IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Review On AI-Powered Detection Of Deepfake Media With Real-Time Insights

[1]Mrs.U.A.S Gani, [2]Siddhi Chindhalore, [3]Sanskruti Pote, [4]Shreya Walde, [5]Suchita Pawar

[1]Professor, [2,3,4,5] Final Year B.Tech. Student

[1,2,3,4,5] Department of Artificial Intelligence & Data Science, [1]Priyadarshini College of Engineering, Nagpur, Maharashtra, India

Abstract: A number of methods for altering faces in films have been effectively created and made publicly accessible in recent years (e.g., Face Swap, deepfake, etc.). Using these technologies, it is possible to facilitate face video modifications with inaccurate results. It is employable in almost all fields. However, the overemphasis of all technologies is fatal prone to have a certain effect in society which may be negative (e.g., fake news, and cyberbullying revenge porn). It is therefore important to be able to tell whether a person's face in a video has been modified, subjectively. To be able to address the problem of deepfake videos, we focus on the problem of face alteration detection in video sequences. In particular, we focus on the ensembles of several Convolutional Neural Network (CNN) models that have been developed.The proposed methodology attains these objectives through the use of attention layers and data training powerful models derived from a base network, EfficientNetB4. By using two publicly available datasets and combining over 119,000 videos, we show how to be able to address these bezier curves, Detecting face alteration is a crucial field of computer vision remains a challenging task in most scenarios, but we demonstrate that in our case, the combined networks approach highly improves the results.

Index Terms **- Deepfake, Video Forensics, Deep Learning, Attention.**

I. INTRODUCTION

This means that a speaker's identity can be changed with a moderate amount of effort. Digital face editing tools are now easy to use making them accessible to everyone regardless of art or picture retouching experience. Users can now started accessing artificial tools that effectively handle tasks by themselves [4, 5]. New artistic developments help people create better art with their technological tools. Advanced technology enables criminals to produce false videos with relative ease. Face-altering technology poses dangers because attackers can spread fake videos and create illegal revenge pornography. Establishing true identities in video sequences stands as today's major concern because spreading fake content creates serious problems for society [6].

Research around checking if filmmakers change their content has existed for a long time. Experts in multimedia forensics began studying this field long ago with their research about different solution methods. These authors examine film coding details to discover information about movie processing. Research institutions study copy-move detection modifications using dense data blocks. Many experts have created ways to spot when video frames repeat or get removed. All of the above methods rely on the same principle: Each permanent change makes a unique detection mark to help find exactly where the editing took place. The traces forensic scientists look for tend to be hard to see and pick up. Hard-to-detect video edits occur during extreme down sampling or simultaneous complex edits plus strong compression steps [8]. Realistic manipulation techniques create effective obstacles for forensic modelling systems. Current facial transformation techniques prove difficult for forensic experts to identify accurately in modern times [16]. Several different techniques modify face images with no single explanation working for all cases Their technology operates on limited areas within video frames-usually just the face or parts of it. Reference taken from [17] and the Facebook DFDC dataset [18] declare on Kaggle in December 2019 we study how different manipulation tools like deepfakes, Face2Face, Footage Swap and Neural Textures can be identified. We create a new variant of

EfficientNetB4 [19] through our work by adding attention elements from [20]. Researchers find it harder to detect manipulated films because these videos spread on social media platforms apply data compression and coding. We research the challenge of distinguishing face alteration tactics through modern approaches.

Fig. 1. Sample faces extracted from FF++ and DFDC datasets. For each pristine face, we show a corresponding fake sample generated from it.

II. RELATED WORK

In recent years, a number of video forensics methods have been put out for various purposes [7]–[9]. Yet experts have created various ways to spot this kind of fake since the forensics field realized the possible social issues that new face-altering methods could cause [16]. Many of these techniques look at each frame using CNN. One example, Mesonota, is put forward in [21]. This simple CNN aims to find fake faces. The writers in [17] show how Captioned beats this network when retrained on purpose. Other approaches use LSTM analysis to check how video frames change over time. [22] and [23] are examples where a repeating process combines features already picked from frames.

Some techniques take advantage of specific processing traces. The researchers in [24] exploit the idea that deepfake donor faces are warped to match the host film. They suggest a detector that picks up warping traces. Other approaches use frame semantic content analysis to overcome pixel-level analysis limits. [25] offers a method to learn to classify between true and fake head poses. [26] focuses on asymmetrical illumination artifacts instead. [27] describes a system based on eye blinking. Early deepfake movies had many eye artifacts that this approach captures.

As manipulation techniques get better at creating realistic results semantic approaches become less useful. Also several methods provide some localization information. [28] presents a multi-task learning technique that gives a detection score using a segment-station mask. Another way to tackle this would be to use an attention mechanism, as [29] puts forward.

Our work demonstrates two training methods using Siamese network architecture for all the chosen deep models. We build our forensic detection system because real-world implementation remains difficult to execute. Our solution meets DFDC's strict hardware and timeline necessities as documented in [18]. Recent studies introduced FF++ as their attention-based approach the detection system can better explain how frame sections contribute to finding manipulated faces. During the following sections we explain each part of this research. This section reviews the latest published studies related to our research paper.

III. PROPOSED METHOD

This segment introduces the tactic we crafted for identifying if a face in a video shot is authentic ('pristine') or a fabrication. At the heart of our suggested tactic lies the concept of enemy-bling. For quite a while now, folks have realized combining models can lead to better prediction accuracy. With that insight, we are zeroing in on the question of whether we can teach a bunch of CNN models to catch different kinds of high-level semantic details that fill in for each other pretty well. To achieve this, the Efficient Net lineage unveiled in [19] as a bold new strategy to scaling CNNs , is our starting point. This group of architectural surpasses other cutting-edge CNNs in accuracy and efficiency and has been shown to be highly helpful in meeting the time and hardware requirements set by DFDC.

We provide two approaches to make the model useful for the enabling given an Efficient Net design. As an alternative we still propose an introduction of an attention mechanism which would benefit the analyst in observing at what video segment is more informative towards the task of categorization. To gather max details about the data, we must figure out how to add Siamese training ways into the learning method. Below, you'll find more on the Efficient-Net structure, the suggested focus feature, and the way to train the network.

3.1 Effective attention and net mechanism

In one study denoted as [19], this star player got a score of 83.8% for nailing the top spot in identifying pictures on the ImageNet [30] challenge, and it did all this with 19 million bits and pieces and used up 4.2 billion FLOPS. Now, if we take a look at another piece of work tagged as [17], they used the same challenge for a method named Captioned, which managed 79% on hitting the top spot but gulped down twice as many FLOPS at 8.4 billion and had more bits to it with 23 million parameters. If you want to catch a glimpse of what EfficientNetB4's bones look like just peek at the blue area in Figure 2. . There you'll see all its parts laid out with the same names they were given when they first popped up in study [19]. A color picture squared, I, or as we looked at it, the face we pulled out of a video snap, is what the network starts with. , to get the classifying part more on point, the folks in [17] suggest following the face details rather than chucking the entire snap into the network. The various applications of attention mechanisms in computer vision and natural language processing motivated the proposed adaptation of the standard EfficientNetB4 Archi-texture.

However, when the network has more information in making the decision, the detection of the fake part will yet be useful, i.e., it will detect the fake part when the network is given more information in its input to detect.

3.2 Network Drilling

The two models assist in extraction a feature descriptor by emphasizing analogy among samples of the same kind using a generalization power available class via presentation of such generalization potential by the channels.The overall goal is to distinguish samples (rotten faces) of the real and fake classes. The two models we have for training against any of our staff are (i) end-to-end and (ii) Siamese. Other evaluation tactics were also used, such as the DFDC contest methodology.

1) End-to-end training: The network presents us with a face. Once a sample face is entered into the y-related score yˆ, y. Note that no Sigmoid activation function has been applied to this score. Weight updates take place using the well-known log loss formula, which is L = − 1 N[y log (S(yˆ)) + (1 − y) log (1 − S(yˆ))] yi ∈ { 0, 1 } means the corresponding face label in where.

2) Training in Siamese: We train with the loss function triplet margin loss, which was first discussed in[35] and is motivated from computer vision research that operate CNNs to produce local feature descriptors. The non-linear dimension of f (I) racksen coding from an input face the network gives I, as indicated in Figure 2 further means the L2 norm, LT = triplet margin loss reformulated as max (0, mean + δ+ − δ–).

3) sThe losses δ- = f (Ia) – f (In) ¹ 2 and δ+ = f (Ia) −

Strictly positive f (Ip) μ 2ours is now the following Ia, Ip, and In: Ip belongs to the same group of positive samples. as Ia.

Fig.2. Effect of the attention on faces under analysis. Given some faces to analyze (top row), the attention network tends to selec regions like eyes, mouth and nose (bottom row).Faces have been taken out of the FF++ dataset.

IV.EXPERIMENTS

We provide all the information about the experimental setup and datasets utilized in this section.

4.1 Dataset

FF++ [17] and DFDC [18] are the two datasets on which we test the suggested approach. Each technique is used on 1000 high quality prism videos that were manually chosen to show topics that are almost front facing and free of occlusions after being downloaded from YouTube. There are at least 280 frames in each sequence. A constant rate quantization value of 23 and 40, respectively, is used to create high-quality videos. There are at least 280 frames in each sequence. A constant rate quantization value of 23 and 40, respectively, is used to create high-quality videos.

The DFDC is an initial dataset that was made available for the similar Kaggle competition. These particular video clips were created using both real and fake copies of over 19,000 videos. The actors in actual videos are framed against randomly chosen backdrop to create visual diversity and variability in a number of parameters (gender, skin colour, age, etc.). Some of the videos are authentic, while others are erroneous and all are created utilizing Deepfake techniques. We won't be able to determine the precise algorithms that were used to create the fake

videos because the public and private evaluation sequences, as well as an example of how they were prepared, have not yet been made public.

4.2 Networks

We take into account the following networks in our experiments:

• captioned, as it is the model that performed the best in [17], making it the ideal benchmark for our testing campaign; • EffectiveNetB4, which outperforms other current techniques in terms of accuracy and efficiency [19];

• EffectiveNetB4Att, which ought to separate pertinent from irrelevant facial sample components. Every model is independently trained and assessed among the two sets of data being reviewed. For FF++, in particular, we evaluate only films that have been quantized with constant rate of 23. The two Efficient Net models are both trained by the assistant two approaches noted in Section III-B, Capturing is trained with the same style as in}{17. This is for our four trained models are EfficientNetB4ST and EfficientNetB4AttST models trained with the NiceNetB4 and Siamese strategy and Trained with the efficiency net . Conventional from end to end technique. All of these models derived from EfficientNetB4 might help in the final assembly.

V.RESULTS

Here, we have gathered every one of our needs project effort.

5.1 A. Explainability of EfficientNetB4Att:

The output of the Sigmoid layer of the attention block is a 2D map of 28 x 28 with respect to Fig. 9.8 two. We combine this map with the input face artifacted at input face size 224 ×, 224.es, and we show the generated attention map on a few faces of FF++.

Using this straightforward attention-grabbing approach. Inversely, the flat network does not help the network areas having little gradient data. Study after studies have shown that face most of the proprietary traits create the artifacts generated by deepfake generation techniques [16]. Let's combine this concept with blockchain technology. Aside from the main components of these techniques, the major traits sketchy eyes and too many fragilely done teeth white spots.

5.2 Characteristics of Siamese:

To determine whether the features generated by the network's encoding are discriminative for the task, we used the well-known tSNE [39] technique to calculate a projection over a restricted area during the Siamese fashion training of the network. Starting at 20 FF++, Figure 5 displays the projection based on EfficientNetB4Att. Naturally, frames from the same videos cluster into little subregions. Additionally, the chart is set up with the real samples at the top and the phony samples at the bottom. Then, by clustering its frames, the same video is segmented into smaller subregions.

5.3 The Independence of architecture:

For resolution of networks it can be in an ensemble ,where independent models can record the scores. In Figure 6, each plot below the diagonal highlights how several networks offer distinct scores for each frame. In practice, the point clouds are not sufficiently coordinated in a shape that can be represented by a simple join. This urges us use all of the learned models at once.

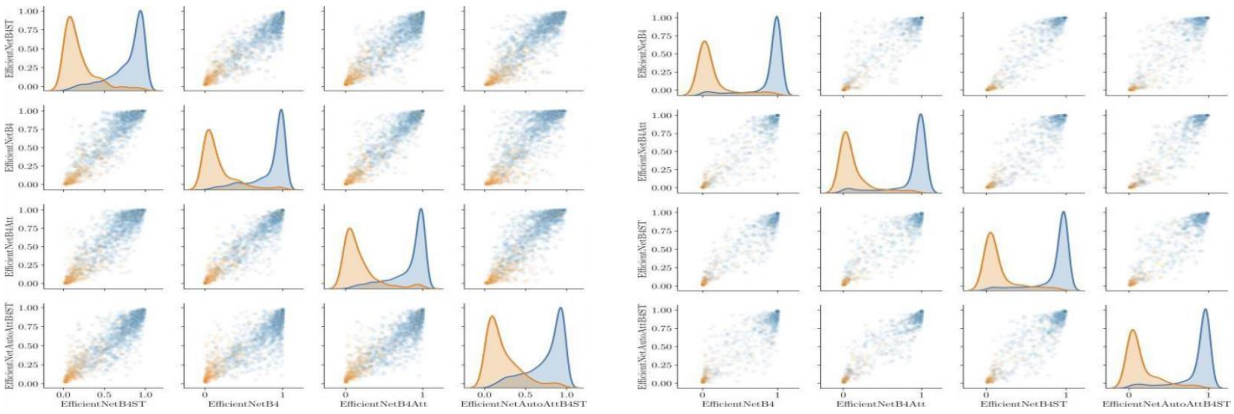
5.4 The Ability to identify face manipulation:

The average results for the baseline network (captioned) and the four proposed models (EfficientNetB4AttST) are shown in this section. There is a theory consisting of an ensemble that has one, two or some even numbers of models. In this particular instance, summing the scores produced by each distinct model yields the greatest score linked to a face. Table I shows the log loss ob.-tainted and AUC (the decisions made by binarizing the network output with various thresholds) for our trials. The out comes are shown for every frame .It is crucial to keep in mind that the model-assembling method usually produces positive outcomes when analysing these findings.

5.5 Results from Kaggle

Team of ISPL, Participated in the DFDC challenge on Kaggle [18] to have a better understanding of the performance of the suggested solution. The final objective of the competition was to develop a system that could distinguish between a real and a false video.

The training dataset made available by the competition host is the DFDC dataset utilized in this work, whereas the two distinct testing datasets are employed for evaluation: (i) the public test.



(a)FF+ (b)DFDC

Fig.3 presents the score distribution (pair-plot) for each pair of networks on the FF++ (a) and DFDC (b) datasets for real (orange •) and fake (blue •) samples.

VI.CONCLUSION

Since video plays such a crucial role in the daily routine and mass communications, it becomes very important to be able to determine if the video contains manipulated content. By using deep learning we move forward to find face manipulation in video sequences in computer graphics and phony films. In this project ,we have used the Efficient Net model and it helps a lot in developing new solutions. This project is used for detecting whether the images are real or fake. This approach also allows deployment of model into streamlit.

Future research will examine how to improve the model selection criteria even further, incorporate multimodal capabilities, and optimize the architecture for conversational tasks that are even more complex. The spread of deepfake photos and videos has serious ramifications for digital media trust, privacy, and national security. Using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs), recent developments in deep learning-based techniques have demonstrated impressive performance in identifying deepfakes. However, the generation of diverse, high-quality datasets and the development of strong detection models that can resist adversarial attacks continue to be pressing research areas. Moreover, real-world deployment and the advancement of a more secure and reliable digital environment depend on explainability, interpretability, and ongoing model changes.

VII.REFERENCES

[1] M. Zoller, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Prez, M. Stamm Inger, M. Niner, and C. Theobald, “State of the art on monocular 3d face reconstruction, tracking, and applications,” *Computer Graphics Forum*, vol. 37, pp. 523–550, 2018.

[2] J. Thies, M. Zoll Hofer, M. Stamm Inger, C. Theobald, and M. Neuner, “Face2face: Real-time face capture and reenactment of grub videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[3] J. Thies, M. Solnhofen, and M. Neuner, “Deferred neural rendering: Image synthesis using neural textures,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[4] “Deepfakes GitHub,
https://github.com/deepfakes/faces
wap.

[5]“Faceswap,” https://github.com/MarekKowalski/FaceSwap/.

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, “Protecting world leaders against deep fakes,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[7] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, “Vision of the unseen: Current trends and challenges in digital image and video forensics,” *ACM Computing Surveys*, vol. 43, no. 26, pp. 1–42, 2011.

[8] S. Milani, M. Fontana, P. Betaine, M. Barni, A. Piva, M. Pagliacci, and S. Tufaro, “An overview on video forensics,” *APSIPA Transactions on Signal and Information Processing*, vol. 1, p. e2, 2012.

[9] M. C. Stamm, Min Wu, and K. J. R. Liu, “Information forensics: An overview of the first decade,” *IEEE Access*, vol. 1, pp. 167–200, 2013.

[10] P. Betaine, S. Milani, M. Tallahatchie, and S. Tufaro, "Codec and gop identification in double compressed videos," *IEEE Transactions on Image Processing (TIP)*, vol. 25, pp. 2298–2310, 2016.
[11] D. Quezada, M. Fontane, D. Shalane, F. Prez-Gonzlez, A. Piva, and M. Barni, "Video integrity verification and gap size estimation via generalized variation of prediction footprint," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 15, pp. 1815–1830,2020.
[12] P. Betaine, S. Milani, M. tagasaste, and S. tuber, "Local tamper- in detection in video sequences," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2013.
[13] L. Damianos, D. Cozzolino, G. Poggi, and L. Verdolaga, "A patch match based dense-field algorithm for video copy move detection and localized in," *IEEE Transactions on Circuits and Systems for Video Technology(TCSVT)*, vol. 29, pp. 669–682, 2019. [14] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 7, pp. 1315–1329, 2012.
[15] A. Gerona, M. Fontanne, T. Bianchi, A. Piva, and M. Barni, "A video forensic technique for detecting frame deletion and insertion," in *2014IEEE International Conference on Acoustics, Speech and Signal Procasing (ICASSP)*, 2014, pp. 6226–6230. [16] L. Verdolaga, "Media forensics and deepfakes: an overview," 2020.
[17] A. Roesler, D. Cozzolino, L. verdolaga, C. Riess, J. Thies, and M. Neuner, "Face Forensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.
[18] "Deepfake Detection Challenge (DFDC),"https://deepfakedetectionchallenge.ai/, 2019.
[19] M. Tan and Q. V. Le, "Efficient net: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning, (ICML) 2019*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114.
[20] A. Vaswani, N. Shabeer, N. Parmar, J. Ozokerite, L. Jones, A. N. Gomez, L. Kaiser, and I. Polishing, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, I. Guyon, U. V. Lux burg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.
[21] D. Achaar, V. Nozick, J. Yamagishi, and I. Echizen, "mesonota: a compact facial video forgery detection network," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.
[22] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition assessment and detection," *Corr*, vol. abs/1812.08685, 2018.
[23] D. Guvera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2019.
[24] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *IEEE Conference on Computer Vision and Pattern Recognitions Workshops (CVPRW)*, 2019.
[25] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *IEEE International Conference on Acoustics, Speech and Processing (ICASSP)*, 2019.
[26] F. Matern, C. Riess, and M. Steiniger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.
[27] Y. Li, M. Chang, and S. Lyu, "In cite oculi: Exposing AI created fake videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.
[28] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," *Corri*, vol. abs/1906.06876, 2019.
[29] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the detection of digital face manipulation," 2019.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
[31] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[32] V. Bassarisks, Y. Kartune, A. Vanunu, K. Raveendran, and M. Grundmann, "Blueface: Sub-millisecond neural face detection on mobile opus," *Corr*, vol. abs/1907.05047, 2019. [Online]. Available :http://arxiv.org/abs/1907.05047
[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2018, pp.7132–7141.
[35] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393. [36] E. K. V. I. I. A. Busload, A. Barinov and A. A. Kalinin, "Augmentations:

fast and flexible image augmentations," *Arrive e-prints*, 2018.

[37] A. Paczki, S . Gross, F . Massa, A . Lerer, J . Bradbury, G. Chanan, T . Killeen, Z . Lin, N . Gimel Shein, L . Antigo, A. Desma Ison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chellamuthu, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Porch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Bergelmir, F. d'Alene´-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019,pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/ 9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[38] D. Kingma and J. Ba, "Adam: a method for stochastic optimization. arrive: 14126980," 2014.

[39] L. van der Maten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9,pp.2579– 2605,2008. [Online].Available:http://www.jmlr.org/papers/v9/vandermaaten08a. html