTM 9
LSTM 12
Ψ
A
A
A
A
x_i

Figure 4 The LSTM network to capture the global information from ResNet-50 and 3D ResNext-101 extracted features. A denotes the soft attention mechanism

The sequence of captured features will be based on the fused representations $X = \{x_1, x_{2,} \ldots x_m\}$. The LSTM cell presents the predicted sequence to capture the global information and it is depicted as follow:

$$\{h_t, z_t\} = \mathrm{LSTM}\left(\left[E_{(o-1)}, \phi_t(X, h_{t-1})\right], h_{t-1}\right), \tag{5}$$

$$Pr(c_o|c_{<t}, I; \theta) = \text{softmax}(Wh_t\ +\ b),$$

where h_t and z_t denote the hidden state and memory cell, respectively. The E denotes the matrix for extracted feature representation tags and the $E_{(o-1)}$ presents the embedding vector of extracted features. θ, W and b presents the learnable parameters by the network. $\Pr(c_o|c_{<t}, V; \theta)$ denotes the probability of predicting the correct feature presentation tags c_t given the previous tags $c_{<t} = \{c_1, c_2 \dots c_{t-1}\}$ and input image sequence I.

The ϕ_t symbol in Eq. (5) denotes the soft attention mechanism based on [20] which yields a vector representation with different weights on X:

$$\phi_t(X, h_{t-1}\,) = \sum_{i=1}^{m} a_{t,i} x_i, \tag{6}$$

where $\sum_{i=1}^{m} a_{t,i} = 1$ and $a_{t,i}$'s is computed at each time step t and presents the attention weights. The attention weights reflect the relevance of the fused representations extracted from ResNet-50 and 3D ResNext-101 to select the most useful information predicted by LSTM at the current step and returns the unnormalized relevance score $e_{t,i}$:

$$e_{t,i} =\ w^T \tanh(Wh_{t-1} + Ux_i + b)\,, \tag{7}$$

where w, W, U and b are the learnable parameters. Once the relevance scores $e_{t,i}$ are computed for all the images $i = 1, \dots, n$, utilizing the softmax attention, we normalize the relevance scores to obtain $a_{t,i}$'s:

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{m} \exp(e_{t,i})}. \tag{8}$$

The attention mechanism allows the LSTM network to focus on only important subset of frames by increasing the attention weights of the corresponding temporal feature. The Ψ symbol in Figure 4 denotes the last hidden state and it is expected to capture the global information from ResNet-50 and 3D ResNext-101 extracted features and later used to guide the sequence generator LSTM network to predict the future beam indices.

2.2.3. Soft attention mechanism

The present section is dedicated to the importance of the soft attention mechanism for capturing the global information from extracted features of the images and suggesting different attention functions to replace Eq. (8).

The reason of not directly feeding extracted features to the sequence generator model is to not treat the whole image equally. The attention mechanism instead of using the image as an input, it considers paying attention to the particular areas or objects of the interest related to the input image as it is provided in Eq. (6), Eq. (7) and Eq. (8).

The attention softmax equation presented in Eq. (8), when used in attention mechanisms is not normally distributing the relevance scores which leads to a gradient vanishing problem making the training difficult [28]. Hence, we are considering replacing the softmax attention to alleviate the gradient problem and thus providing more accurate importance to the features. The suggestions can be listed as below:

- **Taylor softmax:** The Taylor softmax function is proposed by [29] and uses the second order Taylor series approximation, that is

$$a_{t,i} = \frac{1 + e_{t,i} + 0.5e_{t,i}^2}{\sum_{k=1}^{m}\left(1 + e_{t,k} + 0.5e_{t,k}^2\right)}, \tag{9}$$

- **Soft-margin softmax:** Soft-margin (SM) softmax reduces intra-class distances but enhances inter-class discrimination, by introducing a distance margin into the logits [29]:

$$a_{t,i} = \frac{\exp\left(e_{t,i} - \beta\right)}{\sum_{k\neq i}^{m} \exp\left(e_{t,k}\right) + \exp\left(e_{t,k} - \beta\right)}, \tag{10}$$

where β is manually set, and when β is set to zero, SM-Softmax simplifies to the original softmax.

- **Sin-Max-Constant/Cos-Max:** This attention function is a periodic function, and it can be defined as below [28]:

$$a_{t,i} = \frac{1 + \sin\left(e_{t,i}\right)}{d + M + \sin\left(e_{t,i}\right)}, \quad \text{and} \quad M = \sum_{j\neq i}^{m} \sin\left(e_{t,i}\right). \tag{11}$$

- **Sin2-Max-Shifted:** This function is defined as below [28]:

$$a_{t,i} = \frac{\sin^2\left(e_{t,i}\right)}{M + \sin^2\left(e_{t,i}\right)}, \quad \text{and} \quad \mathrm{M} = \sum_{\mathrm{j}\neq \mathrm{i}}^{\mathrm{m}} \sin^2\left(e_{t,j}\right). \tag{12}$$

- **Sin-Softmax:** This attention function is defined as below [28]:

$$a_{t,i} = \frac{e^{\sin\left(e_{t,i}\right)}}{M + e^{\sin\left(e_{t,i}\right)}}, \quad \text{and} \quad M = \sum_{j\neq i}^{m} e^{\sin\left(e_{t,j}\right)}. \tag{13}$$

- **Siren-max:** This function is defined as below [28]:

$$a_{t,i} = \frac{\dfrac{1 + \sin\left(e_{t,i}\right)}{2 - 2\sin\left(e_{t,i}\right)}}{M + \dfrac{1 + \sin\left(e_{t,i}\right)}{2 - 2\sin\left(e_{t,i}\right)}}, \quad \text{and} \quad M = \sum_{j\neq i}^{m} \frac{1 + \sin\left(e_{t,j}\right)}{2 - 2\sin\left(e_{t,j}\right)}. \tag{14}$$

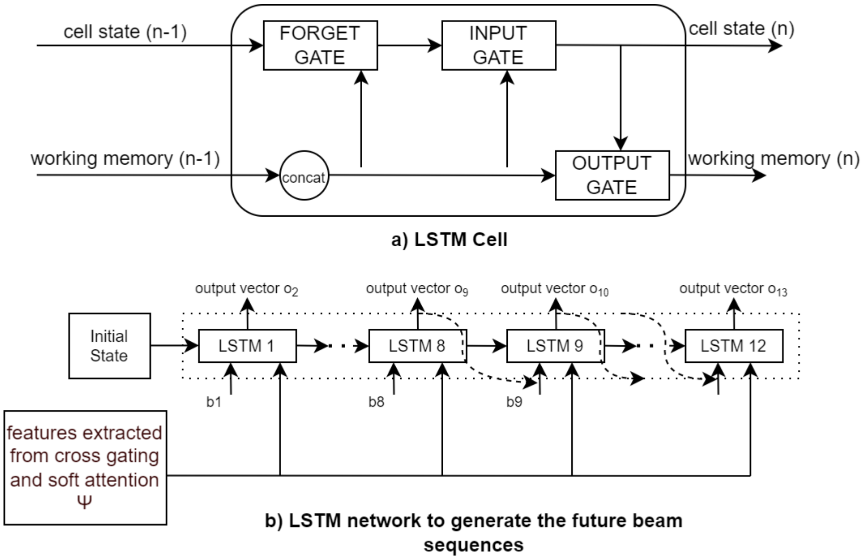
2.2.4. Sequence generator module

In the ViWi-BT challenge it is required to predict 1, 3, and 5 future beams by previously observed beam sequences. The LSTM network [30] has shown great success

on tasks that contains time-series data such as prediction and text generation and thus it is considered in this paper to be used as predictive model.
In Figure 5, we are presenting an LSTM cell and LSTM network for the prediction task.

The main components of an LSTM cell are the forget gate, input gate, and the output gate. The cell state is the core idea of the LSTM cell, and the role is to act as a long-term memory, persevering the information under all iterations of the node. The cell node can decide to remove unnecessary information or to decide if the information is important and thus to keep it. The role of the forget gate is to help cell state to decide which information to remove from the cell state. This operation is completed by performing calculations on the concatenated input values and applying this operation to the cell state. Input gate based on the concatenated input values decides what information is important and what should be included to the cell state. The output gate responsibility is to decide what the next hidden state should be based on calculations on the current cell state and the concatenated working input value.

The LSTM network to generate the sequence of beams consists of twelfth LSTM cells. The inputs of the LSTM cell are initial state, merged feature vector from capture the global information module, cell state (current state), working memory (hidden state) where the forget gate, input gate, output gate tries to decide which information should be kept or removed to provide better future predictions. In Figure 5, b represents the beam index $B = \{b_1, b_{2,} \dots b_{12}\}$ which are given as an input to every LSTM cell where every LSTM cell outputs new long-term and short-term memory which the last outputs $o_9 \dots o_{13}$ are considered as the prediction of the beam sequence.



a) LSTM Cell

b) LSTM network to generate the future beam sequences

Figure 5 The representation of a) LSTM cell and b) the LSTM network to generate future beam sequences.

2.3. Training and testing

In this section, we introduce the process of training and evaluating the deep learning framework.

To train the deep learning framework, firstly we input and utilize the eight images. These images are given as an input to the pretrained ResNet-50 and 3D ResNext-101 to help the model to obtain the visual and motion features such as location, motion, and blockage information. For each sample, we extract 2048-dimensional visual and 8192-dimensional features. The second step is to merge the extracted features from ResNet-50 and 3D ResNet-101 by using the capture the global information module. After merging the extracted features, the model forms a 463-dimensional vector. The third step is to input the output of the capture the global information module to each LSTM cell. The merged feature vectors of the first twelve beam indices visit every LSTM cell to update the hidden state of each LSTM cell and generate the output vectors. The last step is to train the network by using the twelve output vectors to calculate the training loss using the ground truth values.

The testing process of the proposed framework is similar to the training process. The second step differs in the way that the merged vectors of the first eight beam indices visit the first eight beam indices from the first to seventh LSTM cells to update the hidden states. The eighth and twelfth LSTM cells help to obtain the prediction of the future beam indices. The last step of the training process is skipped and not used for the testing process. Due to the training process, we optimize the model by using the Adam optimizer [31] and the learning rate is set to $4\ x\ 10^{-4}$ reduced by half on every eight epochs. The training process consists of the batch size of 256 and the loss function is set to the cross-entropy loss.

3. Results

In this section, we conduct experiment to evaluate the performance of the framework on ViWi-BT dataset. We aim to investigate the effect of the visual data on training process and compare the framework with the baseline solution for future beam predictions. The experiments are completed under very limited set of the hardware on PyTorch environment using one NVIDIA 1060 6GB GPU.

The ViWi-BT dataset contains the training set denoted by $\mathcal{D}_t$ of 281100 user instances, the validation dataset denoted by $\mathcal{D}_v$ of 120468 user instances, and testing dataset denoted by $\mathcal{D}_{test}$ of 10000 user instances. Each instance is represented by $S_u(t)$ which contains the 8 pair of the beam indices and corresponding images of the street view. The camera images in $S_u(t)$ contains the appearance of the target user and the appearance of different object such as cars, trucks, buildings, trees, etc. The localization of the target user is not provided in $\mathcal{D}_t$, $\mathcal{D}_v$, and $\mathcal{D}_{test}$ so the appearance of the target user can happen in every possible t instance. Moreover, the first eight beam indices and corresponding camera images are the observed information for the target user and the last five beam indices are the groundtruth data containing the future beam indices and corresponding camera images. In the conducted experiment the aim is to find a prediction function $f_{\odot}(S_u(t))$ to maximize the probability of the generated five beam indices comparing with the last five beam indices by providing as an input the first eight beam indices and corresponding camera images.

In this section we also compare our framework with the baseline solution [1] on validation dataset by using the top-1 accuracy and exponential decay score metrics.
The top-1 accuracy can be defined in Eq. (9) as it presented in [1].

$$\mathrm{Acc}_{\mathrm{top-1}}^{(\mathrm{n})} = \frac{1}{\mathrm{U}} \sum_{\mathrm{i=1}}^{\mathrm{U}} \mathbb{1}\left\{\mathrm{p}_{\mathrm{i}}^{(\mathrm{m})} = \mathrm{t}_{\mathrm{i}}^{(\mathrm{m})}\right\} \tag{15}$$

where U is the number of the instances in the validation dataset, $\mathbb{1}\{\cdot\}$ presents the indicator function with value of 1 only when the condition is met, $t_i^{(m)}$ is the i^{th} target beam sequence which includes the correct future beams, and $p_i^{(m)}$ is the sequence of the predicted future m beams.
The exponential decay score metric is defined in Eq. (10) as it is presented in [1].

$$\mathrm{score}^{(\mathrm{m})} = \frac{1}{\mathrm{U}} \sum_{\mathrm{i=1}}^{\mathrm{U}} \mathrm{e}^{-\frac{\left\|\mathrm{p}_{\mathrm{i}}^{(\mathrm{m})} - \mathrm{t}_{\mathrm{i}}^{(\mathrm{m})}\right\|_1}{\mathrm{m}\sigma}} \tag{16}$$

where the $\|\cdot\|_1$ presents the first norm, and σ is a penalization factor that is set to 0.5. Our comparison with the baseline solution [1] are illustrated in Table 2.

Table 2 The comparison of our framework with the baseline solution

	Top-1 Accuracy			**Exponential Decoy Score**		
	1 future beam	*3 future beams*	*5 future beams*	*1 future beam*	*3 future beams*	*5 future beams*
Our leveraged method	0.8989	0.7043	0.6205	0.8229	0.7492	0.6909
Baseline method [1]	0.85	0.60	0.50	0.86	0.68	0.60

The results from top-1 accuracy shows that our method outperforms the baseline solution [1] on 1 feature beam, 3 feature beams, and 5 future beams predictions. The overall score of the proposed model is increased by ~0.10 and thus we can say that leveraging the visual data it can highly help the framework to achieve better results.

The results from exponential decay score surprisingly showed worse results on prediction for the next 1 future beam. This performance failure can open future discussions and challenges related to that if the visual data can really help the model to achieve better results if we use the visual data for predicting 1 future beam. The results from exponential decay on predicting the 3 feature beams, and 5 feature beams, yields better results than the baseline solution [1]. These results are expected since the feature extraction module can start to be more powerful at future time instances and help the model to learn the features behind the environment that came from the image and thus provide more accurate predictions.

Overall, observing the top-1 accuracy and the exponential decay score, we conclude that leveraging visual data to extract environment features of the location and blockage of the target user, and features of the environment details, in combination with good predictive network of LSTM shows superior performances for beam tracking and predictions tasks.

4. Discussion and Conclusion

This paper presents soft attention as an additional mechanism to deep learning to perform mmWave beam predictions by previously observed beam indices and images using different feature extraction techniques and LSTM network as predictive model. The experimental results conducted on ViWi-BT dataset show that leveraging both wireless and visual data for beam prediction tasks can help to achieve better prediction accuracies than using the wireless data only.

Our current work has several possible extensions. First, introducing extra concepts in data preprocessing like data cleaning and prefetching can help the framework to achieve better results. Second, clustering the visual data according to their link LOS or NLOS-like environment can help us to conduct better training strategies. Third, we observe that considering different feature extraction models may help the framework to extract more important features related to the user trajectories and the environment. Consequently, considering better network architectures like encoder-decoder models to replace with the gated fusion and sequence generator module may yield better results.

5. Acknowledge

This publication was made possible by Tachyonic Telecommunications Electronics Computer Technologies Inc. The statements made herein are solely the responsibility of the author[s].

References

1. Alrabeiah, M., Booth, J., Hredzak, A., & Alkhateeb, A. (2020). Viwi vision-aided mmwave beam tracking: Dataset, task, and baseline solutions. arXiv preprint arXiv:2002.02445.
2. The official website of the vision-aided beam tracking data competition at IEEE ICC 2022: https://www.viwi-dataset.net/viwi-bt.html.

3. Rappaport, T. S., Xing, Y., MacCartney, G. R., Molisch, A. F., Mellios, E., & Zhang, J. (2017). Overview of millimeter wave communications for fifth-generation (5G) wireless networks—With a focus on propagation models. IEEE Transactions on antennas and propagation, 65(12), 6213-6230.
4. Niu, Y., Li, Y., Jin, D., Su, L., & Vasilakos, A. V. (2015). A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges. Wireless networks, 21(8), 2657-2676.
5. Rappaport, T. S., Sun, S., Mayzus, R., Zhao, H., Azar, Y., Wang, K., ... & Gutierrez, F. (2013). Millimeter wave mobile communications for 5G cellular: It will work!. IEEE access, 1, 335-349.
6. Ly, A., & Yao, Y. D. (2021). A review of deep learning in 5G research: Channel coding, massive MIMO, multiple access, resource allocation, and network security. IEEE Open Journal of the Communications Society, 2, 396-408.
7. Mollel, M. S., Abubakar, A. I., Ozturk, M., Kaijage, S. F., Kisangiri, M., Hussain, S., ... & Abbasi, Q. H. (2021). A survey of machine learning applications to handover management in 5G and beyond. IEEE Access, 9, 45770-45802.
8. Alrabeiah, M., & Alkhateeb, A. (2020). Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels. IEEE Transactions on Communications, 68(9), 5504-5518.
9. Xu, W., Gao, F., Jin, S., & Alkhateeb, A. (2020). 3D scene-based beam selection for mmWave communications. IEEE Wireless Communications Letters, 9(11), 1850-1854.
10. Charan, G., Alrabeiah, M., & Alkhateeb, A. (2021, June). Vision-aided dynamic blockage prediction for 6G wireless communication networks. In 2021 IEEE International Conference on Communications Workshops (ICC Workshops) (pp. 1-6). IEEE.
11. Charan, G., Alrabeiah, M., & Alkhateeb, A. (2021). Vision-aided 6G wireless communications: Blockage prediction and proactive handoff. IEEE Transactions on Vehicular Technology, 70(10), 10193-10208.
12. Reus-Muns, G., Salehi, B., Roy, D., Jian, T., Wang, Z., Dy, J., ... & Chowdhury, K. (2021, December). Deep Learning on Visual and Location Data for V2I mmWave Beamforming. In 2021 17th International Conference on Mobility, Sensing and Networking (MSN) (pp. 559-566). IEEE.
13. Roy, D., Salehi, B., Banou, S., Mohanti, S., Reus-Muns, G., Belgiovine, M., ... & Chowdhury, K. (2022). Going Beyond RF: How AI-enabled Multimodal Beamforming will Shape the NextG Standard. arXiv preprint arXiv:2203.16706.
14. Salehi, B., Reus-Muns, G., Roy, D., Wang, Z., Jian, T., Dy, J., ... & Chowdhury, K. (2022). Deep Learning on Multimodal Sensor Data at the Wireless Edge for Vehicular Network. arXiv preprint arXiv:2201.04712.
15. Tian, Y., & Wang, C. (2021, September). Vision-Aided Beam Tracking: Explore the Proper Use of Camera Images with Deep Learning. In 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall) (pp. 01-05). IEEE..
16. Hu, Z., & Han, C. (2021, October). Image and index fused sequence-to-sequence algorithm for vision-aided millimeter-wave beam tracking. In Proceedings of the 5th ACM Workshop on Millimeter-Wave and Terahertz Networks and Sensing Systems (pp. 7-12).
17. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

18. Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

19. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., & Liu, W. (2019). Controllable video captioning with pos sequence guidance based on gated fusion network. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2641-2650).

20. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In Proceedings of the IEEE international conference on computer vision (pp. 4507-4515).

21. Tian, Yu, Gaofeng Pan, and Mohamed-Slim Alouini. "Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges." IEEE Open Journal of the Communications Society 2 (2020): 132-143.

22. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence-video to text. In Proceedings of the IEEE international conference on computer vision (pp. 4534-4542).

23. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

24. Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1933-1941).

25. Hara, K., Kataoka, H., & Satoh, Y. (2017). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? arXiv preprint. arXiv preprint arXiv:1711.09577.

26. Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR.

27. Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

28. Wang, S., Liu, F., & Liu, B. (2021). Escaping the Gradient Vanishing: Periodic Alternatives of Softmax in Attention Mechanism. IEEE Access, 9, 168749-168759.

29. Banerjee, K., Gupta, R. R., Vyas, K., & Mishra, B. (2020). Exploring alternatives to softmax function. arXiv preprint arXiv:2011.11538.

30. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

31. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980