Hello, my topic is time series adversarial attack and defense.

I divide my presentation into 4 parts. I would first introduce my **motivation** of researching time series adversarial attack. Then I will summarize some attacks and defenses. In the end, I will **conclude my research**.

To make adversarial attack more **understandable**, I would give an example of **self-driving**. Self-driving is a hot topic in recent years. During self-driving, **predicting the future trajectory** of other vehicle is a big task, which will affect the **driving decision** of **Autonomous Vehicles**. The history trajectory is actually a combination of time points and locations. This is time series, a series of data points indexed in time order. To Process such a huge volume of data, researchers use deep learning models, whose performance grows by the amount of data. In this case, researchers let deep learning models learn the history trajectory of other vehicle to predict its future trajectory.

However, researchers find out that a tiny perturbation of the input may lead to wrong output. If the history trajectory of other vehicle is crafted, deep learning model will predict it wrongly. In the picture, we can see that deep learning model predicts the other vehicle will change the lane, thus autonomous vehicle takes a brake. However, the other vehicle won’t change its lane. Taking a brake is a false decision. The crafted history, in other word, the perturbation intentionally generated by human, is called adversarial attack. This may lead to serious consequence to human life. Thus, researching on adversarial attack and its defense is essential.

Before introducing adversarial attack method. I would like to explain an important formula: Xquotation equals to X plus eta. X is the original time series, eta is perturbation and Xquotaion is the adversarial samples. There are mainly 2 types of adversarial attack methods. In White-box attacks, the attacker has access to all the information about the targeted model. Attacking with gradient is a common way. In Black-box attacks, the attacker has no information about the target model.

One common method in white-box attack is Fast gradient sign method. The perturbation is generated by a one-step gradient update. Let’s see the formular.

Epsilon is the magnitude of the perturbation. L is the loss function and its meaning is to evaluate the degree of prediction error. Thus, this sign function can decide in which direction the perturbation will lead to the maximum error. The result of this eta is positive epsilon or negative epsilon.

Basic iterative method is based on Fast gradient sign method. The perturbation will be iterated in smaller step sizes. Small perturbation, maximum perturbation and number of iterations will be set here. Let’s see the picture. The number of iterations is the number of the steps the person can go. During each step, the person will decide which direction to step on. This is smaller FGSM in each iteration. And the person can’t step over the maximum step length epsilon. Finally, the person will find out which is the smallest total steps. This is the result of Basic iterative method and it will be smaller or on worst equal epsilon.

Compare to the iteration in BIM, FGSM needs only one step, thus its generating time is much shorter. However, the effect of adversarial samples generated by BIM is much more effective than FGSM.

One easy method in black-box attack is Random noise attack. Researchers generate random noise data and add them to original data. Even a tiny perturbation can misguide the decision of the model. In the picture, the original example was classified in class 0. When the noise was added, it is now classified in class 2 with high confidence.

Another method is Boundary attack. The goal of it is to get the whole decision boundary of the label of the original data. There’s an original data which is classified in class A and we choose a starting data in class B. The starting data then gets close to original data. When it is classified to B, then change its direction to orthogonal direction and go for a short step. After that go towards original data again. After this iteration, we will get the whole decision boundary and can get the smallest effective perturbation according to the boundary.

Four attack methods have one thing in common: They all modify the whole original time seires. However, it’s not necessary. Researchers find out the some parts of one time series will affect more in prediction.

This is called susceptible region. Class activation map is a method to find it. We use the formular to evaluate how susceptible the region is. Am(t) is the multivariate time series with m variables and w is the weight between label c and variable m. We can see from the picture that red parts are most susceptible.

We also can evaluate the importance of adversarial sample. Firstly, we will generate adversarial samples. Then, calculate the distance between prediction yi and yi-quotation. The bigger the distance, the more important the adversarial sample is. After that, rank them in descending order and determine a proportion P. Finally, replace P percent samples to original data.

The goal of researching attack is to find the better way to defend them. In this section, I will introduce 3 kinds of defenses.

Adversarial training is one of the most widely used defense. The model will be trained by more adversarial samples than original data. Its concept is to find the parameter vector with high sensitivity and then minimize it.

In Feature denoising, the researchers will add denoising layers before max-pooling layers. This layer can be trained and the main function of it is completed by a denoising operator. Researchers often use Gaussian non-local means as denoising operator.

The third method is non-linear transfer. It consists of two parts: Thermometer coding and encode-decode model. Thermometer coding is a variant of one-hot coding and it can transfer continuous input to discrete input. Compare to one-hot coding, it can keep the shape information of the original time series data, which is essential in recovering original data.

Encode-decode model is a model to recover the data, which is encoded by thermometer coding. We can see from the picture that the data after recovery has almost the same shape as original data. Then a time-series model will be trained by recovered data. In this method, people don’t need to concern the perturbation on original data, because all the perturbation will be filtered by thermometer coding.

I have already summarized various adversarial attacks and different kinds of defenses. With the improvement of robustness, normal adversarial samples can’t now misguide models. Thus, I think researchers can pay more attention to optimizing adversarial attacks. For example, using CAM to get the effective perturbation wth only a little change on original data.