

PROACTIVE: Self-Attentive Temporal Point Process Flows for Activity Sequences

Vinayak Gupta
IIT Delhi
New Delhi, India
vinayak.gupta@cse.iitd.ac.in

Srikanta Bedathur
IIT Delhi
New Delhi, India
srikanta@cse.iitd.ac.in

ABSTRACT

Any human activity can be represented as a temporal sequence of actions performed to achieve a certain goal. Unlike machine-made time series, these action sequences are highly disparate as the time taken to finish a similar action might vary between different persons. Therefore, understanding the dynamics of these sequences is essential for many downstream tasks such as activity length prediction, goal prediction, *etc.* Existing neural approaches that model an activity sequence are either limited to visual data or are task-specific, *i.e.*, limited to next action or goal prediction. In this paper, we present PROACTIVE, a neural marked temporal point process (MTPP) framework for modeling the continuous-time distribution of actions in an activity sequence while simultaneously addressing three high-impact problems – next action prediction, sequence-goal prediction, and *end-to-end* sequence generation. Specifically, we utilize a self-attention module with temporal normalizing flows to model the influence and the inter-arrival times between actions in a sequence. Moreover, for time-sensitive prediction, we perform an *early* detection of sequence goal via a constrained margin-based optimization procedure. This in-turn allows PROACTIVE to predict the sequence goal using a limited number of actions. Extensive experiments on sequences derived from three activity recognition datasets show the significant accuracy boost of PROACTIVE over the state-of-the-art in terms of action and goal prediction, and the first-ever application of end-to-end action sequence generation.

CCS CONCEPTS

• Information systems → Data mining.

KEYWORDS

marked temporal point process; continuous time sequences; activity modeling; goal prediction; sequence generation

ACM Reference Format:

Vinayak Gupta and Srikanta Bedathur. 2022. PROACTIVE: Self-Attentive Temporal Point Process Flows for Activity Sequences. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3534678.3539477>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '22, August 14–18, 2022, Washington, DC, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9385-0/22/08...\$15.00
<https://doi.org/10.1145/3534678.3539477>

1 INTRODUCTION

A majority of data generated via human activities, *e.g.*, running, playing basketball, cooking, *etc.*, can be represented as a sequence of actions over a continuous-time. These actions denote a step taken by a user towards achieving a certain goal and vary in their start and completion times, depending on the user and the surrounding environment [20, 27, 28]. Therefore, unlike synthetic time series, these continuous-time action sequences (CTAS) can vary significantly even if they consist of the same set of actions. For *e.g.*, one person making omelets may take a longer time to cook eggs while another may prefer to cook for a short time¹; or Xavi may make a quicker pass than Pirlo in a football game – although in both cases the goals are same, and they all have to perform the same sequence of actions. In addition, modeling the dynamics of these CTAS is increasingly challenging due to the limited ability of modern recurrent and self-attention-based approaches in capturing continuous action times [16, 40]. This situation is further exacerbated due to the asynchronous occurrence of actions and the large variance in action-*times* and *types*. Therefore, the problem of modeling a CTAS has been overlooked by the past literature.

In recent years, neural marked temporal point processes (MTPP) have shown a significant promise in modeling a variety of continuous-time sequences in healthcare [33, 34], finance [3, 47], education [43], and social networks [29, 48, 49]. However, standard MTPP have a limited modeling ability for CTAS as: (i) they assume a homogeneity among sequences, *i.e.*, they cannot distinguish between two sequences of similar actions but with different time duration; (ii) in a CTAS, an action may finish before the start of the next action and thus, to model this empty time interval an MTPP must introduce a new action type, *i.e.*, *NULL* or *end-action* which may lead to an unwarranted increase in the types of actions to be modeled; and (iii) they cannot encapsulate the additional features associated with an action, for *e.g.*, minimum time for completion, necessary previous actions, or can be extended to sequence generation.

1.1 Our Contribution

In this work, we present PROACTIVE (Point Process flows for Activity Sequences), a normalizing flow-based neural MTPP framework, designed specifically to model the dynamics of a CTAS. Specifically, PROACTIVE addresses three challenging problems – (i) action prediction; (ii) goal detection; and (iii) the first-of-its-kind task of *end-to-end* sequence generation. We learn the distribution of actions in a CTAS using temporal normalizing flows (NF) [26, 32] conditioned on the dynamics of the sequence as well as the action-features (*e.g.*, minimum completion time, *etc.*). Such a flow-based

¹<https://bit.ly/3F5aEwX> (Accessed January 2022)

formulation provides PROACTIVE flexibility over other similar models [36, 37] to better model the inter-action dynamics within and across sequences. Moreover, our model is designed for *early* detection of the sequence goal, *i.e.*, identifying the result of a CTAS without traversing the complete sequence. We achieve this via a time-bounded optimization, *i.e.*, by incrementally increasing the probability of identifying the true goal using a *margin*-based and a weighted factor-based learning [15, 25]. Such an optimization procedure allows PROACTIVE to model the goal-action hierarchy, *i.e.*, the *necessary* set of actions in a CTAS towards achieving a particular goal, and simultaneously, the order of actions in CTAS with similar goals.

To the best of our knowledge, in this paper we present the first-ever application of MTPP via *end-to-end* action sequence generation. Specifically, given the resultant goal, PROACTIVE can generate a CTAS with the necessary set of actions and their occurrence times. Such a novel ability for MTPP models can reinforce their usage in applications related to bio-signals [13], sensor-data [1], *etc.*, and overcome the modeling challenge due to scarcity of activity data [5, 10, 24]. Buoyed by the success of attention models in sequential applications [40], we use a self-attention architecture in PROACTIVE to model the inter-action influences in a CTAS. In summary, the key contributions we make in this paper via PROACTIVE are:

- (1) We propose PROACTIVE, a novel temporal flow-based MTPP framework designed specifically for modeling human activities with a time-bounded optimization framework for early detection of CTAS goal.
- (2) Our normalizing flow-based modeling framework incorporates the sequence and individual action dynamics along with the action-goal hierarchy. Thus, PROACTIVE introduces the first-of-its-kind MTPP application of end-to-end action CTAS generation with just the sequence-goal as input.
- (3) Finally, we empirically show that PROACTIVE outperforms the state-of-the-art models for all three tasks – action prediction, goal detection, and sequence generation.

1.2 Organization

We present a problem formulation and a background on necessary techniques in Section 2. Section 3 gives an overview followed by an detailed development of all components in PROACTIVE. Section 4 contains in-depth experimental analysis and qualitative studies over all datasets. Lastly, Section 5 reviews a few relevant works before concluding in Section 6.

2 PRELIMINARIES

In this section, we present a background of MTPP and normalizing flows, and then present the problems addressed in this paper.

2.1 Background

Marked Temporal Point Processes. MTPP[14] are probabilistic generative models for continuous-time event sequences. An MTPP can be represented as a probability distribution over sequences of variable length belonging to a time interval $[0, T]$. Equivalently, they can be described using a counting process, say $N(t)$, and are characterized by the underlying conditional intensity function, $\lambda^*(t)$ which specifies the likelihood of the next event, conditioned

on the history of events. The intensity function $\lambda^*(t)$ computes the infinitesimal probability that an event will happen in the time window $(t, t + dt]$ conditioned on the history as:

$$\mathbb{P}(dN(t) = N(t + dt) - N(t) = 1) = \lambda^*(t), \quad (1)$$

Here, $*$ denotes a dependence on history. Given the conditional intensity function, we obtain the probability density function as:

$$p^*(\Delta_{t,i}) = \lambda^*(t_{i-1} + \Delta_{t,i}) \exp\left(-\int_0^{\Delta_{t,i}} \lambda^*(t_{i-1} + r) dr\right), \quad (2)$$

where, $\Delta_{t,i}$ denotes the inter-event time interval, *i.e.*, $t_i - t_{i-1}$. In contrast to other neural MTPP models that rely on the intensity function [6, 29, 47, 49] we replace the intensity function with a *log-normal* flow. Such a formulation facilitates closed-form and faster sampling as well as more accurate prediction than the intensity-based models [26, 37].

Normalizing Flows. Normalizing flows[26, 32, 37] are generative models that are used for density estimation as well as event sampling. They work by mapping simple distributions to complex ones using multiple bijective functions. In detail, if our goal is to estimate the density function $p_{\mathbf{X}}$ of a random vector $\mathbf{X} \in \mathbb{R}^D$, then the normalizing flows assign a new distribution $\mathbf{X} = g_{\phi}(\mathbf{Z})$, with $g_{\phi} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ is a bijective function, and $\mathbf{Z} \in \mathbb{R}^D$ is a random vector sampled from a simple density function $p_{\mathbf{Z}}$. Using NFs, sampling an event from $p_{\mathbf{X}}$ is done by a two step procedure of first sampling from the simple distribution $\mathbf{z} \sim p_{\mathbf{Z}}$ and then applying the bijective function $\mathbf{x} = g_{\phi}(\mathbf{z})$. Such a procedure supports closed form sampling. Moreover, modern approaches[18, 19, 32] represent the function $g_{\phi}(\bullet)$ via a neural network.

2.2 Problem Formulation

As mentioned in Section 1, we represent an activity via a continuous-time action sequence, *i.e.*, a series of actions undertaken by users and their corresponding time of occurrences. We derive each CTAS from annotated frames of videos consisting of individuals performing certain activities. Specifically, for every video, we have a sequence of activity labels being performed in the video along with timestamps for each activity. Therefore, each CTAS used in our dataset is derived from these sequences of a video. Formally, we provide a detailed description of a CTAS in Definition 1.

Definition 1 (Continuous Time Action Sequence). *We define a continuous-time action sequence (CTAS) as a series of action events taken by a user to achieve a particular goal. Specifically, we represent a CTAS as $S_k = \{e_i = (c_i, t_i) | i \in [k], t_i < t_{i+1}\}$, where $t_i \in \mathbb{R}^+$ is the start-time of the action, $c_i \in C$ is the discrete category or mark of the i -th action, C is the set of all categories, $\Delta_{t,i} = t_i - t_{i-1}$ as the inter-action time, and S_k denotes the sequence of first k actions. Each CTAS has an associated result, $g \in \mathcal{G}$, that signifies the goal of the CTAS. Here, \mathcal{G} denotes the set of all possible sequence goals.*

To highlight the relationship between sequence goal and actions consider the example of a CTAS with the goal of ‘*making-coffee*’, would comprise of actions – ‘*take-a-cup*’, ‘*pour-milk*’, ‘*add-coffee-powder*’, ‘*add-sugar*’, and ‘*stir*’ – at different time intervals. Given the aforementioned definitions, we formulate the tasks of next action, sequence goal prediction, and sequence generation as:

Input. A CTAS of all actions, S_k , consisting of categories and times of different actions that lead to a goal g .

Output. A probabilistic prediction model with three distinct tasks – (i) to estimate the likelihood of the next action e_{k+1} along with the action category and occurrence time; (ii) to predict the goal of the CTAS being modeled, *i.e.*, \hat{g} ; and (iii) a generative model to sample a sequence of actions, $\widehat{\mathcal{S}}$ given the true sequence goal, g .

3 PROACTIVE MODEL

In this section, we first present a high-level overview of the PROACTIVE model and then describe the neural parameterization of each component in detail. Lastly, we provide a detailed description of its optimization and sequence generation procedure.

3.1 High Level Overview

We use an MTPP denoted by $p_\theta(\cdot)$, to learn the generative mechanism of a continuous-time action sequence. Moreover, we design the sequence modeling framework of $p_\theta(\cdot)$ using a self-attention based encoder-decoder model [40]. Specifically, we embed the actions in a CTAS, *i.e.*, \mathcal{S}_k , to a vector embedding, denoted by \mathbf{s}_k , using a weighted aggregation of all past actions. Therefore, \mathbf{s}_k signifies a compact neural representation of the sequence history, *i.e.*, all actions till the k -th index and their marks and occurrence times. Recent research [47, 49] has shown that an attention-based modeling choice can better capture the long-term dependencies as compared to RNN-based MTPP models [6, 29, 30, 37]. A detailed description of the embedding procedure is given in Section 3.2.

We use our MTPP $p_\theta(\cdot)$ to estimate the generative model for the $(k+1)$ -th action conditioned on the past, *i.e.*, $p(e_{k+1})$ as:

$$p_\theta(e_{k+1}|\mathbf{s}_k) = \mathbb{P}_\theta(c_{k+1}|\mathbf{s}_k) \cdot \rho_\theta(\Delta_{t,k+1}|\mathbf{s}_k), \quad (3)$$

where, $\mathbb{P}_\theta(\cdot)$ and $\rho_\theta(\cdot)$ denote the probability distribution of marks and the density function for inter-action arrival times respectively. Note that both the functions are conditioned on \mathbf{s}_k and thus PROACTIVE requires a joint optimizing procedure for both – action time and mark prediction. Next, we describe the mechanism used in PROACTIVE to predict the next action and goal detection in a CTAS.

Next Action Prediction. We determine the most probable mark and time of the next action, using $p_\theta(\cdot)$ via standard sampling techniques over $\mathbb{P}_\theta(\cdot)$ and $\rho_\theta(\cdot)$ respectively [6, 37].

$$\widehat{e}_{k+1} \sim p_\theta(e_{k+1}|\mathbf{s}_k), \quad (4)$$

In addition, to keep the history embedding up-to-date with the all past actions, we iteratively update \mathbf{s}_k to \mathbf{s}_{k+1} by incorporating the details of action e_{k+1} .

Goal Detection. Since the history embedding, \mathbf{s}_k , represents an aggregation of all past actions in a sequence, it can also be used to capture the influences between actions and thus, can be extended to detect the goal of the CTAS. Specifically, to detect the CTAS goal, we use a non-linear transformation over \mathbf{s}_k as:

$$\widehat{g} \sim \mathbb{P}_{g' \in \mathcal{G}}(\Phi(\mathbf{s}_k)), \quad (5)$$

where, \mathbb{P}_\bullet denotes the distribution over all sequence goals and $\Phi(\cdot)$ denotes the transformation via a fully-connected MLP layer.

3.2 Neural Parametrization

Here, we present a detailed description of the neural architecture of our MTPP, $p_\theta(\cdot)$, and the optimization procedure in PROACTIVE. Specifically, we realize $p_\theta(\cdot)$ using a three layer architecture:

Input Layer. As mentioned in Section 2, each action $e_i \in \mathcal{S}_k$ is represented by a mark c_i and time t_i . Therefore, we embed each action as a combination of all these features as:

$$\mathbf{y}_i = \mathbf{w}_{y,c}c_i + \mathbf{w}_{y,t}t_i + \mathbf{w}_{y,\Delta}\Delta_{t,i} + \mathbf{b}_y, \quad (6)$$

where $\mathbf{w}_{\bullet,\bullet}$, \mathbf{b}_\bullet are trainable parameters and $\mathbf{y}_i \in \mathbb{R}^D$ denotes the vector embedding for the action e_i respectively. In other sections as well, we denote weight and bias as $\mathbf{w}_{\bullet,\bullet}$ and $\mathbf{b}_{\bullet,\bullet}$ respectively.

Self-Attention Layer. We use a *masked* self-attention layer to embed the past actions to \mathbf{s}_k and to interpret the influence between the past and the future actions. In detail, we follow the standard attention procedure [40] and first add a trainable positional encoding, \mathbf{p}_i , to the action embedding, *i.e.*, $\mathbf{y}_i \leftarrow \mathbf{y}_i + \mathbf{p}_i$. Such trainable encodings are shown to be more scalable and robust for long sequence lengths as compared to those based on a fixed function [16, 22]. Later, to calculate an attentive aggregation of all actions in the past, we perform three independent linear transformations on the action representation to get the *query*, *key*, and *value* embeddings, *i.e.*,

$$\mathbf{q}_i = \mathbf{W}^Q\mathbf{y}_i, \quad \mathbf{k}_i = \mathbf{W}^K\mathbf{y}_i, \quad \mathbf{v}_i = \mathbf{W}^V\mathbf{y}_i, \quad (7)$$

where, \mathbf{q}_\bullet , \mathbf{k}_\bullet , \mathbf{v}_\bullet denote the query, key, and value vectors respectively. Following standard self-attention model, we represent \mathbf{W}^Q , \mathbf{W}^K and \mathbf{W}^V as trainable *Query*, *Key*, and *Value* matrices respectively. Finally, we compute \mathbf{s}_k conditioned on the history as:

$$\mathbf{s}_k = \sum_{i=1}^k \frac{\exp(\mathbf{q}_k^\top \mathbf{k}_i / \sqrt{D})}{\sum_{i'=1}^k \exp(\mathbf{q}_k^\top \mathbf{k}_{i'} / \sqrt{D})} \mathbf{v}_i, \quad (8)$$

where, D denotes the number of hidden dimensions. Here, we compute the attention weights via a softmax over the interactions between the query and key embeddings of each action in the sequence and perform a weighted sum of the value embeddings.

Now, given the representation \mathbf{s}_k , we use the attention mechanism in Eqn. (8) and apply a feed-forward neural network to incorporate the necessary non-linearity to the model as:

$$\mathbf{s}_k \leftarrow \sum_{i=1}^k [\mathbf{w}_{s,m} \odot \text{RELU}(\mathbf{s}_i \odot \mathbf{w}_{s,n} + \mathbf{b}_{s,n}) + \mathbf{b}_{s,m}],$$

where, $\mathbf{w}_{s,m}$, $\mathbf{b}_{s,m}$ and $\mathbf{w}_{s,n}$, $\mathbf{b}_{s,n}$ are trainable parameters of the outer and inner layer of the point-wise feed-forward layer.

To support faster convergence and training stability, we employ: (i) layer normalization; (ii) stacking multiple self-attention blocks; and (iii) multi-head attention. Since these are standard techniques [2, 40], we omit their mathematical descriptions in this paper.

Output Layer. At every index k , PROACTIVE outputs the next action and the most probable goal of the CTAS. We present the prediction procedure for each of them as follows:

Action Prediction: We use the output of the self-attention layer, \mathbf{s}_k to estimate the mark distribution and time density of the next event, *i.e.*, $\mathbb{P}_\theta(e_{k+1})$ and $\rho_\theta(e_{k+1})$ respectively. Specifically, we model the $\mathbb{P}_\theta(\cdot)$ as a softmax over all other marks as:

$$\mathbb{P}_\theta(c_{k+1}) = \frac{\exp(\mathbf{w}_{c,s}^\top \mathbf{s}_k + \mathbf{b}_{c,s})}{\sum_{c'=1}^{|C|} \exp(\mathbf{w}_{c',s}^\top \mathbf{s}_k + \mathbf{b}_{c',s})}, \quad (9)$$

where, $\mathbf{w}_{\bullet,\bullet}$ and $\mathbf{b}_{\bullet,\bullet}$ are trainable parameters.

In contrast to standard MTPP approaches that rely on an intensity-based model [6, 29, 47, 49], we capture the inter-action arrival times via a *temporal* normalizing flow (NF). In detail, we use a *LogNormal*

flow to model the temporal density $\rho_\theta(\Delta_{t,k+1})$. Moreover, standard flow-based approaches [26, 37] utilize a common NF for all events in a sequence, *i.e.*, the arrival times of each event are determined from a single or mixture of flows trained on all sequences. We highlight that such an assumption restricts the ability to model the dynamics of a CTAS, as unlike standard events, an action has three distinguishable characteristics – (i) every action requires a minimum time for completion; (ii) the time taken by a user to complete an action would be similar to the times of another user; and (iii) similar actions require similar times to complete. For example, the time taken to complete the action ‘*add-coffee*’ would require a certain minimum time of completion and these times would be similar for all users. Intuitively, the time for completing the action ‘*add-coffee*’ would be similar to those for the action ‘*add-sugar*’.

To incorporate these features in PROACTIVE, we identify actions with similar completion times and model them via independent temporal flows. Specifically, we cluster all actions $c_i \in \mathcal{C}$ into M non-overlapping clusters based on the *mean* of their times of completion and for each cluster we define a trainable embedding $\mathbf{z}_r \in \mathbb{R}^D \forall r \in \{1, \dots, M\}$. Later, we sample the start-time of the future action by conditioning our temporal flows on the cluster of the current action as:

$$\widehat{\Delta_{t,k+1}} \sim \text{LOGNORMAL}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2), \quad (10)$$

where, $[\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2]$, denote the mean and variance of the log-normal temporal flow and are calculated via the sequence embedding and the cluster embedding as:

$$\boldsymbol{\mu}_k = \sum_{r=1}^M \mathcal{R}(e_k, r) (\mathbf{w}_\mu (\mathbf{s}_k \odot \mathbf{z}_{c,i}) + \mathbf{b}_\mu), \quad (11)$$

$$\boldsymbol{\sigma}_k^2 = \sum_{r=1}^M \mathcal{R}(e_k, r) (\mathbf{w}_\sigma (\mathbf{s}_k \odot \mathbf{z}_{c,i}) + \mathbf{b}_\sigma), \quad (12)$$

where $\mathbf{w}_\bullet, \mathbf{b}_\bullet$ are trainable parameters, $\mathcal{R}(e_k, r)$ is an indicator function that determines if event e_k belongs to the cluster r and \mathbf{z}_r denotes the corresponding cluster embedding. Such a cluster-based formulation facilitates the ability of model to assign similar completion times for events in a same cluster. To calculate the time of next action, we add the sampled time difference to the time of the previous action e_k , *i.e.*,

$$\widehat{t_{k+1}} = t_k + \widehat{\Delta_{t,k+1}} \quad (13)$$

where, $\widehat{t_{k+1}}$ denotes the predicted time for the action e_{k+1} .

Goal Detection: In contrast to other MTPP approaches [6, 29, 37, 47, 49], an important feature of PROACTIVE is identifying the goal of a sequence, *i.e.*, a hierarchy on top of the actions in a sequence, based on the past sequence dynamics. To determine the goal of a CTAS, we utilize the history embedding \mathbf{s}_k as it encodes the inter-action relationships of all actions in the past. Specifically, we use a non-linear transformation via a feed-forward network, denoted as $\Phi(\cdot)$ over \mathbf{s}_k and apply a softmax over all possible goals.

$$\Phi(\mathbf{s}_k) = \text{RELU}(\mathbf{w}_{\Phi,s} \mathbf{s}_k + \mathbf{b}_{\Phi,s}), \quad (14)$$

where, $\mathbf{w}_{\Phi,s}, \mathbf{b}_{\Phi,s}$ are trainable parameters. We sample the most probable goal as in Eqn. (5).

We highlight that we predict the CTAS goal at each interval, though a CTAS has only one goal. This is to facilitate *early* goal detection in comparison to detecting the goal after traversing the entire CTAS. More details are given in Section 3.3 and Section 3.4.

3.3 Early Goal Detection and Action Hierarchy

Here, we highlight the two salient features of PROACTIVE– early goal detection and modeling the goal-action hierarchy.

Early Goal Detection. Early detection of sequence goals has many applications ranging from robotics to vision [15, 35]. To facilitate early detection of the goal of a CTAS in PROACTIVE, we devise a ranking loss that forces the model to predict a *non-decreasing* detection score for the correct goal category. Specifically, the detection score of the correct goal at the k -th index of the sequence, denoted by $p_k(g|\mathbf{s}_k, \Phi)$, must be more than the scores assigned the correct goal in the past. Formally, we define the ranking loss as:

$$\mathcal{L}_{k,g} = \max(0, p_k^*(g) - p_k(g|\mathbf{s}_k, \Phi)), \quad (15)$$

where $p_k^*(g)$ denotes the maximum probability score given to the correct goal in all past predictions.

$$p_k^*(g) = \max_{j \in \{1, k-1\}} p_j(g|\mathbf{s}_j, \Phi), \quad (16)$$

where $p_k(g)$ denotes the probability score for the correct goal at index k . Intuitively, the ranking loss $\mathcal{L}_{k,g}$ would penalize the model for predicting a smaller detection score for the correct CTAS goal than any previous detection score for the same goal.

Action Hierarchy. Standard MTPP approaches assume the category of marks as independent discrete variables, *i.e.*, the probability of an upcoming mark is calculated independently [6, 29, 47, 49]. Such an assumption restricts the predictive ability while modeling CTAS, as in the latter case, there exists a hierarchy between goals and actions that lead to the specific goal. Specifically, actions that lead to a common goal may have similar dynamics and it is also essential to model the relationships between the actions of different CTAS with a common goal. We incorporate this hierarchy in PROACTIVE along with our next action prediction via an action-based ranking loss. In detail, we devise a loss function similar to Eqn. 15 where we restrict the model to assign non-decreasing probabilities to all actions leading to the goal of CTAS under scrutiny.

$$\mathcal{L}_{k,c} = \sum_{c' \in C_g^*} \max(0, p_k^*(c') - p_k(c'|\mathbf{s}_k)), \quad (17)$$

where C_g^* , $p_k(c'|\mathbf{s}_k)$ denote a set of all actions in CTAS with the goal g and the probability score for the action $c' \in C_g^*$ at index k respectively. Here, $p_k^*(c')$ denotes the maximum probability score given to action c' in all past predictions and is calculated similar to Eqn. 16. We regard $\mathcal{L}_{k,g}$ and $\mathcal{L}_{k,c}$ as *margin* losses, as they aim to increase the difference between two prediction probabilities.

3.4 Optimization

We optimize the trainable parameters in PROACTIVE, *i.e.*, the weight and bias tensors ($\mathbf{w}_{\bullet,\bullet}$ and $\mathbf{b}_{\bullet,\bullet}$) for our MTPP $p_\theta(\cdot)$, using a two channels of training consisting of action and goal prediction. Specifically, to optimize the ability of PROACTIVE for predicting the next action, we maximize the the joint likelihood for the next action and the log-normal density distribution of the temporal flows.

$$\mathcal{L} = \sum_{k=1}^{|\mathcal{S}|} \log(\mathbb{P}_\theta(c_{k+1}|\mathbf{s}_k) \cdot \rho_\theta(\Delta_{t,k+1}|\mathbf{s}_k)), \quad (18)$$

where \mathcal{L} denotes the joint likelihood, which we represent as the sum of the likelihoods for all CTAS. In addition to action prediction, we optimize the PROACTIVE parameters for *early* goal detection via

a temporally weighted cross entropy (CE) loss over all sequence goals. Specifically, we follow a popular reinforcement recipe of using a time-varying *discount* factor over the prediction loss as:

$$\mathcal{L}_g = \sum_{k=1}^{|\mathcal{S}|} \gamma^k \cdot \mathcal{L}_{\text{CE}}(p_k(g|\mathbf{s}_k)), \quad (19)$$

where $\gamma \in [0, 1]$, $\mathcal{L}_{\text{CE}}(p_k(g|\mathbf{s}_k))$ denote the decaying factor and a standard softmax-cross-entropy loss respectively. Such a recipe is used exhaustively for faster convergence of reinforcement learning models [31, 38]. Here, the discount factor penalizes the model for taking longer times for detecting the CTAS goal by decreasing the gradient updates to the loss.

Margin Loss. In addition, we minimize the margin losses given in Section 3.3 with the current optimization procedure. Specifically, we minimize the following loss:

$$\mathcal{L}_m = \sum_{k=1}^{|\mathcal{S}|} \mathcal{L}_{k,g} + \mathcal{L}_{k,c}, \quad (20)$$

where $\mathcal{L}_{k,g}$ and $\mathcal{L}_{k,c}$ are margin losses defined in Eqn. 15 and Eqn. 17 respectively. We learn the parameters of PROACTIVE using an Adam [17] optimizer for both likelihood and prediction losses.

3.5 Sequence Generation

A crucial contribution of this paper via PROACTIVE is an end-to-end generation of action sequences. Specifically, given the CTAS goal as input, we can generate a most probable sequence of actions that may lead to that specific goal. Such a feature has a range of applications from sports analytics [28], forecasting [4], identifying the duration of an activity [27], *etc.*

A standard approach for training a sequence generator is to sample future actions in a sequence and then compare with the true actions [46]. However, such a procedure has multiple drawbacks as it is susceptible to noises during training and deteriorates the scalability of the model. Moreover, we highlight that such a sampling based training cannot be applied to a self-attention-based model as it requires a fixed sized sequence as input [40]. Therefore, we resort to a two-step generation procedure that is defined below:

- (1) **Pre-Training:** The first step requires training all PROACTIVE parameters for action prediction and goal detection. This step is necessary to model the relationships between actions and goals and we represent the set of optimized parameters as θ^* and the corresponding MTPP as $p_{\theta^*}(\cdot)$ respectively.
- (2) **Iterative Sampling:** We iteratively sample events and update parameters via our trained MTPP till the model predicts the correct goal for CTAS or we encounter an <EOS> action. Specifically, using $p_{\theta^*}(\cdot)$ and the first *real* action (e_1) as input, we calculate the detection score for the correct goal, *i.e.*, $p_1(g|\mathbf{s}_k)$ and while its value is highest among all probable goals, we sample the mark and time of next action using Eqn. 9 and Eqn. 10 respectively.

Such a generation procedure harnesses the fast sampling of temporal normalizing flows and simultaneously is conditioned on the action and goal relationships. A detailed pseudo-code of sequence generation procedure used in PROACTIVE is given in Algorithm 1.

Algorithm 1: Sequence Generation with PROACTIVE

```

1 Input:  $g$ : Goal of CTAS
2  $e_1$ : First Action
3  $p_{\theta^*}(\cdot)$ : Trained MTPP
4 Output:  $\widehat{\mathcal{S}}$ : Generated CTAS  $\mathcal{S}_1 \leftarrow e_1$ 
5  $k = 1$ 
6 while  $k < \text{max\_len}$  do
7   Sample the mark of next action:  $\widehat{c}_{k+1} \sim \mathbb{P}_{\theta^*}(\mathbf{s}_k)$ 
8   Sample the time of next action:  $\widehat{t}_{k+1} \sim \rho_{\theta^*}(\mathbf{s}_k)$ 
9   Add to CTAS:  $\mathcal{S}_{k+1} \leftarrow \mathcal{S}_k + e_{k+1}$ 
10  Update the MTPP parameters  $\mathbf{s}_{k+1} \leftarrow p(\mathbf{s}_k, e_{k+1})$ 
11  Calculate most probable goal:  $\widehat{g}_k = \max_{g'}(p_k(g'|\mathbf{s}_k))$ 
12  if  $\widehat{g}_k \neq g$  or  $\widehat{c}_{k+1} == \text{EOS}$  > then
13    Add EOS mark:  $\widehat{\mathcal{S}} \leftarrow \mathcal{S}_{k+1} + \text{EOS}$ 
14    Exit the sampling procedure: BREAK
15  Increment iteration:  $k \leftarrow k + 1$ 
16 Return generated CTAS: return  $\widehat{\mathcal{S}}$ 

```

4 EXPERIMENTS

In this section, we present the experimental setup and the empirical results to validate the efficacy of PROACTIVE. Through our experiments we aim to answer the following research questions:

- RQ1** What is the action-mark and time prediction performance of PROACTIVE in comparison to the state-of-the-art baselines?
- RQ2** How accurately and quickly can PROACTIVE identify the goal of an activity sequence?
- RQ3** How effectively can PROACTIVE generate an action sequence?
- RQ4** How does the action prediction performance of PROACTIVE vary with different hyperparameters values?

4.1 Datasets

To evaluate PROACTIVE, we need time-stamped action sequences and their goals. Therefore, we derive CTAS from three activity modeling datasets sourced from different real-world applications – cooking, sports, and collective activity. The datasets vary significantly in terms of origin, sparsity, and sequence lengths. We highlight the details of each of these datasets below:

- **Breakfast [20].** This dataset contains CTAS derived from 1712 videos of different people preparing breakfast. The actions in a CTAS and sequence goals can be classified into 48 and 9 classes respectively. These actions are performed by 52 different individuals in 18 different kitchens.
- **Multi-THUMOS [45].** A sports activity dataset that is designed for action recognition in videos. We derive the CTAS using 400 videos of individuals involved in different sports such as discus throw, baseball, *etc.* The actions and goals can be classified into 65 and 9 classes respectively and on average, there are 10.5 action class labels per video.
- **Activity-Net [7].** This dataset comprises of activity categories collected from 591 YouTube videos with a total of 49 action labels and 14 goals.

We highlight that in Activity-Net, many of the videos are shot by amateurs in many uncontrolled environments, the variances within the CTAS of the same goal are often large, and the lengths of CTAS vary and are often long and complex.

Table 1: Performance of all the methods in terms of action prediction accuracy (APA) and mean absolute error (MAE) across all datasets. Bold fonts and underline indicate the best performer and the best baseline respectively. Results marked † are statistically significant (i.e., two-sided t-test with $p \leq 0.1$) over the best baseline.

	Action Prediction Accuracy (APA)			Mean Absolute Error (MAE)		
	Breakfast	Multi-THUMOS	Activity-Net	Breakfast	Multi-THUMOS	Activity-Net
NHP [29]	0.528±0.024	0.272±0.019	0.684±0.034	0.411±0.019	<u>0.017±0.002</u>	0.796±0.045
AVAE [27]	0.533±0.028	0.279±0.022	0.678±0.036	0.417±0.021	0.018±0.002	0.803±0.049
RMTTP [6]	0.542±0.022	0.274±0.017	0.683±0.034	<u>0.403±0.018</u>	<u>0.017±0.002</u>	<u>0.791±0.046</u>
SAHP [47]	0.547±0.031	0.287±0.023	0.688±0.042	0.425±0.031	0.019±0.003	0.820±0.072
THP [49]	<u>0.559±0.028</u>	<u>0.305±0.018</u>	0.693±0.038	0.413±0.023	0.019±0.002	0.806±0.061
PROACT-c	0.561±0.027	0.297±0.020	0.698±0.038	0.415±0.027	0.015±0.002	0.774±0.054
PROACT-t	0.579±0.025	0.306±0.018	0.722±0.035	0.407±0.025	0.015±0.002	0.783±0.058
PROACTIVE	0.583±0.027†	0.316±0.019	0.728±0.037†	0.364±0.028†	0.013±0.002†	0.742±0.059†

4.2 Baselines

We compare the action prediction performance of PROACTIVE with the following state-of-the-art methods:

RMTTP [6]: A recurrent neural network that models time differences to learn a representation of the past events.

NHP [29]: Models an MTPP using continuous-time LSTMs for capturing the temporal evolution of sequences.

AVAE [27]: A variational auto-encoder based MTPP framework designed specifically for activities in a sequence.

SAHP [47]: A self-attention model to learn the temporal dynamics using an aggregation of historical events.

THP [49]: Extends the transformer model [40] to include the *conditional* intensity of event arrival and the inter-mark influences. We omit comparison with other continuous-time models [14, 30, 37, 41, 42] as they have already been outperformed by these approaches.

4.3 Evaluation Criteria

Given the dataset \mathcal{D} of N action sequences, we split them into training and test set based on the goal of the sequence. Specifically, for each goal $g \in \mathcal{G}$, we consider 80% of the sequences as the training set and the other last 20% as the test set. We evaluate PROACTIVE and all baselines on the test set in terms of (i) mean absolute error (MAE) of predicted times of action, and (ii) action prediction accuracy (APA) described as:

$$\text{MAE} = \frac{1}{|\mathcal{S}|} \sum_{e_i \in \mathcal{S}} \|t_i - \hat{t}_i\|, \quad \text{APA} = \frac{1}{|\mathcal{S}|} \sum_{e_i \in \mathcal{S}} \#(c_i = \hat{c}_i), \quad (21)$$

where, \hat{t}_i and \hat{c}_i are the predicted time and type the i -th action in test set. Moreover, we follow a similar protocol to evaluate the sequence generation ability of PROACTIVE and other models. For goal prediction, we report the results in terms of accuracy (ratio) calculated across all sequences. We calculate confidence intervals across 5 independent runs.

4.4 Experimental Setup

All our implementations and datasets are publicly available at: <https://github.com/data-iitd/proactive/>.

System Configuration. All our experiments were done on a server running Ubuntu 16.04. CPU: Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz, RAM: 125GB and GPU: NVIDIA Tesla T4 16GB DDR6.

Parameter Settings. For our experiments, we set $D = 16$, $M = 8$, $\gamma = 0.9$ and weigh the margin loss \mathcal{L}_m by 0.1. In addition, we set a l_2 regularizer over the parameters with coefficient value 0.001.

4.5 Action Prediction (RQ1)

We report on the performance of action prediction of different methods across all our datasets in Table 1. In addition, we include two variants of PROACTIVE– (i) PROACT-c, represents our model without the goal-action hierarchy loss (Eqn. 17) and cluster-based flows (Eqn. 11 and 12), and (ii) PROACT-t, represents our model without cluster-based flows. From Table 1, we note the following:

- PROACTIVE consistently yields the best prediction performance on all the datasets. In particular, it improves over the strongest baselines by 8-27% for time prediction and by 2-7% for action prediction. These results signify the drawbacks of using standard sequence approaches for modeling a temporal action sequence.
- RMTTP [6] is the second-best performer in terms of MAE of time prediction in almost all the datasets. We also note that for Activity-Netdataset, THP [49] outperforms RMTTP for action category prediction. However, PROACTIVE still significantly outperforms these models across all metrics.
- Neural MTPP methods that deploy a self-attention for modeling the distribution of action – namely THP, SAHP, and PROACTIVE, achieve better performance in terms of category prediction.
- Despite AVAE [27] being a sequence model designed specifically for activity sequences, other neural methods that incorporate complex structures using self-attention or normalizing flows easily outperform it.

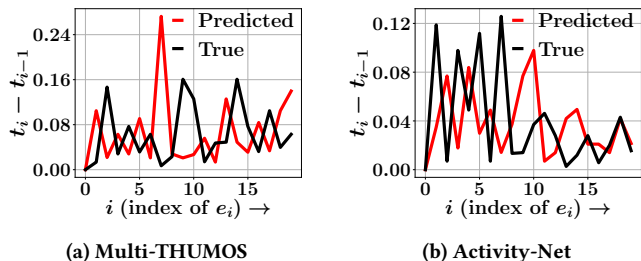


Figure 1: Real life true and predicted inter-arrival times $\Delta_{t,k}$ of different events e_k for (a) Multi-THUMOS and (b) Activity-Net datasets. The results show that the true arrival times match with the times predicted by PROACTIVE.

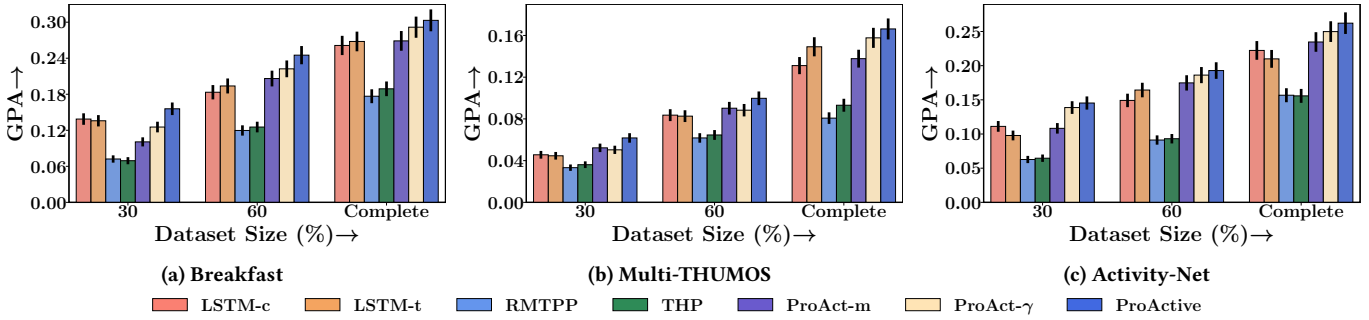


Figure 2: Sequence goal prediction performance of PROACTIVE, its variants – PROACT-m and PROACT- γ , and other baseline models. The results show that PROACTIVE can effectively detect the CTAS goal even with smaller test sequences as input.

To sum up, our empirical analysis suggests that PROACTIVE can better model the underlying dynamics of a CTAS as compared to all other baseline models. Moreover, the performance gain over PROACT-c and PROACT-t highlights the need for modeling action hierarchy and cluster-based flows.

Qualitative Assessment. We also perform a qualitative analysis to highlight the ability of PROACTIVE for modeling the inter-arrival times for action prediction. Specifically, we plot the actual inter-arrival time differences and the time-difference predicted by PROACTIVE in Figure 1 for Multi-THUMOS and Activity-Net datasets. From the results, we note that the predicted inter-arrival times closely match with the true inter-arrival times and PROACTIVE is even able to capture large time differences (peaks). For brevity, we omitted the results for Activity-Net dataset.

4.6 Goal Prediction (RQ2)

Here, we evaluate the goal detection performance of PROACTIVE along with other baselines. To highlight the *early* goal detection ability of our model, we report the results across different variants of the test set, *i.e.*, with the initial 30% and 60% of the actions in the CTAS in terms of goal prediction accuracy (GPA). In addition, we introduce two novel baselines, LSTM-c, and LSTM-t, that detect the CTAS goal using just the types and the times of actions respectively. We also compare with the two best-performing MTPP baselines – RMTTP and THP which we extend for the task of goal detection by a k-means clustering algorithm. In detail, we obtain the sequence embedding, say s_χ using the MTPP models and then cluster them into $|\mathcal{G}|$ clusters based on their cosine similarities and perform a maximum *polling* across each cluster, *i.e.*, predict the most common goal for each cluster as the goal for all CTAS in the same cluster. In addition, we introduce two new variants of our model to analyze the benefits of early goal detection procedures in PROACTIVE– (i) PROACTIVE-m, represents our model without the goal-based margin loss given in Eqn. 15 and (ii) PROACTIVE- γ , is our model without the discount-factor weight in Eqn. 19. We also report the results for the complete model PROACTIVE.

The results for goal detection in Figure 2, show that the complete design of PROACTIVE achieves the best performance among all other models. We also note that the performance of MTPP based models deteriorates significantly for this new task which shows the unilateral nature of the prediction prowess of MTPP models, unlike PROACTIVE. Interestingly, the variant of PROACTIVE without the



Figure 3: Sequence Generation results for PROACTIVE and other baselines in terms of APA for action prediction.

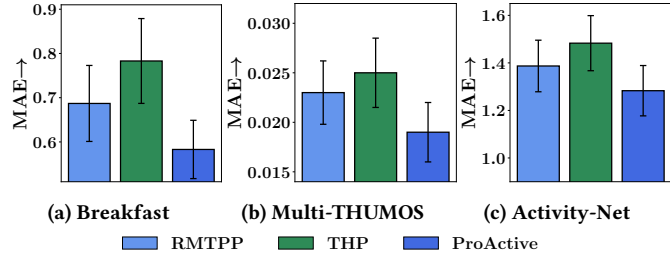


Figure 4: Sequence Generation results for PROACTIVE and other baselines in terms of MAE for time prediction.

margin loss PROACTIVE-m performs poorly as compared to the one without the discount-factor, PROACTIVE- γ . This could be attributed to better convergence guarantees with a margin-based loss over the latter. Finally, we observe that standard LSTM models are easily outperformed by our model, thus reinforcing the need for joint training of types and action times.

4.7 Sequence Generation (RQ3)

Here, we evaluate the sequence generation ability of PROACTIVE. Specifically, we generate all the sequences in the test set by giving the *true* goal of the CTAS and the first action as input to the procedure described in Section 3.5. However, there may be differences in lengths of the generated and true sequences, *i.e.*, the length of generated sequences is usually greater than the true CTAS. Therefore, we compare the actions in the true sequence with the initial $|S|$ generated actions. Such an evaluation procedure provides us the flexibility of comparing with other MTPP models such as RMTTP [6] and THP [49]. As these models cannot be used for end-to-end sequence generation, we alter their underlying model for *forecasting*

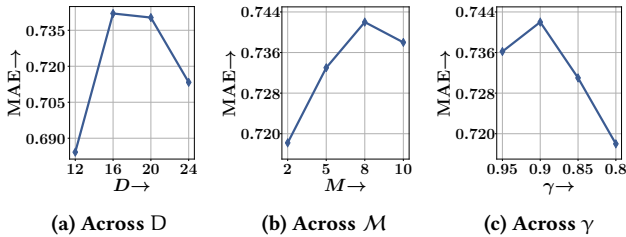


Figure 5: PROACTIVE sensitivity for Activity-Net dataset with different hyper-parameter values.

future actions given the first action and then iteratively update and sample from the MTPP parameters. We report the results in terms of APA and MAE for action and time prediction in Figure 3 and Figure 4 respectively. The results show that PROACTIVE can better capture the generative dynamics of a CTAS in comparison to other MTPP models. We also note that the prediction performance deteriorates significantly in comparison to the results given in Table 1. This could be attributed to the error that gets compounded in farther predictions made by the model. Interestingly, the performance advantage that PROACTIVE has over THP and RMTTP is further widened during sequence generation.

Length Comparison. Here, we report the results for length comparison of the generated sequence and the true sequence. Specifically, we identify the count of instances where PROACTIVE was able to effectively capture the generative mechanism of a sequence as:

$$CL = \frac{1}{N} \sum_{\forall S} \#(|S| = |\hat{S}|), \quad (22)$$

where CL denotes the *Correct-Length* ratio with values 0.21, 0.11, and 0.16 for datasets Breakfast, Multi-THUMOS, and Activity-Net respectively. On a coarse level these results might seem substandard, however, given the difficulty associated with the problem of sequence generation using just the CTAS goal, we believe these values are satisfactory. Moreover, we believe that the sequence generation procedure of PROACTIVE opens up new frontiers for generating action sequences.

4.8 Parameter Sensitivity (RQ4)

Finally, we perform the sensitivity analysis of PROACTIVE over key parameters: (i) D , the dimension of embeddings; (ii) M , no. clusters for log-normal flow; and (iii) γ , the discount factor described in Eqn.19. For brevity purposes, we only report results on Activity dataset, but other datasets displayed a similar behavior. From Figure 5 we show the performance of PROACTIVE across different hyper-parameter values. We note that as we increase the embedding dimension the performance first increases since it leads to better modeling. However, beyond a point the complexity of the model increases requiring more training to achieve good results, and hence we see its performance deteriorating. We see a similar trend for M , as increasing the number of clusters leads to better results before saturating at a certain point. We found $M = 5$ to be the optimal point across datasets in our experiments. Finally across γ , we notice that smaller values for gamma penalize the loss function for detecting the goal late, however, it deteriorates the action prediction performance of PROACTIVE. Therefore, we found $\gamma = 0.9$ as the best trade-off between goal and action prediction.

Scalability. For all datasets, the runtimes for training PROACTIVE are within 1 hour and thus are within the practical range for deployment in real-world scenarios. These running times further reinforce our design choice of using a neural MTPP due to their faster learning and closed-form sampling [26, 37].

5 RELATED WORK

In this section, we introduce key related work for this paper. It mainly falls into – activity prediction and temporal point processes.

5.1 Activity Prediction

Activity modeling in videos is a widely used application with recent approaches focusing on frame-based prediction. Lan et al. [21] predicts the future actions via hierarchical representations of short clips, Tahmida Mahmud and Roy-Chowdhury [39] jointly predicts future activity and the starting time by capturing different sequence features and a similar procedure is adopted by [44] that predicts the action categories of a sequence of future activities as well as their starting and ending time. Ma et al. [25] propose a method for early classification of a sequence of frames extracted from a video by maximizing the margin-based loss between the correct and the incorrect categories, however, it is limited to visual data and cannot incorporate the action-times. This limits its ability for use in CTAS, and sequence generation. A recent approach [27] proposed to model the dynamics of action sequences using a variational auto-encoder built on top of a temporal point process. We consider their work as most relevant to PROACTIVE as it also addressed the problem of CTAS modeling. However, as shown in our experiments PROACTIVE was able to easily outperform it across all metrics. This could be attributed to the limited modeling capacity of VAE over normalizing flows. Moreover, their sampling procedure could not be extended to sequence generation. Therefore, in contrast to the past literature, PROACTIVE is the first application of MTPP models for CTAS modeling and end-to-end sequence generation.

5.2 Temporal Point Processes

In recent years neural Marked Temporal Point Processes (MTPP) have shown a significant promise in modeling a variety of continuous-time sequences in healthcare [12, 33], finance [3, 47], education [43], and social networks [9, 11, 23, 29, 48, 49]. However, due to the limitations of traditional MTPP models, in recent years, neural enhancements to MTPP models have significantly enhanced the predictive power of these models. Specifically, they combine the continuous-time approach from the point process with deep learning approaches and thus, can better capture complex relationships between events. The most popular approaches [6, 29, 37, 47, 49] use different methods to model the time- and mark-distribution via neural networks. Specifically, Du et al. [6] embeds the event history to a vector representation via a recurrent encoder that updates its state after parsing each event in a sequence; Mei and Eisner [29] modified the LSTM architecture to employ a continuous-time state evolution; Shchur et al. [37] replaced the intensity function with a mixture of *log-normal* flows for closed-form sampling; Zhang et al. [47] utilized the transformer architecture Vaswani et al. [40] to capture the long-term dependencies between events in the history embedding and Zuo et al. [49] used the transformer architecture

for sequence embedding but extended it to graph settings as well. However, these models are not designed to capture the generative distribution of future events in human-generated sequences.

6 CONCLUSION

Standard deep-learning models are not designed for modeling sequences of actions localized in continuous time. However, neural MTPP models overcome this drawback but have limited ability to model the events performed by a human. Therefore, in this paper, we developed a novel point process flow-based architecture called PROACTIVE for modeling the dynamics of a CTAS. PROACTIVE solves the problems associated with action prediction, goal prediction, and for the first time, we extend MTPP for end-to-end CTAS generation. Our experiments on three large-scale diverse datasets reveal that PROACTIVE can significantly improve over the state-of-the-art baselines across all metrics. Moreover, the results also reinforce the novel ability of PROACTIVE to generate a CTAS. We hope that such an application will open many horizons for using MTPP in a wide range of tasks. As a future work, we plan to incorporate a generative adversarial network [8, 41] with action sampling and train the generator and the MTPP model simultaneously.

ACKNOWLEDGMENTS

This work was partially supported by a DS Chair of AI fellowship to Srikanta Bedathur.

REFERENCES

- [1] Moustafa Alzantot, Supriyo Chakraborty, and Mani Srivastava. 2017. Sensegen: A deep learning architecture for synthetic sensor data generation. In *PerCom (Workshops)*.
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [3] Emmanuel Bacry, Jacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes processes in finance. *arXiv preprint arXiv:1502.04592* (2015).
- [4] Prathamesh Deshpande, Kamlesh Marathe, Abir De, and Sunita Sarawagi. 2021. Long Horizon Forecasting With Temporal Point Processes. In *WSDM*.
- [5] Seth R Donahue, Li Jin, and Michael E Hahn. 2020. User Independent Estimations of Gait Events With Minimal Sensor Data. *IEEE Journal of Biomedical and Health Informatics* 25, 5 (2020), 1583–1590.
- [6] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*.
- [7] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Nibbles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*.
- [8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. In *NeurIPS*.
- [9] Vinayak Gupta and Srikanta Bedathur. 2021. Region Invariant Normalizing Flows for Mobility Transfer. In *CIKM*.
- [10] Vinayak Gupta and Srikanta Bedathur. 2022. Doing More with Less: Overcoming Data Scarcity for POI Recommendation via Cross-Region Transfer. *arXiv preprint arXiv:2201.06095* (2022).
- [11] Vinayak Gupta, Srikanta Bedathur, Sourangshu Bhattacharya, and Abir De. 2021. Learning Temporal Point Processes with Intermittent Observations. In *AISTATS*.
- [12] Vinayak Gupta, Srikanta Bedathur, and Abir De. 2022. Learning Temporal Point Processes for Efficient Retrieval of Continuous Time Event Sequences. In *AAAI*.
- [13] Shota Haradal, Hideaki Hayashi, and Seiichi Uchida. 2018. Biosignal data augmentation based on generative adversarial networks. In *EMBC*.
- [14] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.
- [15] Minh Hoai and Fernando De la Torre. 2012. Max-margin early event detectors. In *CVPR*.
- [16] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *ICDM*.
- [17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- [18] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*.
- [19] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In *NeurIPS*.
- [20] H. Kuehne, A. B. Arslan, and T. Serre. 2014. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *CVPR*.
- [21] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *ECCV*.
- [22] Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time Interval Aware Self-Attention for Sequential Recommendation. In *WSDM*.
- [23] Ankita Likhvani, Vinayak Gupta, PK Srijiith, P Deepak, and Srikanta Bedathur. 2020. Modeling Implicit Communities from Geo-tagged Event Traces using Spatio-Temporal Point Processes. In *WISE*.
- [24] Yue Luo, Sarah M Coppola, Philippe C Dixon, Song Li, Jack T Dennerlein, and Boyi Hu. 2020. A database of human gait performance on irregular and uneven surfaces collected by wearable sensors. *Scientific data* 7, 1 (2020), 1–9.
- [25] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In *CVPR*.
- [26] Nazanin Mehrasa, Ruizhi Deng, Mohamed Osama Ahmed, Bo Chang, Jiawei He, Thibaut Durand, Marcus Brubaker, and Greg Mori. 2019. Point Process Flows. *arXiv preprint arXiv:1910.08281* (2019).
- [27] Nazanin Mehrasa, Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. A Variational Auto-Encoder Model for Stochastic Point Processes. In *CVPR*.
- [28] Nazanin Mehrasa, Yatao Zhong, Frederick Tung, Luke Bornn, and Greg Mori. 2017. Learning person trajectory representations for team activity analysis. *arXiv preprint arXiv:1706.00893* (2017).
- [29] Hongyuan Mei and Jason M Eisner. 2017. The neural hawkes process: A neurally self-modulating multivariate point process. In *NeurIPS*.
- [30] Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. 2019. Fully Neural Network based Model for General Temporal Point Processes. In *NeurIPS*.
- [31] Silviu Pitis. 2019. Rethinking the Discount Factor in Reinforcement Learning: A Decision Theoretic Approach. In *AAAI*.
- [32] Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *ICML*.
- [33] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. 2018. SIR-Hawkes: on the Relationship Between Epidemic Models and Hawkes Point Processes. In *WWW*.
- [34] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be hip: Hawkes intensity processes for social media popularity. In *WWW*.
- [35] M. Ryoo. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*.
- [36] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD*.
- [37] Oleksandr Shchur, Marin Bilos, and Stephan Günnemann. 2020. Intensity-Free Learning of Temporal Point Processes. In *ICLR*.
- [38] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*.
- [39] Mahmudul Hasan Tahmida Mahmud and Amit K. Roy-Chowdhury. 2017. Joint prediction of activity labels and starting times in untrimmed videos. In *ICCV*.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- [41] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, and Hongyuan Zha. 2017. Wasserstein Learning of Deep Generative Point Process Models. In *NeurIPS*.
- [42] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M Chu. 2017. Modeling the Intensity Function of Point Process Via Recurrent Neural Networks. In *AAAI*.
- [43] M. Yao, S. Zhao, S. Sahebi, and R. Feyzi Behnagh. 2021. Stimuli-Sensitive Hawkes Processes for Personalized Student Procrastination Modeling. In *WWW*.
- [44] Alexander Richard Yazan Abu Farha and Juergen Gall. 2018. When Will You Do What? - Anticipating Temporal Occurrences of Activities. In *CVPR*.
- [45] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2015. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *arXiv preprint arXiv:1507.05738* (2015).
- [46] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. 2019. Time-series Generative Adversarial Networks. In *NeurIPS*.
- [47] Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. 2020. Self-attentive Hawkes processes. In *ICML*.
- [48] Q. Zhao, M. Erdogdu, H. He, A. Rajaraman, and J. Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *KDD*.
- [49] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer Hawkes Process. In *ICML*.