Supplementary Material for In Defense of LSTMs for addressing Multiple Instance Learning Problems

Kaili Wang[1], Jose Oramas[2], and Tinne Tuytelaars[1]

[1] KU Leuven, ESAT-PSI
{first name.last name}@esat.kuleuven.be
[2] University of Antwerp, imec-IDLab
{first name.last name}@uantwerpen.be

This documment constitutes supplementary material for the ACCV 2020 oral paper "In Defense of LSTMs for addressing Multiple Instance Learning Problems". It extends the original submssion in three fronts. First, it includes details of the network architectures used in each experiment (Sec. 1). Second, we provide additional examples for instance-level detection of the Colon cancer experiment (Sec. 2). Third, a presentation of the explanation capabilities of the proposed model is given (Sec. 3). Then, we show the derivation of the mutual information estimation used in our paper (Sec. 4). Finally, we show additional explanation examples for the MNIST-based experiment. (Sec. 5)

1 Architecture

In this section we present the architecture of the models we use in the submitted manuscript.

1.1 Single/Multi. Digit Occurrence and Digit Outlier Detection

The Instance Description Unit is based on LeNet [1]. The Iterative Bag Pooling Unit uses a Bi-direction LSTM with 500 dimension input and hidden state. The Prediction Unit is a classifier. Binary Cross-Entropy loss is used to supervise the MIL learning and we maximize the Mutual information between Layer 1 and 5.

1.2 Digit Counting

The only difference w.r.t. the previous setting is the Prediction Unit which is now a regressor. Accordingly, we the Mean Square Error (MSE) loss is used.

1.3 Cross-domain retrieval

The Instance Description Unit is based on VGG-16. The Iterative Bag Pooling Unit uses a Bi-direction LSTM with 2048 dimension input and cell state. The output of the LSTM (33-1) is used to learn the triplet embedding supervised by Triplet loss. The output of (33-2) is used to learn a classifier supervised by Cross-Entropy loss.

Layer	Type
1	conv(5,1,0)-20 + ReLU
2	maxpool(2,2)
3	conv(5,1,0)-50 +ReLU
4	maxpool(2,2)
5	fc-500 + ReLU
6	LSTM (500, 500)
7	Dropout(0.5)+ fc-1 + sign

Table 1: Architecture of our Single/Multi Digit Occurrence and Digit Outlier Detection model.

Layer	Type
1	conv(5,1,0)-20 + ReLU
2	maxpool(2,2)
3	conv(5,1,0)-50 +ReLU
4	maxpool(2,2)
5	fc-500 + ReLU
6	LSTM (500, 500)
7	Dropout(0.5)+ fc-1

Table 2: Architecture of our Digital Counting model.

1.4 Colon Cancer

The Instance Description Unit is based on [2]. The Iterative Bag Pooling Unit uses a Bi-direction LSTM with 512 dimension input and cell state. The Prediction Unit is a classifier. Binary Cross-Entropy loss is used to supervise the MIL learning and we maximize the Mutual information between Layer 1 and 7.

2 More examples for Instance level prediction

Here we show more examples for the instance level prediction of Colon cancer experiment by using our MIL model. Similar to Figure 5 in the main paper, the first column shows the original H&E image, the second and third column are the epithelial nuclei patches (Ground-Truth) and the epithelial nuclei patches detected by our MIL model, respectively.

3 Explaining Model Prediction

3.1 Methodology

In the main paper, we present an iterative method to make predictions from a bag-level representation S_j through the use of a prediction function $g(.)$. While

Layer	Type
1-31	VGG-16
32	ReLU + Dropout(0.5) + fc-2048
33-1	LSTM (2048, 2048)
33-2	ReLU + Dropout(0.5) + fc-2000

Table 3: Architecture of our Cross-domain retrieval model.

Layer	Type
1	conv(4,1,0)-36 + ReLU
2	maxpool(2,2)
3	conv(3,1,0)-48 +ReLU
4	maxpool(2,2)
5	fc-512 + ReLU
6	Dropout(0.5)
7	fc-512 + ReLU
8	Dropout(0.5)
9	LSTM (500, 500)
10	Dropout(0.5)+ fc-1

Table 4: Architecture of our Colon cancer model.

being able to make accurate predictions is of importance, being able to provide an explanation supporting the prediction made is a desirable property for any automatic system. In MIL algorithms, these explanations usually come in the form of highlighting the elements or instances x_i^* of the bag which determine the predicted bag label $\hat{y_j}$.

In the proposed approach this can be achieved by probing the bag representation S_j after each of the elements x_i are embedded on it. More specifically, on an initial step we can push every element x_i through the bag pooling unit and store the bag representation S_j^i computed after the embedding of the i^{th} element. Then, the relevant elements x_i^* can be highlighted by identifying the elements x_i with strong effect in the computed bag representation S_j^i. Finally, the selection of elements x_i^* can be further verified, by the response $\hat{y}_j^i{=}g(S_j^i)$ that their corresponding bag-level representations S_j^i produce when evaluated by the prediction unit.

3.2 Experiment

In this section we analyze the explanation capabilities of our method. Towards this goal, in Fig. 3 we show the predicted output after observing each element of the bag. Since the bag pooling unit utilizes a Bi-LSTM, which processes forward and backward directions of the bag together, we show the two directions of the bag. In addition, we verify the capabilities of the proposed bag representation to encode the underlying MI assumption. This could be indicated by reflecting

significant variations in the S_j^i when observing the elements x_i involved in the MI assumption. We ease the visualization of the high-dimensional S_j^i representation by plotting its corresponding t-SNE [3] projection in Fig. 2.

Discussion: In Fig. 3, we can notice that each time one of the elements that determine the MI assumption are observed, the bag representation S_j^i is updated in such a way that there is a significant change in the prediction made by the model. This is further supported by the state of the internal representation S_j^i as shown by the corresponding t-SNE visualizations (Fig. 2). For the *Single Digit Occurrence* case, we notice that the representation gets updated to a different region of the space when the digit of interest is observed. More specifically, from the third row of Fig. 2, it is clear that the space is divided into two parts: the bag representation of negative bags changes within the bottom-left region, while for positive bags, once the digit of interest occurs, the bag representation jumps to the top-right region and ends there. Similarly, for the *Multiple Digit Occurrence*, *Digit Sequences* and *Digit Counting* cases, we notice that the representation shifts, significatively, to specific regions (green and magenta dots) every time one of the digits of interest is observed. Moreover, for *Multiple Digit Occurrence* and *Digit Sequences* the representation seems to always reach a common region once the underlying MI assumption has been completely satisfied.

4 Mutual Information Maximization

We follow the work [4] to estimate the lower boundary of the mutual information between the input and its latent representation. We advice readers to read the original paper [4] for a comprehensive understanding.

We take the derivations from [4] here.

Assume we have inputs X and their latent representation Z. The definition of mutual information is defined as:

$$\begin{aligned} I(X,Z) &= \iint p(z|x)p(x)\log\frac{p(z|x)}{p(z)}dxdz \\ &= KL(p(z|x)p(x)||p(z)p(x)), \end{aligned} \tag{1}$$

where $p(x)$ is the distribution of the inputs, $p(z|x)$ is the distribution of the corresponding latent representations, $p(z)$ is the distribution of the latent space. $p(z) = \int p(z|x)p(x)dx$. In order to maximize the mutual information $I(X,Z)$, we have:

$$p(z|x) = \max_{\theta_e} I(X,Z). \tag{2}$$

$p(z)$ is difficult to calculate, therefore, we try to find an auxiliary distribution $q(z)$ to approximate $p(z)$. We assume $q(z)$ is standard normal distribution. To measure the distance of two distribution, we use KL divergence:

$$KL(p(z)||q(z)) = \int p(z)log\frac{p(z)}{q(z)}dz. \tag{3}$$

According to Eqs. (2) and (3), we have:

$$p(z|x) = \min_{\theta_e} \left\{ - \iint p(z|x)p(x) \log \frac{p(z|x)}{p(z)} dxdz + \alpha \int p(z) \log \frac{p(z)}{q(z)} dz \right\}. \tag{4}$$

It can be further rewritten as:

$$p(z|x) = \min_{\theta_e} \left\{ \iint p(z|x)p(x)[-(\alpha+1) \log \frac{p(z|x)}{p(z)} + \alpha \log \frac{p(z|x)}{q(z)}] dxdz \right\}. \tag{5}$$

According to Eq. (1), the Eq. (5) can be viewed as:

$$p(z|x) = \min_{\theta_e} \{ -\beta I(X,Z) + \gamma \mathbb{E}_{x \sim p(x)}[KL(p(z|x)||q(z))] \}. \tag{6}$$

[4] chooses JS divergence for mutual information maximization since there is no boundary for KL divergence:

$$p(z|x) = \min_{\theta_e} \{ -\beta JS(p(z|x)p(x), p(z)p(x)) + \gamma \mathbb{E}_{x \sim p(x)}[KL(p(z|x)||q(z))] \}. \tag{7}$$

The variational estimation of JS divergence [5] is defined as:

$$JS(p(x)||q(x)) = \max_{T} (\mathbb{E}_{x \sim p(x)}[\log \sigma(T(x))] + \mathbb{E}_{x \sim q(x)}[\log(1 - \sigma(T(x)))]). \tag{8}$$

where $T(x) = \log \frac{2p(x)}{p(x)+q(x)}$ [5]. Here $p(z|x)p(x)$ and $p(z)p(x)$ are utilized to replace $p(x)$ and $q(x)$. As a result, Eq. (7) can be defined as:

$$p(z|x) = \min_{\theta_e} \{ -\beta(\mathbb{E}_{(x,z) \sim p(z|x)p(x)}[\log \sigma(T(x,z))] + \mathbb{E}_{(x,z) \sim p(z)p(x)}[\log(1 - \sigma(T(x,z)))]) + \gamma \mathbb{E}_{x \sim p(x)}[KL(p(z|x)||q(z))] \}. \tag{9}$$

To solve the problem in Eq. (9), [4] considers to use Negative sampling estimation [6]. [6] uses a discriminator to distinguish the real and noisy samples to estimate the distribution of real samples. Therefore, $\sigma(T(x,z))$ can be treated as a discriminator, which is trained to distinguish the positive pairs (x and corresponding latent representation z) and negative pairs (x and its latent representation z with random disturbance in batch dimension). Eq. (9) represents the global mutual information between X and Z.

Follow the idea from [6], [4] also consider to maximize the local mutual information, where original input is replaced by the middle layer of the convolutional network. In the end, both global and local mutual information are maximized, the final objective function is:

$$\begin{aligned} L = & -\beta(\mathbb{E}_{(x,z)\sim p(z|x)p(x)}[\log\sigma(T_1(x,z))] \\ & + \mathbb{E}_{(x,z)\sim p(z)p(x)}[\log(1-\sigma(T_1(x,z)))]) \\ & - \frac{\beta}{hw}\Sigma_{i,j}(\mathbb{E}_{(x,z)\sim p(z|x)p(x)}[\log\sigma(T_2(C_{ij},z))] \\ & + \mathbb{E}_{(x,z)\sim p(z)p(x)}[\log(1-\sigma(T_2(C_{ij},z)))]) \\ & + \gamma\mathbb{E}_{x\sim p(x)}[KL(p(z|x)||q(z))], \\ & = \alpha\cdot MI_{global} + \beta\cdot MI_{local} + \gamma\cdot PriorMatching \end{aligned} \tag{10}$$

where h and w represent the height and width of the feature map. C_{ij} represents the feature vector of the middle feature map at coordinates (i, j) and $q(z)$ is the standard normal distribution.

5 Additional Explanation Examples

Figure 4 shows more t-SNE visualization of the learned bag representation in ***Single Digit Occurrence*** experiment. We binarized the visualized digit bag, '0' refers to the non-interest digits while '1' refers to the digit of interest. ('9' in this experiment). It is clear that the space is divided into two parts: the bag representation of negative bags changes within the bottom-left region, while for positive bags, once the digit of interest occurs, the bag representation jumps to the top-right region and ends there.

Figure 5 displays more t-SNE visualization of the learned bag representation in ***Multiple Digit Occurrence*** experiment. '0' refers to the non-interest digits. The first small figure shows the prediction of 20 examples overlaid on the t-SNE space. It is clear that the representation shifts to *green dots region* or *magenta dots region* every time when *digit '3'* or *digit '6'* is observed. In addition, representation seems to always reach a common region (here is the right area) once the underlying MI assumption has been completely satisfied as long as the last digit is not one of the interest ones.

Figure 6 shows more t-SNE visualizations of the learned bag representation in the ***Digit Sequence*** experiment. Similar trend can be observed in this experiment. In addition, when the model observes the wrong order of the digits of interest, even if the representation jumps to the specific regions, the end point is different from that that is reached in case the correct order is observed. This suggests that there is a region in the space to indicate whether the underlying assumption has been satisfied.

Figure 7 shows more t-SNE visualizations of the learned bag representation in the ***Digit Counting*** experiment. Please note the *green dots* is the region where the representations shift to when the digit of interest is observed.

References

1. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE (1998)
2. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Transactions on Medical Imaging (2016)
3. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research (2008)
4. Yang, X., Deng, C., Zheng, F., Yan, J., Liu, W.: Deep spectral clustering using dual autoencoder network. CVPR (2019)
5. Nowozin, S., Cseke, B., Tomioka, R.: f-gan: Training generative neural samplers using variational divergence minimization. In: Advances in Neural Information Processing Systems (NIPS). (2016)
6. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: International Conference on Learning Representations. (2019)

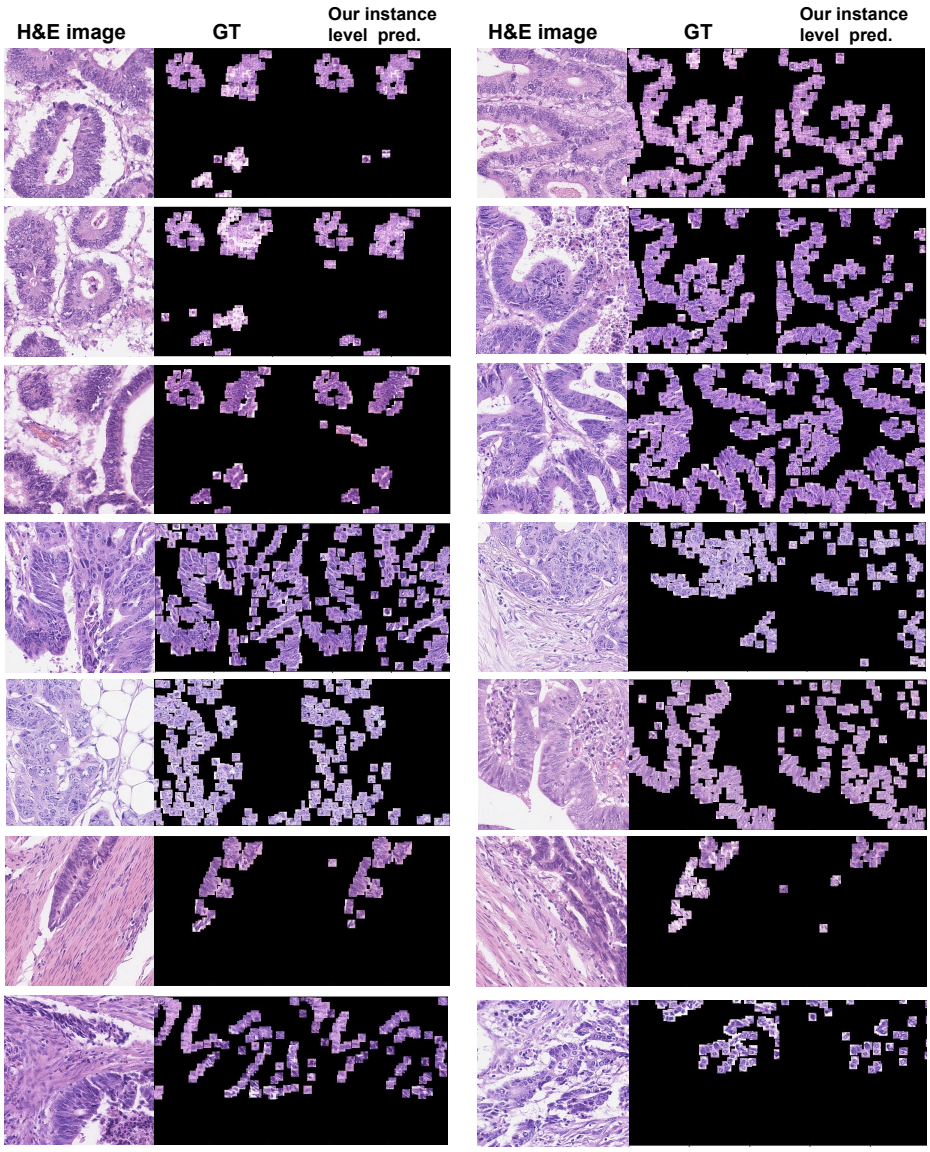

Fig. 1: More examples for Instance level prediction of Colon cancer experiment by using our MIL model.

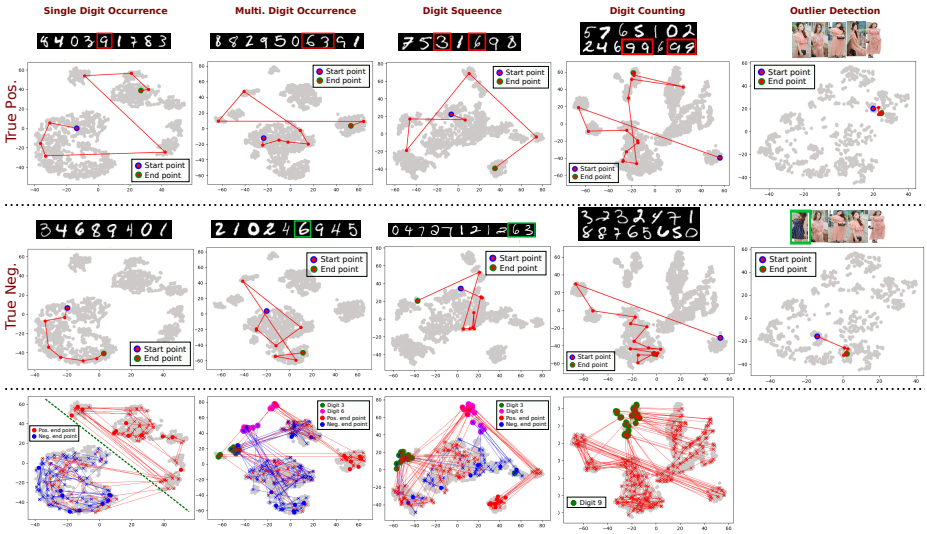

Fig. 2: t-SNE visualization of the learned bag representation. The first two rows show examples of predictions on true positive and true negative bags, except for the *Digit Counting* experiment, which shows two bags containing 4 and 0, elements of interest, respectively. The third row shows the prediction of 20 examples overlaid on the t-SNE space for the digit-based experiments.

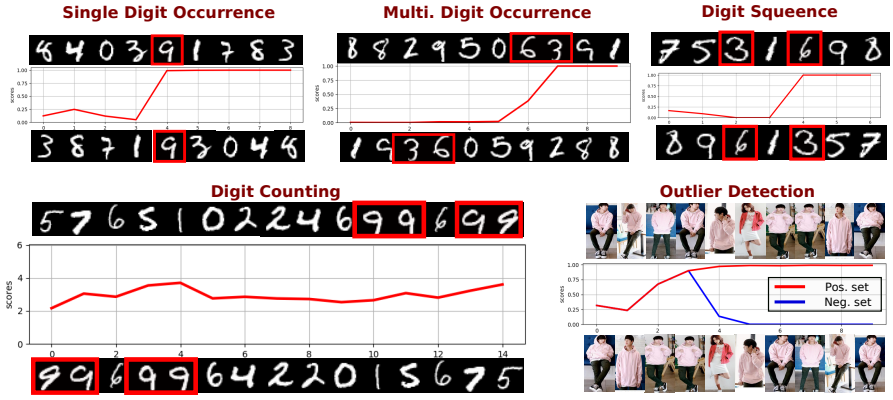

Fig. 3: Prediction score post bag representation update after observing each element from the bag. For reference, we present both the forward (top) and backward (bottom) directions in which the elements of the bag are observed.

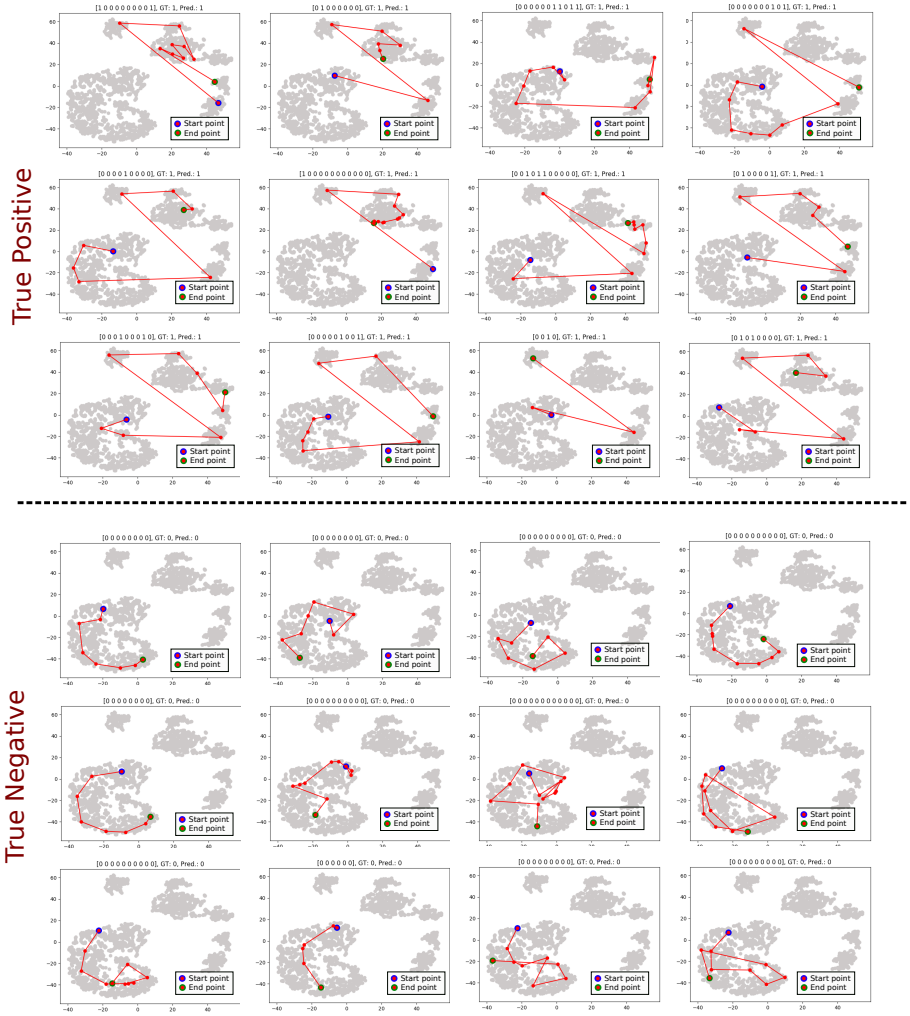

Fig. 4: t-SNE visualization of the learned bag representation in ***Single Digit Occurrence*** experiment.On top of each t-SNE representation, the visualized bag is presented in binary form where '0' refers to the background digits while '1' refers to the [witness] digit of interest ('9' in this experiment).

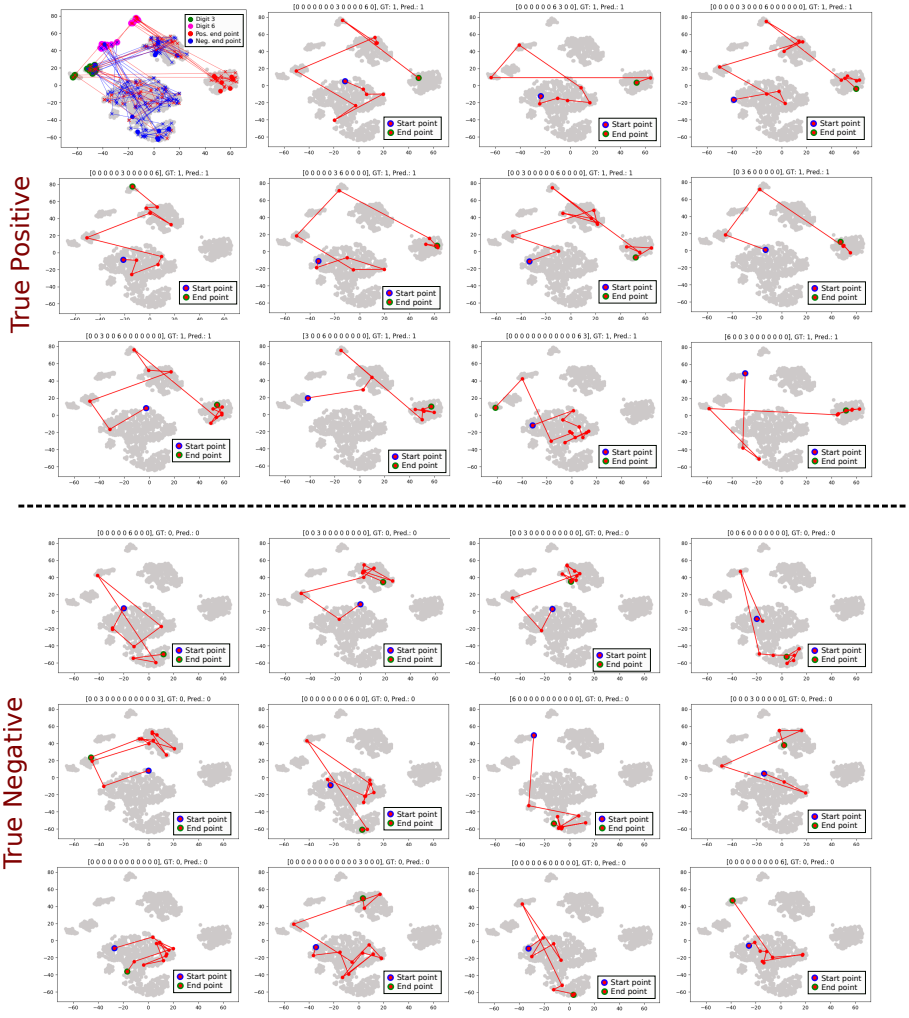

Fig. 5: t-SNE visualization of the learned bag representation in ***Multiple Digit Occurrence*** experiment. On top of each t-SNE representation, '0' refers to the background digits.

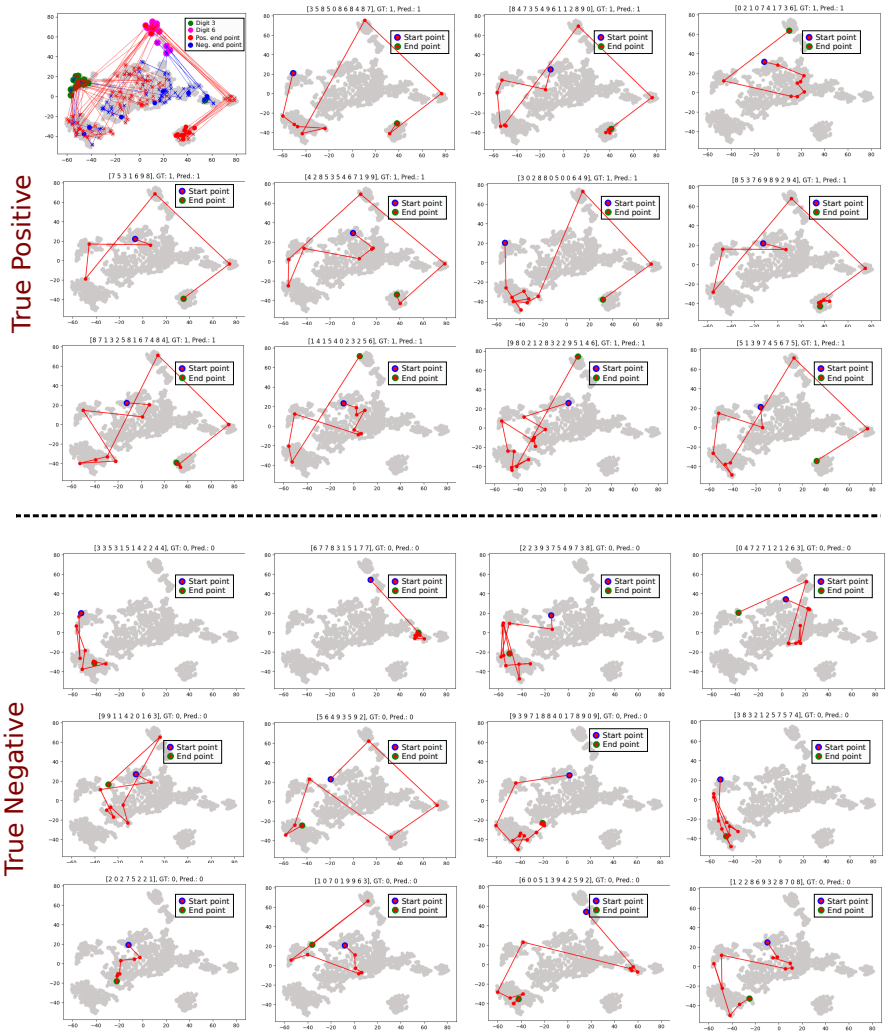

Fig. 6: t-SNE visualization of the learned bag representation in ***Digit Sequence*** experiment.

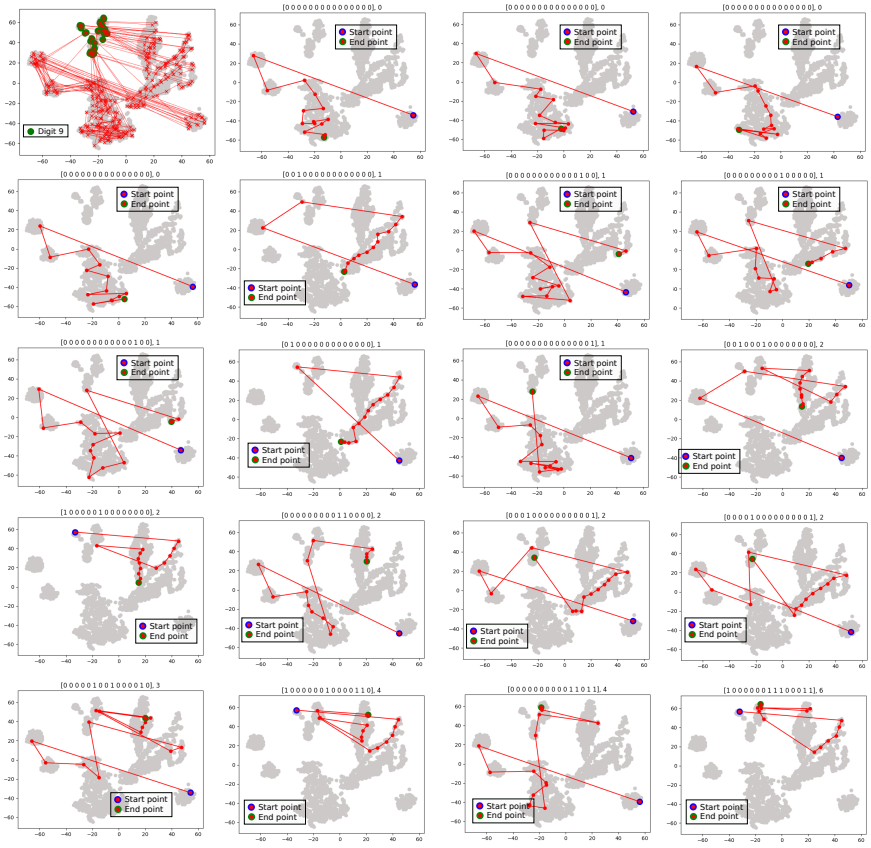

Fig. 7: t-SNE visualization of the learned bag representation in ***Digit Counting*** experiment. On top of each t-SNE representation, the visualized bag is presented in binary form where '0' refers to the background digits while '1' refers to the [witness] digit of interest ('9' in this experiment).