

Multi-Modal Trajectory Prediction of NBA Players

Sandro Hauri¹ Nemanja Djuric² Vladan Radosavljevic³ Slobodan Vucetic¹
¹Temple University ²Uber ATG ³Spotify
 {sandro.hauri, nemanja, vladan, vucetic}@temple.edu

Abstract

National Basketball Association (NBA) players are highly motivated and skilled experts that solve complex decision making problems at every time point during a game. As a step towards understanding how players make their decisions, we focus on their movement trajectories during games. We propose a method that captures the multi-modal behavior of players, where they might consider multiple trajectories and select the most advantageous one. The method is built on an LSTM-based architecture predicting multiple trajectories and their probabilities, trained by a multi-modal loss function that updates the best trajectories. Experiments on large, fine-grained NBA tracking data show that the proposed method outperforms the state-of-the-art. In addition, the results indicate that the approach generates more realistic trajectories and that it can learn individual playing styles of specific players.

1. Introduction

In recent years, advances in artificial intelligence and computer vision started revolutionizing how athletic performance and results are being analyzed and understood, which includes the use of fine-grained player tracking data during sporting events. In our research we focus on developing new methods aimed at deeper understanding of the behavior of athletes in team sports, with a particular focus on their motion prediction. This is a particularly important task in invasion sports, such as soccer, football, or basketball, where knowledge of how and where the players will move, especially when it comes to those from the opposing team, is of critical importance for gaining a tactical advantage during the game [20]. Beyond this use case the benefits of accurate motion prediction extend to other applications, such as postgame analysis [12] or improving TV broadcasting of games by optimizing camera movement [4, 17]. Prediction of human trajectories can also be used to improve tracking accuracy [18], and has recently become a vibrant topic of research in the computer vision community [1, 8, 9, 14].

Using mathematics, statistics, and artificial intelligence to analyze sports performance is not a novel idea. It has been famously explored in baseball [23] and applied with great success to soccer [19], with authors uncovering useful patterns that have been used to move the needle in this highly competitive field. As these advanced tools have been proven successful in practice, statistical analysis has been adopted by top-performing teams regardless of the sport they play. Today, elite teams from across the globe, such as Golden State Warriors, New York Yankees, and Manchester United, have analytics departments focusing on deriving knowledge from large amounts of data these teams generate. Beyond the sports professionals, even the general public is becoming more accepting of these complex statistical tools, as exemplified by the introduction of the concept of expected goals [26] in some postgame summaries in the Premier League, the English top soccer division. This trend is also exemplified by a number of research publications, as well as high-profile conferences and workshops organized on the topic, such as MIT Sloan SAC or KDD Sports Analytics [3]. These are attended by both the scientific community, world-class athletes and management of professional sports teams, indicating the value that the artificial intelligence is bringing to this multi-billion dollar industry.

In this paper we focus on movement prediction of NBA players during offensive possessions and we assume that at any moment players have freedom to consider several options for their movement. The trajectories depend on the state of a possession, which includes positions and current trajectories of the players and the ball, as well as on individual player preferences. To predict the trajectories, we propose an uncertainty-aware, multi-modal deep learning model. The model is trained to predict multiple player trajectories and probabilities that they will be selected. Figure 1 shows an example of such trajectories and their probabilities, compared to baseline models. We provide an in-depth discussion of Figure 1 in the Results section, and evaluate the proposed method using player tracking data collected during several months of an NBA season. We showcase that with our proposed training regime, the model has the ability of recreating distinct playing styles of individual players.

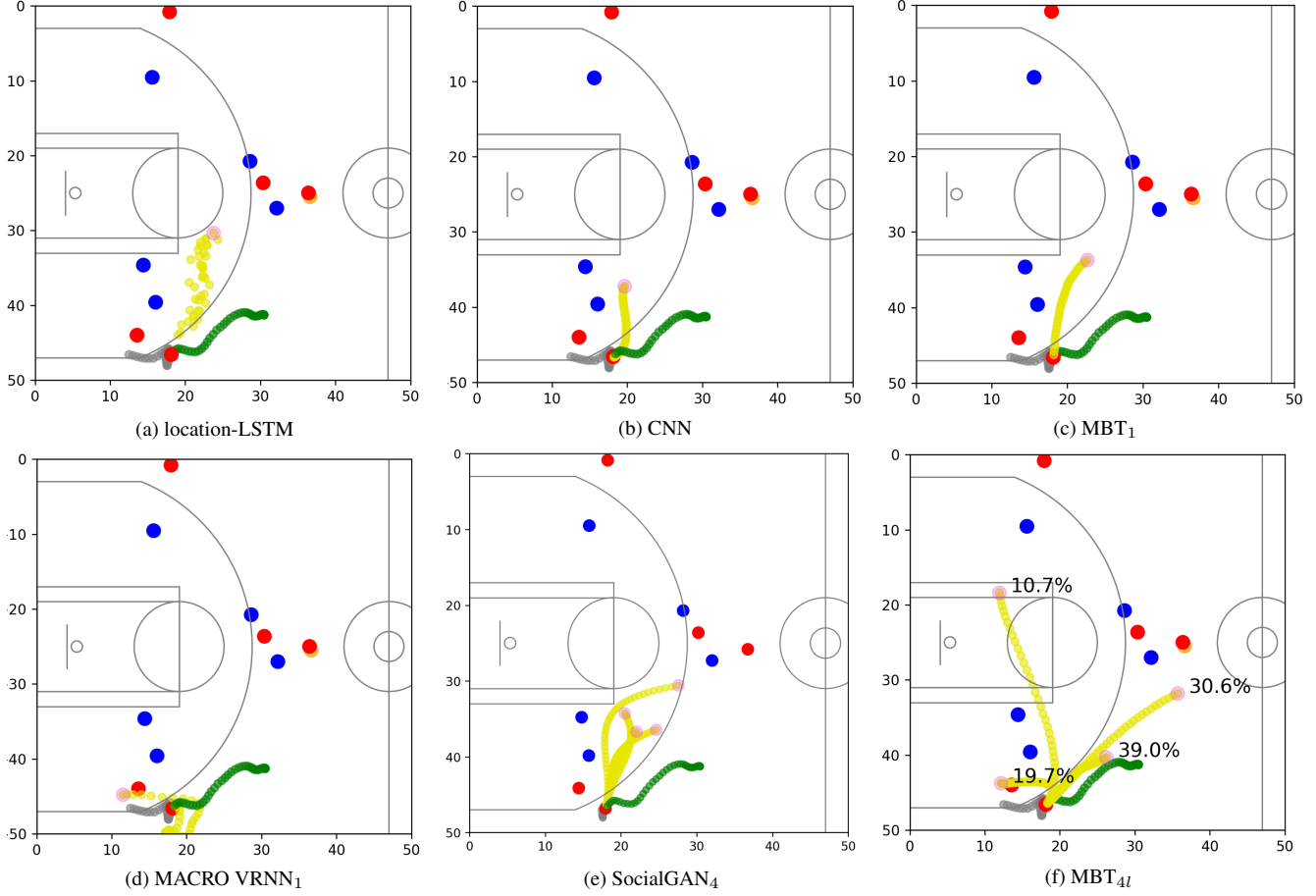


Figure 1: Visualization of predicted trajectories with $H = 40$ using several state-of-the-art methods: a) location-LSTM, b) CNN, c) MBT_1 , d) $MACRO\ VRNN_1$, e) $SocialGAN_4$, f) MBT_{4l} (ours), red: attackers, blue: defenders, orange: ball, grey: input history of predicted player, yellow: prediction, green: ground truth; a video animation is included in the Supplementary Material

2. Related Work

Modeling and predicting human trajectories is an important challenge in a number of scientific areas. Researchers have worked on this problem to develop realistic crowd simulations [24], or to improve vehicle collision avoidance systems [16] through predicting future pedestrian movement. When it comes to traffic applications, pedestrian behavior was usually modeled using attracting and repulsive forces to guide them towards a goal, while simultaneously avoiding obstacles. Human pedestrian prediction was also used to improve accuracy of tracking systems [6, 25, 31] or to study intentions of individuals or groups of people [5, 21, 30]. The advances in deep learning led to data-driven methods, such as Long Short-Term Memory (LSTM) networks [15] with shared hidden states [1], multi-modal Generative Adversarial Networks (GANs) [13], or inverse reinforcement learning [18], outperforming the traditional methods. The work by [13] is particularly related to our study, through its use of a multi-modal loss function and by showing practical

benefits of multi-modal trajectory prediction as compared to single trajectory predictions. Beyond pedestrian movement, recent research on predictive modeling of vehicular trajectories for self-driving car applications also contains ideas of relevance for the current study. In particular, [7] showed that multi-modal trajectory predictions for vehicles generate realistic real-world traffic trajectories. The multi-modal loss function in our approach is inspired by this work, where we adapt ideas from the self-driving domain to modeling of movement of basketball players.

The ubiquitous use of tracking systems in professional sports leagues like the NBA or the English Premier League inspired researchers to analyze and model trajectories of athletes during matches. For example, [9] used Variational Autoencoders (VAEs) to model real-world basketball data and showed for NBA data that the offensive player trajectories are less predictable than the defense. The authors of [22] and [28] used LSTM to predict near-optimal defensive positions for soccer and basketball, respectively. [29] similarly used variants of VAEs to generate trajectories for NBA

players. NBA player trajectory predictions are also studied by [32] and [33], where a deep generative model based on VAE and LSTM and trained with weak supervision was proposed to predict trajectories for an entire team. Macro-intents for each player were inferred, where the players target a spot on the court they want to move to. The authors evaluate the model mostly by human expert preference studies and show they can outperform the baselines, indicating that RNNs can capture information from observational data in sports. However, their trajectories are usually not smooth and no restrictions are set on the position of a player on consecutive time steps, such that the model may output physically unrealistic trajectories. We consider this state-of-the-art approach in our experiments, and show that it is outperformed by the proposed multi-modal method.

3. Methodology

3.1. Problem Setting

Recent advancements in optical tracking have made it possible to track the players and the ball during an NBA game with good enough accuracy and temporal resolution to recreate the trajectories of all ten players and the ball during an entire basketball game. This allows us to extract 2-D location $\ell_t^p = [x_t^p, y_t^p]$ of player p at time step t , with $p \in \{1, \dots, 10\}$, as well as 2-D location of the ball at time t , $\ell_t^b = [x_t^b, y_t^b]$, where x -coordinate represents the length of the field while the y -coordinate represents the width, with the origin at the upper left corner (see Figure 1 for illustration). Using an ordered sequence of previous $L + 1$ time steps we can generate historical trajectory of the p -th player as $\mathbf{h}_t^p = [\ell_{t-L}^p, \dots, \ell_t^p]$, where time steps are equally spaced at an interval of Δ_t . Similarly, we can generate a historical trajectory of the ball as $\mathbf{h}_t^b = [\ell_{t-L}^b, \dots, \ell_t^b]$. As a convention, we will assume that the first 5 players represent the team on the offense and the last 5 players the team on the defense. We are interested in predicting future trajectory of p -th offensive player, represented as a vector $\tau_t^p = [\ell_{t+1}^p, \dots, \ell_{t+H}^p]$, where H is the number of future time steps (or horizon) for which we predict the trajectory. We will assume that the *player of interest* (i.e., the offensive player for which we are predicting future trajectory) is denoted by player index P .

In this paper, we processed the raw tracking data to create labeled data set $\mathcal{D} = \{(\mathbf{u}_t^P, \tau_t^P), t = 1, \dots, T, P = 1, \dots, 5\}$, where one data point is defined for each time step and each offensive player (as indicated by the range $P = 1, \dots, 5$). Here T is the total number of time steps, input vector $\mathbf{u}_t^P = \{\mathbf{h}_t^P, \mathbf{h}_t^{-P}, \mathbf{h}_t^b, s_t\}$ is a set of historical player and ball trajectories, where \mathbf{h}_t^P indicates history of the player of interest, \mathbf{h}_t^{-P} indicates histories of all other 9 players, and s_t is the shot clock defined as the time in seconds remaining until the shot clock expires. Note that

in the input vector the history of the player of interest P always comes first, followed by histories of their 4 teammates and then by 5 opposing players, ordered by a distance to the player of interest. Output vector τ_t^P is a future trajectory of the player of interest P computed at time step t , and objective is to build a predictor that accurately predicts their trajectory given inputs \mathbf{u}_t^P . We emphasize that, in addition to the given inputs, there are other features that potentially might influence the observed trajectories, such as game clock, home vs. away, foul calling, previous plays, or player mismatch. As we demonstrate with the shot clock feature, our approach allows for a straightforward use of any additional feature that a modeller may deem important. However, an in-depth feature analysis is out of scope of this paper, and instead we focus on showing viability of the proposed multi-modal predictive model. In fact, it could be argued that a number of such features are implicitly present in the input representation already. For example, if a team has a large point lead with little game time remaining, they may slow down on the offense and the observed movement history could capture that information.

Lastly, note that an alternative to predicting a sequence of H future locations of the offensive player is predicting a sequence of their velocities. As we know the current location at time t , we can convert trajectory τ_t^P to a velocity vector $\nu_t^P = [\mathbf{v}_{t+1}^P, \dots, \mathbf{v}_{t+H}^P]$ using a direct mapping of velocities to locations, computed for horizon $h \in \{1, \dots, H\}$ as

$$\mathbf{v}_{t+h}^P = [v_{x,t+h}^P, v_{y,t+h}^P] = \left[\frac{x_{t+h}^P - x_{t+h-1}^P}{\Delta_t}, \frac{y_{t+h}^P - y_{t+h-1}^P}{\Delta_t} \right]. \quad (1)$$

Although trajectories and velocity vectors are mathematically interchangeable, a particular choice might have a significant impact on model training. As we will demonstrate experimentally, predicting the next location is more challenging due to the issue in normalization of coordinates.

3.2. Proposed Approach

As noted previously [32], movement of basketball players is inherently multi-modal as the players can decide between multiple plausible trajectories at any given time (e.g., to move towards the basket for a layup or towards a corner for a three-point attempt). In order to account for this multi-modality we train a predictive model that generates output $\mathbf{o}_t^P = [\hat{\nu}_{t,1}^P, \dots, \hat{\nu}_{t,M}^P, \hat{p}_{t,1}^P, \dots, \hat{p}_{t,M}^P]$, which consists of M predicted trajectories $\hat{\nu}_{t,m}^P$ representing M modes, as well as M scalars $\hat{p}_{t,m}^P$ representing probabilities that a corresponding mode is selected by a player. This results in $(2H + 1)M$ output values, since output for each mode consists of a trajectory comprising H 2-D locations and an additional mode probability.

3.2.1 Loss function

Given a ground-truth trajectory ν and predicted trajectory $\hat{\nu}$, we first define the trajectory loss as

$$\mathcal{L}^{\text{MSE}}(\nu, \hat{\nu}) = \frac{1}{2H} \|\nu - \hat{\nu}\|_2^2, \quad (2)$$

defined as a mean squared error (MSE) of the predicted velocity vector. Then, in order to train a model to predict multiple trajectories and their probabilities, we base our approach on an adaptation of the multi-modal loss function presented in [7]. A similar loss function is used by [13] to generate multi-modal pedestrian trajectories within a GAN framework. In particular, we define the Multiple-Trajectory Prediction (MTP) loss for time step t and player P , comprising a linear combination of classification loss $\log \hat{p}_m$ and trajectory loss (2),

$$\mathcal{L}^{\text{MTP}} = \sum_{m=1}^M \delta_\epsilon(m = m^*) \left(\log \hat{p}_m + \alpha \mathcal{L}^{\text{MSE}}(\nu_t^P, \hat{\nu}_{t,m}^P) \right), \quad (3)$$

where \hat{p}_m is an output of a softmax, α is a hyper-parameter used to trade-off the classification and trajectory losses, and m^* is the index of the winning mode that produced the trajectory closest to the ground truth, computed according to a distance function $\text{dist}()$ defined in the next subsection,

$$m^* = \arg \min_{m \in \{1, \dots, M\}} \text{dist}(\nu_t^P, \hat{\nu}_{t,m}^P). \quad (4)$$

Moreover, δ_ϵ is a relaxed Kronecker delta [27] giving the most weight to the best matching trajectory, but also a small weight to the remaining ones,

$$\delta_\epsilon(\text{cond}) = \begin{cases} 1 - \epsilon, & \text{if condition } \text{cond} \text{ is true,} \\ \frac{\epsilon}{M-1}, & \text{otherwise.} \end{cases} \quad (5)$$

Intuitively, the classification loss in (3) forces the probability of the winning mode to 1 (thus pushing probabilities of other modes towards zero due to the softmax), and trajectory loss penalizes prediction error of the winning mode.

We note that [7] used the unrelaxed Kronecker delta (i.e., ϵ was set to 0), which only updates the closest trajectory. In practice, this leads to problems where a randomly initialized path is much worse than the remaining paths. Such poorly initialized modes never get selected through (4) and do not get a chance to improve during training. To prevent this issue we use the relaxed Kronecker delta, where we start from some small value of ϵ that is gradually reduced towards 0 as the training progresses. This phenomenon is well known in generative models and is commonly referred to as mode collapse in GANs or posterior collapse in VAEs. Comparable annealing remedies have been proposed in VAEs [2], but are generally not sufficient to achieve good performance [11]. Our approach was more stable than VAE or GAN training, and we will empirically show that we can outperform state-of-the-art models based on each of those two methods.

3.2.2 Distance functions

As mentioned previously, m^* denotes a trajectory closest to the ground truth, however there are different closeness measures that can be considered. For example, in [13] the closest mode is defined simply as a path with the lowest trajectory loss, computed as

$$\text{dist}_{\text{MSE}}(\nu, \hat{\nu}_m) = \mathcal{L}^{\text{MSE}}(\nu, \hat{\nu}_m). \quad (6)$$

We also considered other distance functions, as [7] concluded that its choice has a large impact on the model performance. Thus, we considered distance function with the smallest overall displacement error, defined as a location error at the last time step and computed as

$$\text{dist}_l(\nu, \hat{\nu}_m) = \left\| \sum_{h=1}^H (\nu_{t+h} - \hat{\nu}_{t+h,m}) \right\|_2. \quad (7)$$

Lastly, we considered using the error of final player velocity (which can be interpreted as player's "heading"), shown in earlier work [7] to be beneficial,

$$\text{dist}_v(\nu_t, \hat{\nu}_{t,m}) = \|\nu_{t+H} - \hat{\nu}_{t+H,m}\|_2. \quad (8)$$

3.2.3 Model architecture

While [7] use the multi-modal loss function to train a CNN model, we will show that on the NBA data LSTM network is more effective. We use a two-layer LSTM architecture, each with a width of 128, to encode the time-series input of recently observed data \mathbf{u}_t^P . The encoder is a fully connected layer and the prediction consists of M trajectories of a single player given as x - and y -velocities for H future time steps, as well as M probabilities that the player will follow the respective trajectory.

Because players differ in their positions, skills, heights, and weights, we would expect them to run at different speeds and along different paths. To take these differences into account, we consider a two-stage training approach to learn specific per-player models. To this end we first train the proposed model on data taken from all players to learn the average behavior of all NBA players. In the second training phase these pre-trained networks can be used to initialize a specialized per-player network fine-tuned on a subset containing only that player's data, so that individual behavior of the player can be learned. In the experiments we evaluate both global and per-player models.

We refer to the proposed multi-modal approach as Multi-modal Basketball Trajectories (MBT). We evaluate different number of modes M and investigate different distance functions in (4), indicating these choices in the subscript. In particular, we denote model variants as MBT_{Md} , with $d \in \{\text{MSE}, l, v\}$, corresponding to (6), (7), and (8), respectively. For example, MBT_{4l} generates 4 paths and uses distance function (7) during training. When using a single mode the distance measure is not used, and we refer to the uni-modal model as MBT_1 .

4. Experiments

4.1. Experimental setting

4.1.1 Data set

We used publicly available movement data collected from 632 NBA games during the 2015-2016 season¹, from which we extracted 114,294 offensive possessions. An offensive possession starts when all players from one team cross into the opponent’s half court, and ends when the first offensive player leaves the half court or the game clock is paused. Possessions shorter than 3s were discarded, resulting in 113,760 possessions. This amounts to 1.1 million seconds of gameplay where player location is captured every 0.04s. We downsampled the data by a factor of 3 to obtain sampling rate of $\Delta_t = 0.12s$, corresponding to a lower bound on human reaction time [10] during which velocity is considered constant. Furthermore, we randomly split the data into train and test sets using 90/10 split. All inputs and outputs were normalized to the $[-1, 1]$ range. To train the specialized networks that predict specific player’s movement we extracted possessions featuring that player. The amount of data for each player is in the order of several thousands (e.g., for Stephen Curry there were 2,767 possessions).

4.1.2 Model training

As discussed previously, we used a 2-layer LSTM with 128 channels in each layer. To learn the general model for all NBA players we trained LSTM in batches of 1,024 samples. The learning rate in Adam optimizer was set to $5 \cdot 10^{-4}$. We set hyper-parameter α in equation (3) to 1, such that the amplitude of the two losses are about equal, and ϵ in (5) to 0.25 which was reduced by a factor of 0.05 per epoch until $\epsilon = 0$. We used ℓ_2 regularization with the weight of $\lambda = 0.001$ and an early stopping mechanism to further prevent overfitting. To specialize the neural network for a specific player we fine-tune the base model on data from that player and adjust the hyper-parameters as follows. We start with $\epsilon = 0.75$ which is reduced by a factor of 0.01 per epoch to make sure that all modes benefit from the information contained in this smaller training set. The initial learning rate in this case was reduced to 10^{-5} .

All training was done on a single computer with Nvidia GeForce GTX 1080 card. It took approximately 60 minutes to train the base model, while specializing the network on a specific player took less than 5 minutes.

4.1.3 Accuracy measures

We report common measures used in pedestrian trajectory prediction, *final displacement error* (FDE) and *average dis-*

placement error (ADE) [1, 13], defined as

$$\begin{aligned} \text{FDE} &= \frac{1}{5T} \sum_{t=1}^T \sum_{P=1}^5 \left\| \ell_{t+H}^P - \hat{\ell}_{t+H}^P \right\|_2 \\ \text{ADE} &= \frac{1}{5HT} \sum_{t=1}^T \sum_{P=1}^5 \sum_{h=1}^H \left\| \ell_{t+h}^P - \hat{\ell}_{t+h}^P \right\|_2. \end{aligned} \quad (9)$$

In other words, FDE considers the location error at the end of the prediction horizon H , while ADE averages location errors over the entire trajectory. We also report MSE error, defined as in equation (2). Unlike FDE and ADE that measure trajectory prediction errors, MSE is a measure of how accurately are the velocities predicted.

To evaluate multi-modal approaches we calculate the metrics for each output trajectory and only choose the path that has the smallest FDE, which is consistent with evaluation procedure commonly used in the literature [13, 27].

4.1.4 Baselines

To establish an upper bound for the proposed error measures we compared our method to a straw-man baseline. Constant velocity (CV) baseline assumes that the player keeps moving in the last observed direction with constant speed.

Baseline CNN refers to an approach that transforms the input to a rasterized trace image and uses a CNN encoder (instead of LSTM) before predicting the future velocities [7]. For the encoder, we used 5 layers with depths [64, 128, 128, 64, 32], 5×5 mask, "same" padding, and 2×2 max pooling. The decoder consisted of 2 densely connected layers with sizes 128 and 64.

To compare different output alternatives we trained the same LSTM architecture used for our model to directly predict player locations, as opposed to predicting velocities. We refer to this model as location-LSTM. We also considered SocialGAN [13], the state-of-the-art in human trajectory prediction. This approach uses an LSTM-based generator, coupled with a social pooling layer to account for nearby actors. We trained this model using the code made available by its authors², using the same NBA data set except that SocialGAN can not use extra information such as ball location or shot clock, therefore only the players trajectories are used. GANs are notoriously hard to train, which resulted in a training time of 28 hours for 50 epochs of training. In addition, we considered the state-of-the-art MACRO VRNN [32], which uses programmatic weak supervision to first predict a location that the player wants to reach and then uses a Variational RNN (VRNN) to predict a trajectory that the player will take to reach it. MACRO VRNN also accounts for the multi-modality of the problem, with the number of generated paths denoted in the subscript. We

¹<https://github.com/sealneaward/nba-movement-data>, last accessed November 2020; we are not associated with the data creator in any way.

²<https://github.com/agrimgupta92/sgan>, last accessed November 2020.

Table 1: Comparison of various models, input steps L , and modes M in terms of error metrics ADE and FDE (in feet) and MSE (in ft^2/s^2)

Method	L	M	$H = 10$			$H = 20$			$H = 40$		
			ADE	FDE	MSE	ADE	FDE	MSE	ADE	FDE	MSE
CV	1	1	1.72	3.92	9.09	4.64	10.97	16.01	11.59	26.14	20.59
CNN	10	1	2.76	5.25	15.80	5.28	9.99	17.48	8.15	13.23	21.95
location-LSTM	10	1	1.61	2.98	10.21	3.43	6.91	15.94	6.79	12.11	29.80
MBT ₁	10	1	1.43	2.98	7.26	3.32	6.92	12.36	6.59	11.97	16.93
MBT ₁	20	1	1.40	2.93	7.25	3.30	6.91	12.41	6.59	11.97	16.74
MBT ₁	30	1	1.39	2.92	7.46	3.33	6.91	12.32	6.58	11.92	16.87
SocialGAN ₁	10	1	1.25	2.75	8.18	3.09	6.67	13.32	6.47	12.35	17.54
MACRO VRNN ₁	10	1	1.70	3.43	13.17	4.46	8.66	19.85	8.48	14.98	25.03
SocialGAN ₄	10	4	1.19	2.61	7.36	2.95	6.33	11.91	6.19	11.54	15.76
MACRO VRNN ₄	10	4	1.07	1.98	5.90	3.14	5.07	11.93	6.40	8.54	19.29
MTP _{4l}	10	4	1.44	2.87	7.91	3.08	6.14	11.36	5.78	10.06	13.52
MBT _{4MSE}	10	4	1.01	1.91	3.82	2.33	4.00	6.35	5.25	6.92	12.46
MBT _{4v}	10	4	1.05	1.93	4.00	2.66	4.31	7.75	6.71	8.74	14.72
MBT _{4l}	10	4	1.01	1.90	3.82	2.33	4.04	6.35	4.89	6.39	11.56

used models provided in [32] trained on roughly the same amount of data. Note that training takes up to 20 hours, as opposed to only 1 hour for our proposed method. Finally, we compare to Multiple-Trajectory Prediction (MTP) [7] which resembles our approach, but instead uses an unrelaxed Kronecker delta (i.e., ϵ was set to 0) and the distance measure from equation (7).

4.2. Results

We first compare the performance of models trained on data containing all possessions, with results across different error measures and time horizons presented in Table 1.

The CV model, which assumes the player will keep moving with the last observed velocity, gives relatively small errors for short time horizons, but deteriorates quickly for longer time horizons. The CNN model outperforms this simple baseline at longer horizons, while the performance is suboptimal at short horizons. Location-LSTM is comparable to MBT₁ model in terms of ADE and FDE metrics, with much worse MSE metric. As we will demonstrate later in qualitative results, this difference in MSE can be explained by the fact that location-LSTM produces trajectories that are not physically achievable by the players.

Next we experiment with the uni-modal MBT₁ model and evaluate the influence of different lengths of historical inputs L . Based on the results we confirm that the MBT₁ models only marginally improve with longer input sequences. As a result, in the remainder of the experiments we use a value of $L = 10$, consistent with [32].

In the following experiment we compare different distance functions used for training MBT methods, where we keep M fixed at 4. We see that the choice of distance function has limited effect on accuracy measures at a shorter horizon of 1.2s. However, as the horizon increases, MBT_{4l} starts outperforming the competing approaches by a con-

siderable margin. Taking this result into account, in further experiments we use the distance function defined in (8).

When we compare the proposed method to the state-of-the-art models MACRO VRNN and SocialGAN, we separate the analysis by comparing the same number of modes. When evaluating a single trajectory, SocialGAN outperforms both our approach and MACRO VRNN in ADE and FDE. However, MBT₁ reaches better MSE than those approaches. When comparing multiple modes, we see that MBT_{4l}, MBT_{4v} and MBT_{4MSE} performance is roughly comparable at shorter horizons, but MBT_{4l} outperforms all other methods across all accuracy measures at longer horizons. Quite notably, MBT_{4l} outperforms the baselines with a large margin in terms of MSE velocity measure. For example, for horizon $H = 40$, our MBT_{4l} model achieves ADE 24% and 21% smaller than MACRO VRNN₄ and SocialGAN₄, respectively. The comparison to MTP_{4l} shows problems arising from using an unrelaxed Kronecker delta during the training process. Observations of the generated paths reveals that some modes are collapsed or not all have a non-zero probability, as the poorly initialized paths are not trained at all.

In Figure 1 we illustrate predicted trajectories for a randomly picked player. Trajectories are generated using a single-path model that predicts locations (location-LSTM, Figure 1a), two single-path models that predicts velocities, one based on a CNN architecture (Figure 1b) and one based on an LSTM architecture (MBT₁, Figure 1c), one sample path of MACRO VRNN (Figure 1d), 4 sampled paths of SocialGAN (Figure 1e), and our proposed method using 4 modes MBT_{4l} (Figure 1f). We can see that location-LSTM output is noisy and does not represent realistic player movements. Player trajectories predicted by the CNN and MBT₁ model are smoother and more realistic, showing the advantage of predicting velocities instead

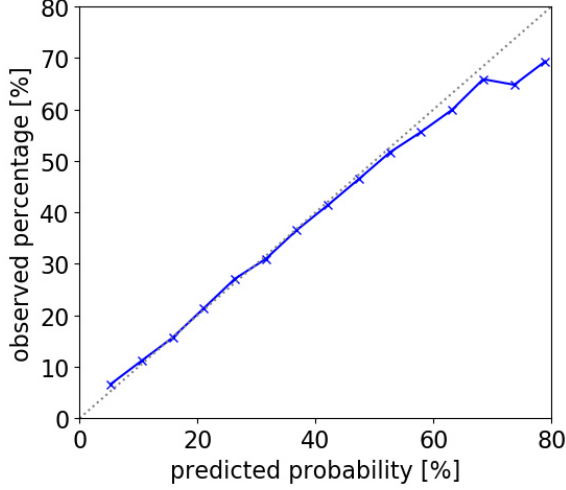


Figure 2: Evaluation of predicted mode probabilities for MBT_{4l}

of locations. While CNN and MBT_1 generate qualitatively similar results, MBT_1 outperforms CNN in the quantitative measures. MACRO VRNN generally produces paths that are less smooth than competing models, explaining the high error in MSE as discussed above. The multiple paths predicted by SocialGAN are smooth and look plausible, but lack the diversity of movement that we would expect in basketball trajectories. MBT_{4l} predicts 4 paths that are very distinct from each other. The highest-probability path ends up very close to the observed final player location, while accurately following the ground-truth trajectory. Other paths produced by the multi-modal model allow for diverse movements, such as an aggressive drive to the basket or supporting the ball-handling teammate near the center of the court.

We also evaluate the quality of inferred mode probabilities produced by the MBT_{4l} model. To this end we compare predicted mode probabilities to empirical ones, computed as a frequency of how often a mode of certain probability had the lowest FDE. We bucketed inferred probabilities in 5% bins and for each computed the empirical probability, with the average per-bucket results presented in Figure 2. We can see that the plot closely follows the identity line, indicating that the predicted mode probabilities are well-calibrated.

To evaluate the hypothesis that the MBT trajectories are more physically realistic, we calculate acceleration of predicted trajectories on the test set. The maximum acceleration of MBT_{4l} is $12.2m/s^2$. We note that the ground truth contains noisy outliers, with accelerations of up to $600m/s^2$ (the 99.9th percentile is $14.5m/s^2$). In contrast, when considering MACRO VRNN we observe accelerations of more than $500m/s^2$ (the 99.9th percentile is $54.86m/s^2$). This indicates that in many cases the baseline trajectories are far from being physically achievable, while the proposed method yielded more realistic outputs.

Table 2: Prediction of specific players with and without fine-tuning for $H = 40$ (4.8 seconds) using the MBT_{4l} model

Player	Fine-tuned?	ADE	FDE	MSE
LeBron James	No	4.78	6.63	9.97
LeBron James	Yes	4.67	6.24	9.91
Stephen Curry	No	6.32	7.80	17.35
Stephen Curry	Yes	6.09	7.51	16.62
Russell Westbrook	No	5.49	7.15	12.43
Russell Westbrook	Yes	5.36	6.90	12.23
DeAndre Jordan	No	4.36	6.01	12.20
DeAndre Jordan	Yes	3.93	4.94	12.56
Andrew Bogut	No	4.54	6.12	9.34
Andrew Bogut	Yes	4.29	5.40	9.03

4.2.1 Evaluation of per-player models

In this section we compare per-player models to the base model trained on all players, as well as the per-player models fine-tuned on players that are playing in the same position, but are known to have distinct playing styles. We first compare the performance of the base and per-player models for several example players, with results presented in Table 2. The per-player models result in improved performance across the board, as they are better capturing playing styles of individual players.

Let us consider a specific game situation where center DeAndre Jordan just set up a pick and roll, shown in Figure 3 and in the animated video in the Supplementary Material. The model trained on all players predicts that the so-called roll man will now move either towards the basket or towards the wide open space on the right-hand side of the court, shown in the first row of Figure 3a. Jordan is a very dynamic and fast center who executes many successful pick and rolls, so our model trained on his data predicts he will drive to the basket faster and with a higher probability than an average player in the same situation, as shown in Figure 3b. We also compare to a model trained on data of Andrew Bogut, a defense specialist who is not as fast as Jordan. According to stats.nba.com³, Bogut only attempts 0.5 pick and rolls per game, while Jordan attempts 2.4. Our model correctly predicts Bogut’s paths to be less dynamic and gives a 25% probability that he would turn around and focus on defending a counter attack, entirely relying on his team mate to capitalize on the pick, as shown in Figure 3c.

The following experiment involves a situation where Stephen Curry has possession of the ball at the top of the circle with a defender to his right, as illustrated in Figure 4 (and in the Supplementary Material). This example shows some limitations of our approach because in actuality Curry first acts like he wants to drive inside, but decides to stop and shoot the ball for a 2-pointer before starting to move backwards. The predicted trajectories are much simpler, but still capture some interesting options that the player may

³<https://on.nba.com/2ulXVau>, last accessed November 2020.

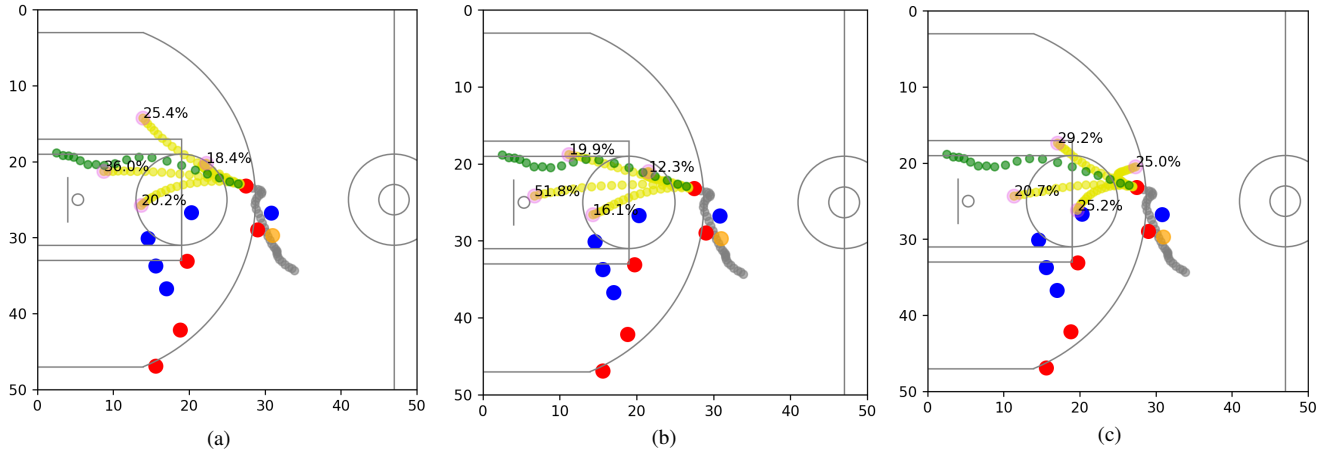


Figure 3: Visualization of predicted trajectories for DeAndre Jordan with $H = 20$ (2.4s) using 3 different networks MBT_{4l} : a) trained on all players, b) retrained with the data of DeAndre Jordan and c) retrained with the data of Andrew Bogut

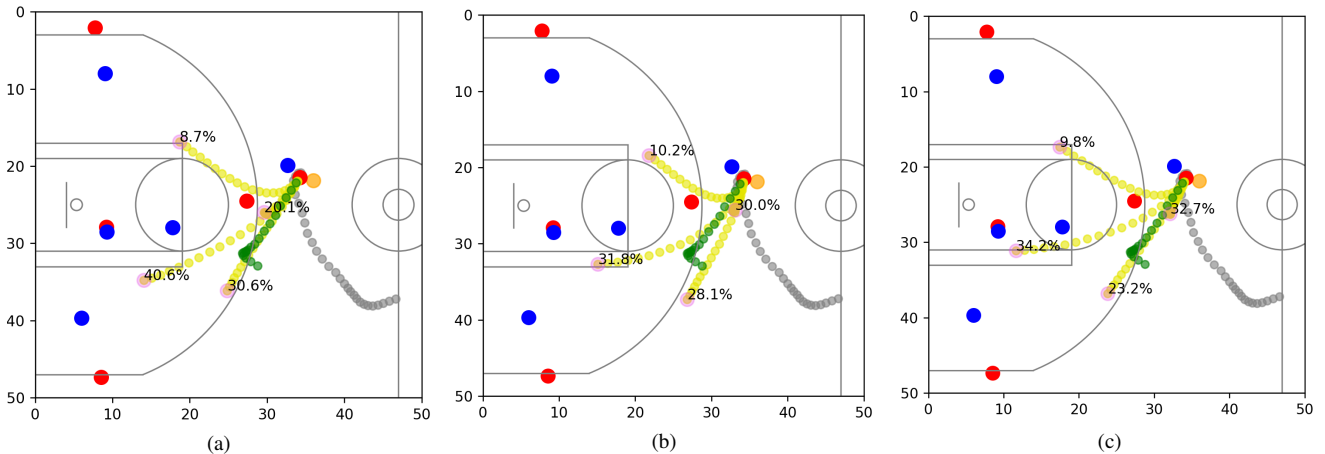


Figure 4: Visualization of predicted trajectories for Stephen Curry with $H = 20$ (2.4s) using 3 different networks MBT_{4l} : a) trained on all players; b) retrained with the data of Stephen Curry; c) retrained with the data of Russel Westbrook

choose. The model that was trained on all players predicts that the player may move towards the basket with about 40% probability as seen in Figure 4a, with other lower-probability options to move along the arc, stay at the top of the arc, or try to circle around the defender. The model that was retrained on data of Stephen Curry shown in Figure 4b slightly adjusts the path along the arc, because Curry often tries to shoot 3-pointers (more specifically, he had the second-most 3-point attempts in the 2015/16 season). As a result the model also gives him a lower probability to drive towards the basket. We evaluate the same situation with a network fine-tuned on data of Russell Westbrook, shown in Figure 4c. Westbrook attempts much fewer 3-pointers than Curry, and instead has more 2-point attempts. He is also a very dynamic player that is excellent at driving to the basket, such that when he makes an attempt he usually gets closer to the basket than an average player would. Thus, when he moves along the arc our model predicts that he

will not stay behind the 3-point line, but will instead try to get closer to the basket. We can see the model successfully managed to capture characteristics of individual players, adjusting the predictions to their own playing styles.

5. Conclusion

In this paper we proposed an LSTM-based model trained using multi-modal loss that can generate multiple paths which accurately predict movement of NBA players. In addition, we showed that per-player fine-tuning can capture interesting and specific behavior of different players. The proposed approach outperformed the state-of-the-art by a large margin, both in terms of standard prediction metrics and velocity error that better captures trajectory realism. As future work, we are exploring ideas to model the multi-modal behavior of the entire team, as well as opponent's strategies that can counter such trajectories.

References

- [1] A Alahi, K Goel, V Ramanathan, A Robicquet, L Fei-Fei, and S Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.
- [2] S R Bowman, L Vilnis, O Vinyals, A M Dai, R Jozefowicz, and S Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.
- [3] U Brefeld and A Zimmermann. Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, 31(6):1577–1579, 2017.
- [4] J Chen, H M Le, P Carr, Y Yue, and J J Little. Learning online smooth predictors for realtime camera planning using recurrent decision trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4688–4696, 2016.
- [5] S Choi, Wand Savarese. Understanding collective activities of people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013.
- [6] W Choi and S Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012.
- [7] H Cui, V Radosavljevic, F Chou, T Lin, T Nguyen, T Huang, J Schneider, and N Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [8] N Djuric, V Radosavljevic, H Cui, T Nguyen, F-C. Chou, T-H. Lin, N Singh, and J Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.
- [9] P Felsen, P Lucey, and S Ganguly. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [10] B Fischer and E Ramsperger. Human express saccades: extremely short reaction times of goal directed eye movements. *Experimentelle Hirnforschung. Expérimentation cérébrale*, 57:191–5, 02 1984.
- [11] H Fu, C Li, X Liu, F Gao, A Celikyilmaz, and L Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.
- [12] J Gudmundsson and M Horton. Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, 50(2):22, 2017.
- [13] A Gupta, J Johnson, L Fei-Fei, S Savarese, and A Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.
- [14] S Haddad and S Lam. Self-growing spatial graph networks for pedestrian trajectory prediction. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1151–1159, 2020.
- [15] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.
- [16] C G Keller and D M Gavrila. Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, 2013.
- [17] K Kim, M Grundmann, A Shamir, I Matthews, J Hodgins, and Irfan Essa. Motion fields to predict play evolution in dynamic sport scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 840–847. IEEE, 2010.
- [18] K M Kitani, B D Ziebart, J A Bagnell, and M Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.
- [19] S Kuper. *Soccernomics: Why England Loses, Why Spain, Germany, and Brazil Win, and Why the US, Japan, Australia and Even Iraq Are Destined to Become the Kings of the World’s Most Popular Sport*. Nation Books, 2014.
- [20] L Lamas, J Barrera, G Otranto, and C Ugrinowitsch. Invasion team sports: strategy and match modeling. *International Journal of Performance Analysis in Sport*, 14(1):307–329, 2014.
- [21] T Lan, Y Wang, W Yang, and G Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in neural information processing systems*, pages 1216–1224, 2010.
- [22] H Le, P Carr, Y Yue, and P Lucey. Data-driven ghosting using deep imitation learning. 03 2017.
- [23] M Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.
- [24] N Pelechano, J M Allbeck, and N I Badler. Controlling individual agents in high-density crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 99–108. Eurographics Association, 2007.

- [25] S Pellegrini, A Ess, and L Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010.
- [26] R Pollard, J Ensum, and S Taylor. Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. *Int. J. Soccer Sci.*, 2, 01 2004.
- [27] C Rupprecht, I Laina, R DiPietro, and M Baust. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [28] T Seidl, A Cherukumudi, A T Hartnett, P Carr, and P Lucey. Bhostgusters : Realtime interactive play sketching with synthesized nba defenses. 2018.
- [29] C Sun, P Karlsson, J Wu, J B Tenenbaum, and K Murphy. Stochastic prediction of multi-agent interactions from partial observations. *arXiv preprint arXiv:1902.09641*, 2019.
- [30] D Xie, T Shu, S Todorovic, and S Zhu. Learning and inferring “dark matter” and predicting human intents and trajectories in videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1639–1652, 2017.
- [31] K Yamaguchi, A C Berg, L E Ortiz, and T L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011.
- [32] E Zhan, S Zheng, Y Yue, L Sha, and P Lucey. Generating multi-agent trajectories using programmatic weak supervision. In *International Conference on Learning Representations*, 2019.
- [33] S Zheng, Y Yue, and P Lucey. Generating long-term trajectories using deep hierarchical networks, 2017.