Graph Neural Networks for Deepfake Video Detection

Zane Peycke, Zhongtian Pan
ZMP2105, ZP2217
EECS E6895
Advanced Big Data Analytics
Milestone 1 Report
Columbia University

February 14, 2020

Abstract

As part of an increasing amount of false information, videos altered or generated with artificial intelligence techniques have reached a level of realism that makes it difficult for human viewers to discern whether a video is real or fake. These videos, called deepfakes because of the deep learning methods often used to create them, pose a significant problem for society, and new tools are needed to detect the authenticity of media. Efficient deepfake detection is a difficult problem because of both the computational complexity and questions surrounding what constitutes a real video. We propose the implementation of graph neural networks to build an accurate deepfake detection model. Graph neural networks are well suited to this problem because of their ability to detect hidden patterns in non-euclidean space, and the spatial and temporal relationships of deepfake alterations. We are not aware of any existing deepfake detection techniques that utilize graph neural networks.

Related Work

After the success of deep learning models on euclidian data, including images and text, many deep learning researchers continued working with datasets that have more complex representations. Mapping interactions between users of a social media platform, recommendation-based systems, and even life-science applications like atomic interactions and protein mapping have structures that are often not well represented in euclidian space[1]. Early graph neural networks (GNN) saught to represent these relationships by learning a target node's representation by iteratively propagating neighbor information.[2] For computer vision applications, graph neural networks have been developed with both spectral[3] and spatial [4] approaches. Applying GNNs for computer vision is especially interesting because of the success achieved by convolutional neural networks, and the opportunity to define videos as graphs with nodes that change over time. Spatial and temporal graph representation has been developed for usage in traffic vision applications[5] and human action recognition[6]. We believe that the spatial-temporal graph neural network is also well suited to deepfake detection.

Project Introduction

When a deepfake video constructed in a way that leads humans to believe it is authentic, there is a path for people to spread false information, and it is easy to see why this has many negative consequences. However, deepfake technology also creates an opportunity for people to claim that real videos are actually fake. Highly believable deepfake videos create multiple paths for people to manipulate viewers, and new tools are required before deepfakes becomes more prevalent. We chose to work on this problem because we believe that it is crucial to deploy efficient deepfake detection as soon as possible. Although we are unaware of any GNN based deepfake detection tool, the previous work on computer vision applications suggests that GNNs are well suited for this problem. We plan to construct a tool based on GNNs that can reliably detect deepfake alterations. Specifically, we plan to utilize spatial-temporal graph representation to train our model and eventually identify the hidden patterns present in deepfakes. The application of spatial-temporal graph neural networks (STGNN) has been successfully applied in the previously mentioned traffic forecasting application. We plan to investigate the successful application of videos as spatial-temporal graphs and research the creation of deepfake tools. There are also several promising graph implementations that utilize other popular machine learning techniques, including convolutional layers and generative adversarial networks (GANs). Many deepfake creation tools use GANs, so it may be possible to utilize GANs for deepfake detection.

Dataset

For this project, we will primarily be using a training set composed of both real videos and deepfakes. The size of this data set is over 470GB and includes a diverse set of deepfake examples. This dataset was put together by The Partnership on AI Steering Committee on AI and Media Integrity, Amazon, Microsoft, and Facebook for an ongoing deepfake detection challenge [7]. As part of this project, there is also a private test set that can be used to evaluate the accuracy of our method and compare our graph-based approach to other methods. After successfully building our GNN model, we will also be testing on new deepfakes that are not included in this dataset. A variety of simple to use tools exist that allow deepfake creation and as a result, there are a large number of publicly listed deepfake videos on Youtube. These publicly listed videos will be of particular interest because of the variety of deepfake techniques in use and the potential applications of a generalized deepfake detection tool.

Plan

Before the second milestone, the majority of our time will be devoted to two large tasks. The first being related to video processing and environment setup. Although we anticipate this could take a significant amount of time and resources, we believe that we can accomplish all processing and setup tasks relatively quickly because of the large number of open-source implementations related to image and video processing. The challenge of deepfake detection has led to a range of methods and efficient processing techniques on this dataset that are well documented.

The second category encompasses all matters related to the graph-based implementation. We first must complete our research on graph convolutional networks, graph representation learning with generative adversarial nets, and other applications of spatial-temporal graph neural networks. This task requires a significant amount of effort, and the results of this research will dictate the direction of our project. Because we have not decided on a specific implementation, we cannot describe the later steps of this project in technical detail at this time. We have set a schedule for our progress and will have a more detailed plan once our methods are well defined.

At the second milestone, we plan to have a graph neural network system that can make predictions on deepfake videos. We do not anticipate having a highly accurate model at this time, but we will have decided on the specific implementation details and will have developed a workflow for training and testing deepfake videos. Our second milestone presentation will have more technical detail about our implementation.

At the third milestone, we plan to have significantly improved the accuracy of our model. We want this implementation to be at least as accurate as the most popular deepfake tools that have been developed at this time. We also plan to start testing the model on deepfake videos outside our original dataset. This will be a test of how well our model generalizes, and whether this approach can be useful to the public.

For our final milestone, we plan to deploy our model for public use. We envision this application functioning as a web browser extension that could automatically label videos as potentially being altered. We are also open to collaborating with other groups and will continue these discussions as the semester goes on.

References

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[2] Claudio Gallicchio and Alessio Micheli. Graph echo state networks. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[3] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[4] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.

[5] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[7] Join the deepfake detection challenge (https://deepfakedetectionchallenge.ai).