Towards Robust and Discriminative Sequential Data Learning: When and How to Perform Adversarial Training?*

Xiaowei Jia[1*], Sheng Li[2], Handong Zhao[3], Sungchul Kim[3], Vipin Kumar[1]
[1]Department of Computer Science and Engineering, University of Minnesota
[2]Department of Computer Science, University of Georgia
[3]Adobe Research
[1]{jiaxx221,kumar001}@umn.edu, [2]sheng.li@uga.edu, [3]{hazhao,sukim}@adobe.com

ABSTRACT

The last decade has witnessed a surge of interest in applying deep learning models for discovering sequential patterns from a large volume of data. Recent works show that deep learning models can be further improved by enforcing models to learn a smooth output distribution around each data point. This can be achieved by augmenting training data with slight perturbations that are designed to alter model outputs. Such adversarial training approaches have shown much success in improving the generalization performance of deep learning models on static data, *e.g.,* transaction data or image data captured on a single snapshot. However, when applied to sequential data, the standard adversarial training approaches cannot fully capture the discriminative structure of a sequence. This is because real-world sequential data are often collected over a long period of time and may include much irrelevant information to the classification task. To this end, we develop a novel adversarial training approach for sequential data classification by investigating *when* and *how* to perturb a sequence for an effective data augmentation. Finally, we demonstrate the superiority of the proposed method over baselines in a diversity of real-world sequential datasets.

KEYWORDS

Adversarial training, sequential data, attention mechanism

ACM Reference Format:
Xiaowei Jia[1*], Sheng Li[2], Handong Zhao[3], Sungchul Kim[3], Vipin Kumar[1]. 2019. Towards Robust and Discriminative Sequential Data Learning: When and How to Perform Adversarial Training?. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330957

1 INTRODUCTION

The study of sequential data has become important nowadays given the prevalence of sequential data in multiple applications, *e.g.,* sequential user behavior modeling in commercial applications [4, 17], semantic analysis in natural language processing [6, 36], and remote sensing for land cover detection [15, 40]. Consider the analysis of web browsing data in the domain of e-Commerce, which serves as one of primary motivations of this work. Effective analysis of the sequential browsing history assists in better predicting purchase behaviors and helps ensure better personalized experience.

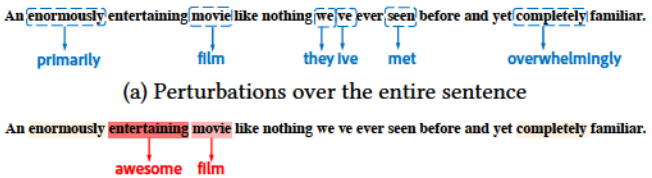

Figure 1: Illustrative figure of a short movie review from Rotten Tomatoes [34] with adversarial perturbations over (a) the entire sentence, versus on (b) the proposal model with discriminative words "entertaining movie". Darker red denotes higher attention score, and vice versa. Particularly, the attention scores of the two most discriminative words "entertaining" and "movie" are 0.62 and 0.12, respectively. We omit coloring the words whose scores are smaller than 0.1.

Advances in deep learning algorithms have provided unrealized potential for classifying sequential data. Most innovations come from the domain of natural language processing, where the recurrent neural networks (RNN) and Convolutional Neural Networks (CNN)-based models are frequently used for a variety of learning tasks, such as machine translation [25], sentiment classification [6, 18], and question answering [21, 48]. Given the success of these techniques in analyzing sequential data, they have been further applied to other applications [4, 9, 15]. The power of these deep learning approaches stems from their ability to model underlying dependencies between data at different time steps. These dependencies across time can provide the context information that enables learning of different sequential behaviors even for two identical data points at a single time step. The modeling of temporal dependencies can also facilitate the extraction of informative feature representation from the entire sequence.

Recent works have shown that the performance of deep learning models can be further improved by augmenting data with deliberate perturbations [11, 31]. These perturbations are selected to adversely alter the model outputs (i.e., along the anisotropic direction), and they are commonly referred to as adversarial perturbations. The learning model is then updated to smooth the output distribution over these perturbed samples. After multiple iterations of data

*This work was done during the author's internship at Adobe Research.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330957

perturbation and model update, this data augmentation strategy can help learn a smooth distribution along every direction around each data point (i.e., an isotropically smooth output distribution). For example, the popular adversarial training (AT) approach [11] adopts the linear-approximation of adversarial perturbations to update model parameters such that the model predictions are not easily altered by the perturbations. This method has shown to successfully improve the generalization performance and the model robustness against adversarial perturbations.

However, existing AT-based approaches mainly focus on static data and do not take into account the structural property of sequential data. Real-world sequential data are often collected over a long period of time while only part of the sequence is critical to the classification. Consider the example of purchase prediction from web browsing data. A user decides to buy a product mostly because he/she is attracted by several relevant web pages to this product while he/she may randomly click on many other web pages. Hence, these important time steps (i.e., web pages) in sequential data contain more discriminative information for purchase prediction. Similarly, a movie review can be a long sentence, but only a few words in this sentence can reflect the sentiment. Consider the sentence in Fig. 1, the words "entertaining movie" can clearly reveal that this is a positive review. Prior work has applied AT to sequential data by directly imposing conventional adversarial perturbations to the entire sequence [30]. However, as shown in Fig. 1 (a), this conventional AT method replaces multiple words that are not relevant for sentiment. Hence, it is less helpful for data augmentation since these perturbed positions are not critical for classification.

To develop an adversarial training approach for sequential data, we need to answer specifically *when* and *how* to perturb the sequence. In particular, we build a robust sequence classification model in two stages. The entire framework is depicted in Fig. 2. First, the proposed model conducts classification with more emphasis on the discriminative periods detected by an LSTM-Attention model (see two panels in Fig. 2). The sequential patterns within these discriminative periods are more influential for the class label than those in other periods. Second, we propose an attention-aware AT method so as to improve the robustness against perturbations on these discriminative periods (see the yellow and blue arrows and the dashed box at the bottom of Fig. 2). In this way, we augment training data by adding more variability to the discriminative period. According to Fig. 1 (c), the proposed method perturbs mostly on the discriminative part of a sequence while also maintaining the semantics of the sentence. This is equivalent to augmenting the training data by adding another training sentence with different but similar discriminative words while maintaining the original meaning.

Furthermore, an effective adversarial training requires that the perturbed data are perceptually similar to original data while they can still alter the model output. Such perceptual similarity is commonly defined based on certain structural property of data. For example, adversarial perturbations in hand-written digits recognition should preserve the structure of digits in the image [38]. Similarly, the perturbations to graph data should preserve graph structural property and node attributes [49]. For sequential data, it is essential to maintain the position of discriminative periods since a successful classifier always needs to refer to these periods for informative patterns. Hence, we define the perceptual similarity of sequential data based on the obtained attention weights from attention model, which reflect the position of discriminative periods. The failure to preserve such perceptual similarity is likely to result in a drastic change of sequential structure and the semantic meaning of data, and thus cannot be used for effective data augmentation.

Since the perturbations used for AT are added to multiple time steps, it is likely that these perturbations can jointly break the perceptual similarity. To address this issue, we propose two algorithms, Selective Filtering (SF) and Conservative Adversarial Training (CAT), to enforce that all the adversarial samples used for training classification model preserve the perceptual similarity.

We evaluate the proposed method on real-world seqeuntial datasets from threee different domains - Digital Marketing dataset, Movie Review dataset and Remote Sensing dataset. Our results demonstrate the superiority of our proposed method over multiple baselines in each domain. We also provide several examples to illustrate the efficacy of the proposed method.

Our contributions can be summarized as follows:

- We propose a novel approach to augment sequential data over the discriminative periods. This technique can be applied to both labeled data and unlabeled data.
- During the augmentation process, we propose two algorithms to effectively prevent drastic change of discriminative structure within the sequence.
- Our implementations in real-world datasets validate the effectiveness on augmenting data for a diversity of domains, including e-Commerce, text and remote sensing.

2 METHOD

2.1 Preliminaries

2.1.1 Adversarial training. Adversarial training (AT) has shown to be able to improve the performance of sophisticated models by learning an isotropically smooth output distribution around each training data point. The AT approach introduces a regularizer to update model parameters such that model outputs cannot be easily altered by adversarial perturbations. Adversarial perturbations are often instantiated as an extra noise r added to the input data x. Given input features x from labeled data X_l and corresponding training labels y, the noise factor r is selected within a small range ϵ such that the training loss $L(x+r, y; \theta)$ is maximized after the perturbation. To improve the model robustness against such perturbations, the AT approach updates model parameters θ to reduce training loss even after adversely perturbing the data. More formally, this can be expressed as a min-max problem as follows:

$$\begin{aligned} &\min_{\theta} L(x, y; \theta) + \lambda L(x + r_{\text{adv}}, y; \theta), \\ &r_{\text{adv}} = \text{argmax}_{r, ||r|| < \epsilon} L(x + r, y; \theta), \end{aligned} \tag{1}$$

where λ is a hyper-parameter to control the balance between the standard training loss and the adversarial training loss.

Assuming L_2 norm constraint is adopted in Eq. 1, previous research [11] computes r by a single-step linear approximation, as:

$$r_{\text{adv}} = \epsilon \frac{\nabla_x L(x, y; \theta)}{||\nabla_x L(x, y; \theta)||}. \tag{2}$$

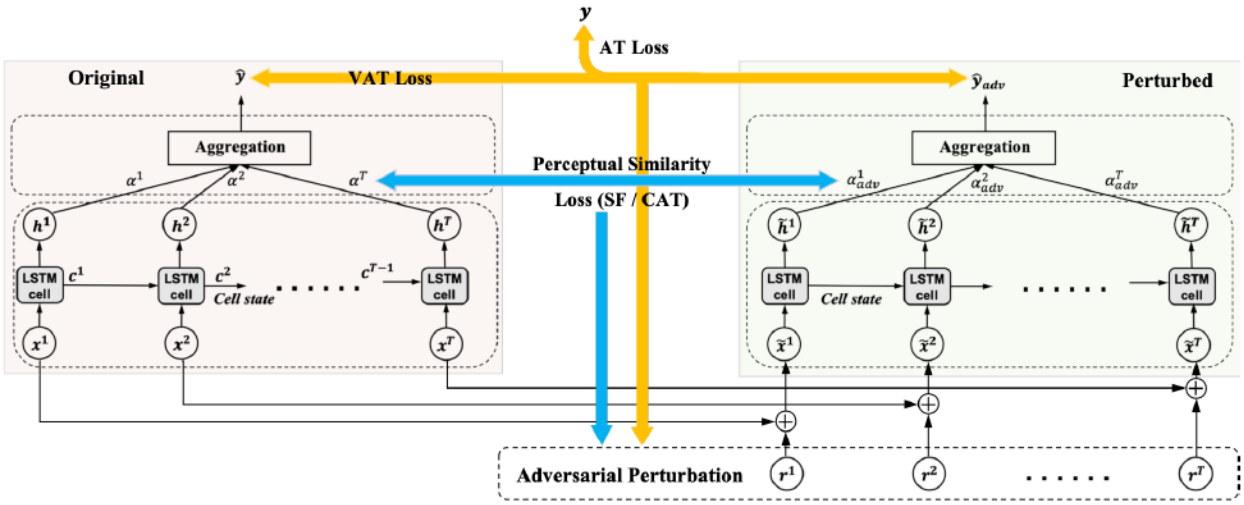

Figure 2: The flow of the proposed learning framework. The input sequential data (x^1,...x^T) and perturbed data ($\tilde{x}^1$,...$\tilde{x}^T$) are classified using LSTM-Attention networks. The adversarial perturbations (r^1,...,r^T) are produced based on the loss of AT/VAT and the loss of the perceptual similarity. The proposed method SF and CAT estimate the loss of perceptual similarity using the original attention weights (α^1,...,α^T) and the attention weights after perturbation (α^1_{adv},...,α^T_{adv}).

This standard AT method can only be applied to labeled data. To overcome this limitation, Virtual Adversarial Training (VAT) approach [31] is proposed, which is an unsupervised extension of AT. VAT computes the adversarial perturbations using the predicted labels $p(\hat{y}|x;\theta)$ instead of the ground-truth labels y. Then it measures the Kullback-Leibler (KL) divergence between the predicted labels before and after perturbing data. VAT can be applied over both labeled data X_l and unlabeled dat X_u, as follows:

$$\min_{\theta} L(x,y;\theta)+\lambda_{\text{vat}}\frac{1}{N_l+N_u}\sum_{X_l\bigcup X_u}\text{KL}[p(\hat{y}|x;\theta_c)||p(\hat{y}|x+r_{\text{vat}};\theta)],$$
$$r_{\text{vat}} = \text{argmax}_{r,||r||\le\epsilon}\text{KL}[p(\hat{y}|x;\theta_c)||p(\hat{y}|x+r;\theta)], \tag{3}$$

where N_l and N_u represent the number of labeled samples and unlabeled samples, respectively. We use θ_c to represent the fixed model parameters in the current iteration, which are not involved in the model update.

2.1.2 LSTM-Attention model. The LSTM-Attention model [25, 45] aims to detect discriminative periods from a sequence and learn representative temporal patterns for classification. Given a sequence of input $\{x^1, x^2, ...x^T\}$, where $x^t \in \mathbb{R}^D$, the LSTM model generates hidden representations $h^t \in \mathbb{R}^H$ at every time step. The hidden representations at different time steps are then combined by the attention model via weighted summation for the final classification. We now briefly introduce the LSTM model and the attention model.

In essence, the LSTM model defines a transition relationship for hidden representation h^t through LSTM cells. Each LSTM cell contains a cell state c^t, which serves as a memory and allows the hidden units h^t to reserve information from the past. The transition of cell state over time forms a memory flow, which enables the modeling of long-term dependencies. Specifically, the LSTM first generates a candidate cell state $\tilde{c}^t$ by combining x^t and h^{t-1} into a tanh(·) function, as follows:

$$\tilde{c}^t = \tanh(W^c_h h^{t-1} + W^c_x x^t). \tag{4}$$

where $W^c_h \in \mathbb{R}^{H\times H}$ and $W^c_x \in \mathbb{R}^{H\times D}$ are weight parameters used to generate candidate cell state. Hereinafter we omit the bias terms as they can be absorbed into weight matrices.

Then a forget gate layer $f^t \in \mathbb{R}^H$ and an input gate layer $g^t \in \mathbb{R}^H$ are generated using sigmoid functions:

$$\begin{aligned} f^t &= \sigma(W^f_h h^{t-1} + W^f_x x^t), \\ g^t &= \sigma(W^g_h h^{t-1} + W^g_x x^t), \end{aligned} \tag{5}$$

where $\{W^f_h \in \mathbb{R}^{H\times H}, W^f_x \in \mathbb{R}^{H\times D}\}$ and $\{W^g_h \in \mathbb{R}^{H\times H}, W^g_x \in \mathbb{R}^{H\times D}\}$ denote two sets of weight parameters for generating forget gate layer f^t and input gate layer g^t, respectively. The forget gate layer is used to filter the information inherited from c^{t-1}, and the input gate layer is used to filter the candidate cell state at time t by entry-wise product $\otimes$. In this way we obtain the new cell state c^t as follows:

$$c^t = f^t \otimes c^{t-1} + g^t \otimes \tilde{c}^t. \tag{6}$$

Then we generate the hidden representation h^t by filtering the obtained cell state using an output gate layer o^t, as:

$$\begin{aligned} o^t &= \sigma(W^o_h h^{t-1} + W^o_x x^t), \\ h^t &= o^t \otimes \tanh(c^t), \end{aligned} \tag{7}$$

where $W^o_h \in \mathbb{R}^{H\times H}$ and $W^o_x \in \mathbb{R}^{H\times D}$ are the weight parameters used to generate the hidden gate layer.

After obtaining the hidden representations $\{h^1, ..., h^T\}$ from LSTM, we utilize the attention model to determine the discriminative period from the sequential data. The attention model aims to enforce the classifier to attend to different time steps with different attention weights. The higher attention weight at a time step indicates more expressive discriminative knowledge at this time step.

Specifically, we measure the attention weight for time t according to the similarity between its hidden representation h^t and a sequence embedding $v \in \mathbb{R}^H$. Here v represents an embedding of

the entire sequence, which has the same dimensionality with the hidden representation, and is jointly learned during the training process [16, 45]. In the simplest case, we can embed $x^{1:T}$ into v using another LSTM.

More formally, the attention weight for time step t is computed as the inner-product between v and h^t. To normalize attention weights over all the time steps, we apply a softmax function on the inner-product, as follows:

$$\alpha^t = \text{softmax}(v \cdot h^t). \tag{8}$$

Then we aggregate h^t from all the time steps based on α^t, and apply a softmax function to compute predicted output:

$$\hat{y} = \text{softmax}(W_y \sum_t \alpha^t h^t), \tag{9}$$

where $W_y \in \mathbb{R}^{C \times H}$ denotes the parameters to transform aggregated hidden representation to the output.

Given the labeled data X_l, we train the LSTM-Attention networks by minimizing the cross-entropy training loss, as:

$$L_{\text{att}} = -\frac{1}{N_l} \sum_{i \in X_l} \sum_k y_{i,k} \log \hat{y}_{i,k}, \tag{10}$$

where the provided label y is expressed in a one-hot representation where $y_{i,k} = 1$ if the i^{th} sample from the source domain belongs to the class k, for k=1 to C.

2.2 Adversarial Training for Sequential Data

We first describe the proposed method for augmenting data on the discriminative period, i.e., *when*. Then we propose two algorithms to preserve perceptual similarity of sequential data, i.e., *how*.

2.2.1 Attention-aware adversarial training. When applied to sequential data, the standard AT approach simultaneously perturbs all the time steps. Such augmentation strategy is less effective for sequential data because the perturbations are very likely to be distributed over non-informative periods.

In this work, we utilize the LSTM-Attention networks to detect the discriminative period. Intuitively, the attention weights indicate the part of sequence that is most critical for the classification. Then we add adversarial perturbations to enrich training data with more variability within the detected discriminative period. After such augmented training, we anticipate the learned model to be robust against any slight perturbations to the discriminative patterns in the sequence.

Specifically, given the attention weights learned from the LSTM-Attention networks, we allow a larger magnitude for adversarial perturbations on more discriminative period. The model is then updated to prevent such perturbations, even selected along the anisotropic directions, to alter the classification outputs.

The final training loss function combines the standard supervised loss and the adversarial loss of augmented samples, which is similar to Eq. 1. If we concatenate the perturbations on all the time steps, i.e., $r = [r^1, r^2, ..., r^T]$, the proposed approach can be summarized as follows:

$$\begin{aligned} &\min_\theta L_{\text{att}}(x, y; \theta) + \lambda L_{\text{att}}(x + r_{\text{adv}}, y; \theta) \\ &r_{\text{adv}} = \text{argmax}_r L_{\text{att}}(x + r, y; \theta), \\ &\text{s.t.} \, ||r^t|| < \alpha^t \epsilon, \quad \text{for} \quad t = 1 \text{ to } T. \end{aligned} \tag{11}$$

From the above equation, we can observe that the allowed perturbation radius is larger for the time steps with higher attention weights. It is noteworthy that attention weights are used for both data augmentation and classification. Hence, the data are augmented over the time periods that seamlessly fit the classification process.

This method can also be extended to unlabeled data. Since the learning model can automatically detect the discriminative period without requiring training labels, we can add VAT-based perturbations over the detected discriminative period, i.e., $||r^t_{\text{vat}}|| < \alpha^t \epsilon$. Then we regularize the KL-divergence between $p(\hat{y}|x; \theta_c)$ and $p(y|x + r_{\text{vat}}; \theta)$.

2.2.2 Perceptual similarity. The adversarial perturbations aim to modify the input data with unnoticeable changes such that the perturbed data cannot be easily differentiated from original data by human but get misclassified by the learning model. Hence, these perturbations should preserve certain important structural properties of data (*e.g.,* the structure of handwritten digits [38] for digits recognition) while pursuing the directions that greatly impact the predictions. If the perturbed data meet this standard, we claim that the perturbed data preserve the perceptual similarity with original data.

For sequential data, we expect that the position of discriminative periods in the sequence should be maintained after applying perturbations. If the time steps within the discriminative period become less important after perturbations, it is highly likely that the structure of entire sequence varies drastically such that the discriminative patterns of the original class cannot be reflected.

Since we add adversarial perturbations to multiple time steps, it is possible that these perturbations jointly result in a severe impact to the discriminative sequential patterns. To address this problem and maintain the perceptual similarity, we propose two approaches, Selective Filtering (SF) and Conservative Adversarial Training (CAT). These methods utilize the obtained attention weights to measure the position of discriminative time steps. To select proper perturbations, they aim to control the variation of attention weights after adversarial perturbations. We now describe these two approaches as follows:

Selective Filtering (SF): In each iteration, we first generate all the adversarial perturbations for every sample x by maximizing the loss function $L_{\text{att}}(x + r, y; \theta)$. Then we remove adversarial samples that lead to large variation of attention weights. Specifically, we compute $\mathcal{R} = \sum_t \alpha^t \log \alpha^t_{\text{adv}}$ for each sample, where α and α_{adv} denote the obtained attention weights before and after we apply adversarial perturbations, respectively. A large value of $\mathcal{R}$ indicates that the perturbed sequence has its attention weights α_{adv} similar to the attention weights α of original sequence. Then, we select the adversarial samples with top $K\%$ of $\mathcal{R}$ values for adversarial training and filter out the remaining adversarial samples that can potentially break the perceptual similarity. If we represent the selected samples as X_{sf}, the training objective can be rewritten as:

$$\begin{aligned} &\min_\theta L_{\text{att}}(x, y; \theta) + \lambda_{\text{sf}} L_{\text{sf}}(x + r_{\text{adv}}, y; \theta), \\ &L_{\text{sf}}(x + r_{\text{adv}}, y; \theta) = -\frac{1}{|X_{\text{sf}}|} \sum_{i \in X_{\text{sf}}} \sum_k y_{i,k} \log \hat{y}_{i,k} \end{aligned} \tag{12}$$

Conservative Adversarial Training (CAT): Rather than filtering all the generated perturbations by the SF approach, an alternative solution is to select perturbations in a more conservative way such that the generated perturbations r_{adv} can already maintain the perceptual similarity. Formally, we first define a loss for the variation of attention weights according to $\mathcal{R}$. Then, we select the adversarial perturbations which maximize the adversarial loss while minimizing the variation of attention weights, as follows:

$$\begin{aligned} r_{\text{adv}} &= \text{argmax}_r L_{\text{att}}(x + r, y; \theta) + \gamma \sum_t \alpha^t \log \alpha^t_{\text{adv}}, \\ \text{s.t. } &||r^t|| < \alpha^t \epsilon, \quad \text{for } t = 1 \text{ to } T, \end{aligned} \tag{13}$$

where γ is a hyper-parameter to balance between maximizing the training loss and maintaining perceptual similarity. The obtained perturbations r_{adv} are then used for adversarial training.

3 EXPERIMENT

In this section, we evaluate the proposed method on three real-world datasets from different domains. We begin by describing these datasets. Then we present our results and discussion pertaining to the efficacy of the proposed framework on each dataset separately.

3.1 Dataset description

Digital Marketing Dataset: This dataset is collected from Adobe.com, and contains web browsing records for different users. Our goal is to predict whether each user finally purchases a product. In total there are 54,155 sequences in the dataset and 3,388 of them finally lead to a conversion (i.e., a purchase behavior). Such data skewness can bring difficulties for effectively training many conventional learning models. The maximum length of sequence in this dataset is 732 and the average length is 29.

Movie Review Dataset: This dataset [34] consists of short movie reviews from Rotten Tomatoes for sentiment classification. There are totally 10,662 short movie reviews. 5,331 reviews are positive and the other 5,331 reviews are negative. The maximum length of reviews is 53 and the average length is 20.

Remote Sensing Dataset: This dataset is a subset of MODIS MOD09A1 multi-spectral data product [1] collected by MODIS instruments onboard NASA's Terra satellites. It provides global data for every 8 days at 500m spatial resolution. At each date, MODIS dataset provides reflectance values on 7 spectral bands for every location. Here we use MODIS product since it provides better temporal coverage for discovering sequential patterns. For other higher-resolution data (*e.g.*, Landsat and Sentinel), they are available much less frequently and over half of the images are blocked by clouds. To better capture short-term patterns, we concatenate spectral features in every 32-days window as a time step and slide the window by 8 days. Totally we have 43 time steps in a year, and the feature dimension is $7 \times 4 = 28$. Our objective is to study temporal growth patterns to distinguish between corn and soybean in 2016 in southwestern Minnesota. Corn and soybean are major crop types that take over 90% cropland area in this region. In total we study 25,139 sequences from different locations, in which 11,911 locations correspond to corn and 13,228 locations correspond to soybean. The ground-truth labels are created by combining ground survey and USDA crop data layer [3].

For each dataset, we utilize 50% data for training and the remaining for testing. Besides the AT-based approach for the supervised learning task, we also extend it to the transductive semi-supervised learning scenario where we apply VAT-based regularization for the unlabeled test data. In the following we present the results on each dataset separately. The selection of hyper-parameters is discussed separately for each dataset in the following sections.

3.2 Evaluation on the Digital Marketing Dataset

Since each web browsing record is a discrete index of web page, we first embed these discrete behaviors into real-valued vectors. Specifically, by treating each browsed web page as a "word" and each sequence of browsing records as a "sentence", we embed each browsed web page into a 120-dimension vector using the Word2Vec [29] technique. Then we fine-tune the embeddings in the classification process. The adversarial perturbations are directly applied to these continuous embeddings. we set the dimension of hidden representation in the LSTM-Attention model to be 70. The λ in Eq. 11 is set to 1.0. In SF, we use 80% adversarial samples in each iteration. In CAT, γ is set as 0.2.

Since this dataset is imbalanced, we measure the performance in terms of the Area Under Curve (AUC) score (see Table 1). We compare against variants of our proposed method as well as the state-of-the-art sequence classification method Temporal Attention-Gated Model (TAGM) [36]. There also exist other works in e-Commerce that classify sequential data using similar techniques to the LSTM-Attention model [33]. The method LSTM+AT corresponds to the method in the prior work [30]. As we aim to improve the sequential classification performance through data augmentation by AT, the works [7, 8, 19, 37] on designing or defending against specific attacks are out of our scope. We run each method five times with random initialization and report the mean and standard deviation.

According to Table 1, the proposed method outperforms baselines by a considerable margin. Besides, we have several observations. First, the attention mechanism is helpful compared to the conventional LSTM model since users randomly view many web pages. Second, the comparisons between attention-aware AT and LSTM-Attention model and between LSTM+AT and LSTM demonstrate that the adversarial training can indeed improve the learning performance. Finally, we show that using the VAT-based approach on the unlabeled data can also boost the performance.

Table 1: Performance on the Digital Marketing Dataset.

Method	AUC
LSTM	0.6964±0.0022
LSTM+AT [30]	0.7055±0.0018
LSTM-Attention	0.7079±0.0018
Attention-aware AT	0.7112±0.0018
Attention-aware AT (SF)	0.7133±0.0013
Attention-aware AT (CAT)	**0.7142**±0.0014
Attention-aware AT+VAT	0.7163±0.0016
Attention-aware AT+VAT (SF)	0.7205±0.0013
Attention-aware AT+VAT (CAT)	**0.7218**±0.0010
TAGM [36]	0.7054±0.0017

To demonstrate how the attention-aware adversarial training helps with data augmentation, we show an illustrative example in Table 2 for a sequence of browsing behaviors for the Photoshop product that finally lead to a conversion. Here the attention model detects that the user is attracted by several tutorial web pages and a download-survey page which are highly relevant to his purchase behavior. We observe that the many tutorial pages (*e.g.*, 7^{th} and 9^{th} behaviors) are highly relevant to professional photography (*PHO*). Hence, it is very likely that the user bought this product for photographic editing. The connections between these detected web pages and the product conversion are confirmed by domain experts who developed this product.

Since we have a relatively small amount of data with purchase behaviors, it would be very helpful to include more data with similar web browsing patterns. In the right column of Table 2, we can see that the attention-aware AT method can automatically replace these most relevant tutorial web pages by other tutorial pages. We show that the original and replaced tutorial pages in the 7^{th} and 9^{th} behaviors are both highly relevant to photography (*PHO*). The original tutorial page in the 4^{th} behavior is about a general operation. Therefore, it is reasonable to replace it by the homepage of tutorials where all the basic operations are listed. The similarity between original tutorial pages and replaced tutorial pages are also confirmed by our collaborators in Adobe. We can conclude that the perturbed behaviors maintain the same browsing patterns with the original sequence while also adding more variability.

Table 2: An example in the Digital Marketing Dataset. The left side shows the original sequence and the right side is the perturbed sequence by Attention-aware AT. The bold items represent perturbed items (with high attention weights) in the sequence. "*GEN*" and "*PHO*" represent general image editing operations and photography-related operations.

ID	original sequence		Perturbed sequence
1	how-to:sharpen-photos		how-to:sharpen-photos
2	downloadsurvey:PS		downloadsurvey:PS
3	using:sharpness-blur		using:sharpness-blur
4	**how-to:remove-obj (*GEN*)**	→	**PS:tutorials**
5	how-to:obj-content-aware		how-to:obj-content-aware
6	**downloadsurvey:PS**	→	**forums:PS**
7	**using:dodge-burn (*PHO*)**	→	**using:adjust-fill-layers (*PHO*)**
8	using:layers		using:layers
9	**how-to:dodge-burn (*PHO*)**	→	**how-to:sharpen-photos (*PHO*)**

We show the sensitivity test for the hyper-parameters in SF and CAT in Fig. 5 (a) and (b). When we use no adversarial samples, the SF approach degenerates into the conventional attention model. In contrast, when using all the adversarial samples, the SF is equivalent to the attention-aware AT method. As we use more and more adversarial samples, the performance will first increase as they can help augment training data. However, the performance starts to decrease after we include sufficient adversarial samples (>80% in this task) because additional adversarial samples can potentially break the perceptual similarity. For the CAT approach, a larger γ can produce adversarial samples that better preserve perceptual similarity, but cannot fully exploit the nearby region in the feature space to find where the classifier can potentially fail. Therefore, when the value of γ is very large, the generated adversarial samples cannot effectively augment training data.

Table 3: Performance on the Movie Review Dataset.

Method	Accuracy
LSTM	0.7221±0.0037
LSTM+AT [30]	0.7422±0.0027
LSTM-Attention	0.7566±0.0054
Attention-aware AT	0.7727±0.0016
Attention-aware AT (SF)	**0.7819**±0.0016
Attention-aware AT (CAT)	0.7803±0.0015
Attention-aware AT+VAT	0.7837±0.0014
Attention-aware AT+VAT (SF)	**0.7901**±0.0014
Attention-aware AT+VAT (CAT)	0.7878±0.0012
NBSVM [44]	0.7643±0.0003
CNN [18]	0.7795±0.0016
TAGM [36]	0.7592±0.0058

3.3 Evaluation on the Movie Review Dataset

Since movie reviews are discrete word inputs, we use an extra embedding layer to transform them to continuous embeddings of 128 dimensions. We pretrain the word-embedding layer and LSTM layer according to previous works [6, 30] using a sequence-to-sequence autoencoder [6]. The pretraining is conducted using a larger unlabeled Amazon Reviews dataset [28]. We set the dimension of hidden representation to be 140. The λ in Eq. 11 is set to 1.0. In SF, we use 60% adversarial samples in each iteration. In CAT, γ is set as 0.2.

The adversarial training-based methods are then applied to the continuous word embeddings. Table 3 shows the accuracy of the proposed method and baselines. The baselines include both the variants of the proposed method and the state-of-the-art sentiment classification algorithms, such as NVSVM [44], CNN [18] and TAGM [36]. It is also noteworthy that the method LSTM+AT corresponds to the method in prior work [30].

We can observe that our proposed method brings 8.3% and 9.4% improvements over the baseline LSTM in the supervised task and the semi-supervised task, respectively. It can be seen that CNN outperforms the attention-aware AT because the CNN model utilizes the pretrained embeddings from Word2Vec Google News. However, after we further refine the augmentation strategy with SF or CAT, the data augmentation becomes much more effective and consequently the performance is comparable to or even better than the CNN baseline which utilizes auxiliary knowledge.

To show how conventional adversarial training can result in significant changes of discriminative period, we give a positive review example in Fig. 3 with the original sentence and the perturbed sentence by LSTM+AT. Here we map embeddings back to words by finding the closest words in the embedding space. After the perturbation, a word "A" is replaced by "B" if the perturbed embedding becomes closer to word "B" and more distant from the original word "A" after the perturbations. We demonstrate the success of the proposed method in detecting the most important words since the words "ambitious", "features" and 'imaginative" indicate positive sentiment. However, after conventional adversarial perturbations, the word "ambitious" is replaced by "open" which does

(a) *Original:* Far more imaginative and ambitious than the trivial cash in features Nickelodeon has made from its other animated TV series.

(b) *Perturbed:* Far more imaginative and *open* than the *playfully* cash *for* *subtlest* Nickelodeon has made from its other animated TV series.

Figure 3: An adversarial sample generated by LSTM+AT in Movie Review Dataset: (a) the original sentence and (b) the perturbed sentence. The italic and underscored words are perturbed words. We mark the top-5 important words detected by the attention model in red color. The darker red color indicates higher attention weight.

Table 4: Performance on the Remote Sensing Dataset.

Method	Accuracy
LSTM	0.8469±0.0083
LSTM+AT [30]	0.8563±0.0075
Attention	0.8566±0.0111
Attention-aware AT	0.8626±0.0022
Attention-aware AT (SF)	**0.8721**±0.0024
Attention-aware AT (CAT)	0.8692±0.0024
Attention-aware AT+VAT	0.8720±0.0011
Attention-aware AT+VAT (SF)	**0.8876**±0.0013
Attention-aware AT+VAT (CAT)	0.8825±0.0011
Random Forest [20]	0.8398±0.0126
SeqRep [32]	0.8505±0.0124
Seq-Recurrent Encoders [40]	0.8618±0.0124
TAGM [36]	0.8623±0.0173

not convey as much positive sentiment as "ambitious". Hence, the attention weight of this position gets reduced after the perturbation. In contrast, the attention weights for non-discriminative positions, *e.g.*, the word "animated", get increased. In this way, the perturbed sentence cannot reflect the correct sentiment by focusing on other non-informative words. Hence, this adversarial sample cannot be used for augmenting training data. Since this example results in a large difference in attention weights after perturbations, it can be filtered by SF or fixed by CAT.

The sensitivity tests in Fig. 5 (c) and (d) shows the similar patterns with the performance on the Digital Marketing Dataset.

3.4 Evaluation on the Remote Sensing Dataset

When we implement the proposed method on this dataset, we set the dimension of hidden representation to be 40. The λ in Eq. 11 is set to 1.0. In SF, we use 40% adversarial samples in each iteration. In CAT, γ is set as 0.6.

We compare our method with baselines in Table 4. The baselines include variants of our method and the state-of-the-art methods for classifying land covers using remote sensing data, such as SeqRep [32], Seq-Recurrent Encoders [40], and TAGM [36]. Since this work focuses on handling sequential data, we do not compare to many existing methods using spatial information. It is noteworthy that the baseline Seq-Recurrent Autoencoder itself also includes convolutional layers over space. Besides, we compare with the Random Forest classifier, which is the most popular method in the domain of remote sensing. Standard LSTM is also widely used in recent remote sensing research [13, 39].

It can be seen that our model outperforms all the baselines. Besides, the attention mechanism performs better than the conventional LSTM model since land covers show their most discriminative patterns only in a short period while the non-discriminative period can introduce much noise. For example, a robust model for crop detection should focus on the period after crops are planted and before they are harvested. The time steps before seeding can adversely impact the classification, since some farmers plant different crop types across years and the residues left on the ground before seeding can belong to a different crop type. After the augmentation by attention-aware AT, the learning performance is even higher than all the state-of-the-art methods in remote sensing.

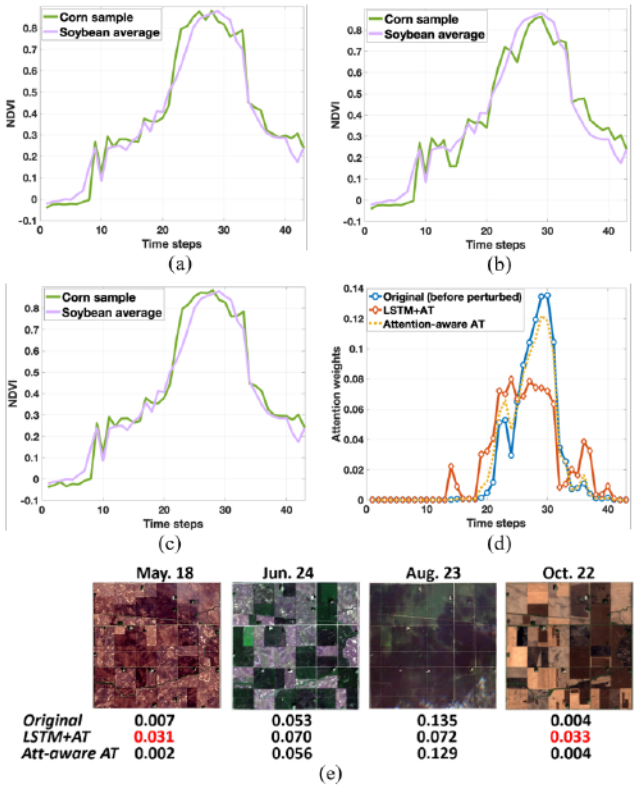

Figure 4: The NDVI series obtained from (a) the original corn sample, (b) the perturbed corn sample using LSTM+AT, and (c) the perturbed corn sample using attention-aware AT. Pink curves are the average of soybean NDVI for comparison. Fig. (d) shows the attention weights obtained from the original corn sample, and the perturbed sample using different methods. (e) illustrative Sentinel images corresponding to important stages: *May. 18* - before seeding, *Jun. 24* - most corns turns green, *Aug. 23* - all crops grow up, and *Oct. 22* - after harvesting. Below is the attention weights obtained from MODIS sequential data in 2016. Red color denotes spuriously high attention weights.

In remote sensing, the Normalized Difference Vegetation Index (NDVI) [2] is widely used as a vegetation/greenness level index

to study vegetation phenology. The NDVI of crop samples will increase rapidly after they are planted, and then decrease after they are harvested. In Fig. 4 (a)-(c), we show the NDVI series of a corn sample to illustrate how adversarial training can cause the variation of attention weights. Besides the corn sample, we also provide the average NDVI of soybeans for comparison (pink curve). In Fig. 4 (d), we show the attention weights obtained from different models. To better visualize important stages in crop growing process, we show Sentinel images (in higher resolution than MODIS) corresponding to these stages in Fig. 4 (e). According to Fig. 4 (d), the attention weight of the original sample has two peaks at the 22^{nd} and the 30^{th} time steps (see Jun. 24 and Aug. 23 in Fig. 4 (e), respectively). On Jun. 24, it can be seen that corn turns into green while other crops remain white in the RGB image. On Aug. 23, even if the RGB images show less difference between corn and soybean, the time steps around Aug. 23 are most important for classification (using the full MODIS spectrum) since the model can better capture the characteristics of corn and soybean after they have fully grown up.

Compared to the original NDVI series (Fig. 4 (a)), in Fig. 4 (b) the corn sample has lower NDVI than soybeans during 24^{th}-28^{th} time steps after perturbations by LSTM+AT. This violates the phenology that corns should turn into green faster than soybean at this early growing stage. Moreover, according to Fig. 4 (d) and (e), the perturbations by LSTM+AT result in spuriously high attention weights before seeding and after harvest (see attention weights around May. 18 and Oct. 22 in Fig. 4 (e)). This is because LSTM+AT cannot perturb data properly at the discriminative time steps. In contrast, attention-aware AT preserves the most important stages according to the NDVI series (Fig. 4 (c)) and attention weights (Fig. 4 (d) and (e)).

In Fig. 5 (e) and (f), we show the results of sensitivity test for this learning task. Compared with previous tasks, the discriminative period in this dataset is more concentrated in a specific short period and the perturbations are more likely to impact the perceptual similarity. Therefore, the optimal performance is achieved using smaller amount of adversarial samples or a larger value of γ.

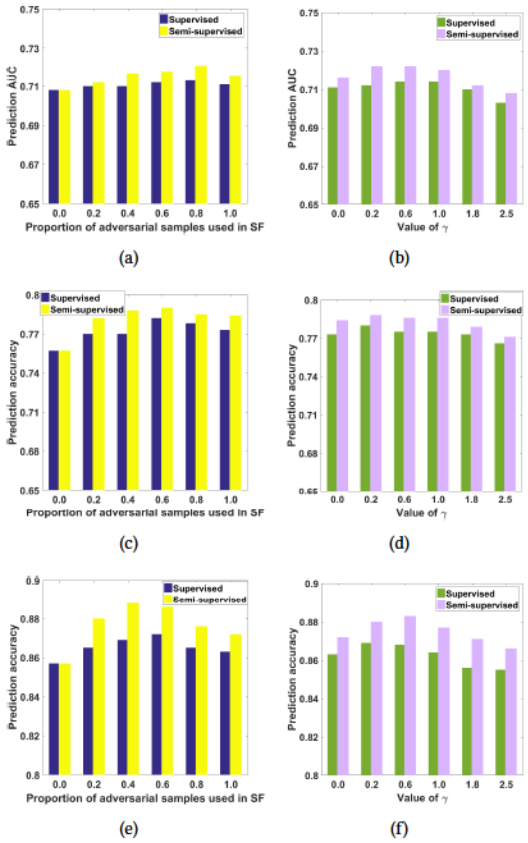

Figure 5: Sensitivity tests with respect to the proportion of data used in SF and the hyperparameter γ in the Digital Marketing Dataset ((a) and (b)), the Movie Review Dataset ((c) and (d)), and the Remote Sensing Dataset ((e) and (f)).

4 RELATED WORK

In past decades, many researchers have shifted their interests from static data to sequential data due to the huge potential therein for discovering knowledge of dynamic patterns. Given that real-world sequential data are collected over long period of time, the attention mechanism is proposed to detect the discriminative data portion [12, 42]. This method has shown to be effective in machine translation [25], image/video captioning [46], question answering [24], and health-care analysis [5, 26, 47].

Recent works have shown that complex deep learning models can produce unreliable predictions with adversarial perturbations [7, 11, 27]. The susceptibility to these perturbations lies in that the learned unsmooth decision boundary can cause sudden changes for model outputs. A variety of defense strategies have been proposed to improve the model robustness against these perturbations, such as AT [11], VAT [31], and other gradient-based approaches [14, 38]. Some defense strategies provide a certified defense or detect adversarial samples given specially designed model structures [19, 37] or specific data structures [10]. These techniques have also shown much success in a variety of real-world applications, including spammer filtering [23], malware detection [43] and medical screening [41].

In this paper, we aim to improve the overall learning performance by developing AT-based approaches, which have the potential to learn a better decision boundary through data augmentation in anisotropic direction [11, 31]. This differs from some aforementioned works in designing or defending against specific adversarial attacks.

Since most of these methods do not take into account the structure of sequential data, they cannot be directly used to improve the robustness of sequential model. Some prior works have explored using AT-based approaches in text mining tasks by directly applying AT on the entire sequence [30], or using an extra mean-pooling layer [35]. Researchers have also developed algorithms to better interpret adversarial text samples through word swapping [8] and erasing [22]. However, these methods cannot capture critical positions and thus the perturbed sequence may not maintain the discriminative patterns of the original sequence. Therefore, these methods cannot fully exploit the nearby feature space without understanding when and how to perturb the sequence.

5 ACKNOWLEDGEMENT

This work was funded by the NSF award 1838159.

6 CONCLUSION

In this paper, we propose a novel attention-aware AT method for sequential data by answering the question "when and how to perform adversarial training?". This method aims to improve the robustness against any perturbations within the discrimiantive periods. Based on this learning method, we further propose SF and CAT to maintain the perceptual similarity of sequential data. Extensive results on real-world datasets demonstrate the effectiveness of our method in improving the classification performance. We give several examples to show how adversarial perturbations can adversely affect the discriminative structure of sequence which supports the intuition of SF and CAT. In summary, we believe this work can provide useful insights for learning robust sequential models and also we anticipate this to be an important stepping stone towards extending adversarial training for different types of structured data.

REFERENCES

[1] 2018. MODIS Product Table. https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod09a1.

[2] 2018. MODIS Vegetation Index Products (NDVI and EVI). https://modis.gsfc.nasa.gov/data/dataprod/mod13.php.

[3] 2018. NASS CDL Program. https://nassgeodata.gmu.edu/CropScape/.

[4] Charles Chen, Sungchul Kim, Hung Bui, Ryan Rossi, Branislav Kveton, Eunyee Koh, and Razvan Bunescu. 2018. Predictive Analysis by Leveraging Temporal User Behavior and User Embeddings. In *Proceedings of the 2018 CIKM*. ACM.

[5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD*. ACM.

[6] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *NIPS*. 3079–3087.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Xiaolin Hu, Jianguo Li, and Jun Zhu. 2017. Boosting adversarial attacks with momentum. *arXiv preprint arXiv:1710.06081* (2017).

[8] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. HotFlip: White-Box Adversarial Examples for NLP. *arXiv preprint arXiv:1712.06751* (2017).

[9] John Cristian Borges Gamboa. 2017. Deep learning for time-series analysis. *arXiv preprint arXiv:1701.01887* (2017).

[10] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. 2018. Adversarial spheres. *arXiv preprint arXiv:1801.02774* (2018).

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints* (Dec. 2014). arXiv:stat.ML/1412.6572

[12] Dichao Hu. 2018. An Introductory Survey on Attention Mechanisms in NLP Problems. *arXiv preprint arXiv:1811.05544* (2018).

[13] Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel. 2017. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters* 14, 10 (2017), 1685–1689.

[14] Daniel Jakubovitz and Raja Giryes. 2018. Improving DNN Robustness to Adversarial Attacks using Jacobian Regularization. *arXiv preprint arXiv:1803.08680* (2018).

[15] Xiaowei Jia, Ankush Khandelwal, Guruprasad Nayak, James Gerber, Kimberly Carlson, Paul West, and Vipin Kumar. 2017. Incremental dual-memory lstm in land cover prediction. In *SIGKDD*. ACM.

[16] Xiaowei Jia, Sheng Li, Ankush Khandelwal, Guruprasad Nayak, Anuj Karpatne, and Vipin Kumar. 2019. Spatial Context-Aware Networks for Mining Temporal Discriminative Period in Land Cover Detection. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 513–521.

[17] Donghyun Kim, Sungchul Kim, Handong Zhao, Sheng Li, Ryan A Rossi, and Eunyee Koh. 2019. Domain Switch-Aware Holistic Recurrent Neural Network for Modeling Multi-Domain User Behavior. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 663–671.

[18] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[19] J Zico Kolter and Eric Wong. 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851* (2017).

[20] Arun D Kulkarni and Barrett Lowe. 2016. Random forest algorithm for land cover classification. (2016).

[21] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. 2017. A context-aware attention network for interactive question answering. In *SIGKDD*. ACM.

[22] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220* (2016).

[23] Ninghao Liu, Hongxia Yang, and Xia Hu. 2018. Adversarial Detection with Model Interpretation. In *Conference on Knowledge Discovery & Data Mining*.

[24] Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. In *Proceedings of the 24th ACM SIGKDD*. ACM.

[25] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[26] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD*. ACM.

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[28] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Recsys*. ACM.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781* (2013).

[30] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint:1605.07725* (2016).

[31] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[32] Guruprasad Nayak, Varun Mithal, Xiaowei Jia, and Vipin Kumar. 2018. Classifying multivariate time series by learning sequence-level discriminative patterns. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM.

[33] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. 2018. Perceive Your Users in Depth: Learning Universal User Representations from Multiple E-commerce Tasks. In *Proceedings of the 24th ACM SIGKDD*. ACM.

[34] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

[35] Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *Military Communications Conference, MILCOM 2016-2016 IEEE*. IEEE, 49–54.

[36] Wenjie Pei, Tadas Baltrušaitis, David MJ Tax, and Louis-Philippe Morency. 2017. Temporal attention-gated model for robust sequence classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 820–829.

[37] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344* (2018).

[38] Andrew Slavin Ross and Finale Doshi-Velez. 2017. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404* (2017).

[39] Marc RuBwurm and Marco Körner. 2017. Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images.. In *CVPR Workshops*. 1496–1504.

[40] Marc Rußwurm and Marco Körner. 2018. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information* (2018).

[41] Mengying Sun, Fengyi Tang, Jinfeng Yi, Fei Wang, and Jiayu Zhou. 2018. Identify Susceptible Locations in Medical Records via Adversarial Attacks on Deep Predictive Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[42] Feng Wang and David MJ Tax. 2016. Survey on the attention based RNN model and its applications in computer vision. *arXiv preprint arXiv:1601.06823* (2016).

[43] Qinglong Wang, Wenbo Guo, Kaixuan Zhang, Alexander G Ororbia II, Xinyu Xing, Xue Liu, and C Lee Giles. 2017. Adversary resistant deep neural networks with an application to malware detection. In *Proceedings of the 23rd ACM SIGKDD*.

[44] Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*.

[45] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 NAACL*.

[46] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.

[47] Yuan Zhang, Xi Yang, Julie Ivy, and Min Chi. 2019. ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling. In *IJCAI 2019*. International Joint Conferences on Artificial Intelligence.

[48] Handong Zhao, Quanfu Fan, Dan Gutfreund, and Yun Fu. 2018. Semantically Guided Visual Question Answering. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1852–1860.

[49] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial attacks on neural networks for graph data. In *SIGKDD*. ACM.