Multi-modal Adversarial Training for Crisis-related Data Classification on Social Media

Qi Chen, Wei Wang
Department of Computer Science and Software Engineering
Xi'an Jiaotong Liverpool University
Suzhou, China
{qi.chen,wei.wang03}@xjtlu.edu.cn.

Kaizhu Huang
Department of Electrical and Electronics Engineering
Xi'an Jiaotong Liverpool University
Suzhou, China
kaizhu.huang@xjtlu.edu.cn.

Suparna De
Computer Science and Networks Department of Digital Technologies
University of Winchester
Winchester, United Kingdom
Suparna.De@winchester.ac.uk.

Frans Coenen
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
Coenen@liverpool.ac.uk.

***Abstract*—Social media platforms such as Twitter are increasingly used to collect data of all kinds. During natural disasters, users may post text and image data on social media platforms to report information about infrastructure damage, injured people, cautions and warnings. Timely and effective processing and analysing tweets can help city organisations gain situational awareness of the affected citizens and take timely operations. With the advances in deep learning techniques, recent studies have significantly improved the performance in classifying crisis-related tweets. However, deep learning models are vulnerable to adversarial examples, which may be imperceptible to the human, but can lead to model's misclassification. To process multi-modal data as well as improve the robustness of deep learning models, we propose a multi-modal adversarial training method for crisis-related tweets classification in this paper. The evaluation results clearly demonstrate the advantages of the proposed model in improving the robustness of tweet classification.**

***Index Terms*—Adversarial training, Crisis-related data classification, Convolutional neural network, Smart city, Deep learning**

I. Introduction

As smartphones and wireless networks become pervasive, ordinary people actively observe, collect, analyse and report information through social media platforms, e.g., Twitter. As of 2019, there are around 500 million tweets posted on Twitter each day. Such large amount of social media data, covering nearly everything happening around the world, is easily accessible and has become valuable resources for research in data mining and knowledge discovery, e.g., sentiment analysis [1], traffic event detection [2], and crisis-related data classification [3], [4].

At times of natural disasters, social media platforms such as Twitter and Facebook are considered vital information sources that contain a variety of useful information such as reports of injured people, infrastructure damage, missing people, etc. [5]. Processing social media data to extract life-saving information which is helpful for organizations in preparedness, response, and recovery of an emergency [6]. Studies [3], [4] build machine learning models, i.e., Deep Neural Networks (DNNs), for information classification and knowledge discovery.

Deep learning is a popular learning paradigm in the machine learning family. With multiple processing layers, representations of raw data with multiple levels of abstraction can be automatically learned without the need for sophisticated feature engineering and tuning [7]. Studies based on deep models have achieved remarkable results in image classification [8]–[10], natural language processing [11], [12] and speech recognition [13], [14]. Meanwhile, deep learning techniques are also considered as ideal candidates for processing and analysing data with multiple modalities, e.g., the research [15], [16] fuses representations learned from text, visual and audio. However, DNN models are vulnerable to adversarial examples. Small perturbations to the input will cause the DNN models to generate incorrect results [17]. To improve the robustness of DNN models, adversarial examples and adversarial training techniques are widely investigated and have become one of the most influential trends in recent deep learning research.

In this paper, we propose a Multi-modal Adversarial Training (MMAT) method for the crisis-related tweets classification. Different from most of the past studies that focused on textual content only, the proposed method is tailored to learn useful representations from both image and text data simultaneously. In particular, adversarial training is applied to improve the robustness of the neural network against adversarial examples. To our best knowledge, this is the first work to classify multi-modal crisis-related social media data with adversarial training techniques. Experimental results demonstrate the proposed MMAT method is able to attain significant improvement of prediction accuracy.

The rest of the paper is organised as follows. In Section II, we review some of the representative work in crisis-related data classification and adversarial training. In Section III, we describe in detail the design of the multi-modal network and its adversarial training process. In Section IV, we conduct a

number of experiments with the proposed model and compared with baseline models on the four datasets. Finally, in Section V, we conclude the paper and point out some future research directions.

II. Related Work

In this section, we present the related work from two perspectives. The first part demonstrates the representative applications of crisis-related data classification on social media. The second part presents some recent studies about adversarial examples and adversarial training for image and text.

Crisis-related tweets classification: During natural disasters, users may post text and image data on social media platforms to report information about infrastructure damage, injured people, cautions and warnings. For example, Figure 1 shows two crisis-related tweets during California wildfires and Mexico earthquake. If twitter information is processed timely and effectively, it will be very valuable for city organisations to gain situational awareness and take timely operations. The work in [3] and [4] presented applications to identify useful or crisis-related tweets from not related ones during crises. A Convolutional Neural Network (CNN) based framework was adopted to capture the salient n-gram information by convolution and pooling operations. In addition, two domain adaptation techniques that utilise the out-of-event data are applied for better classification results when no labelled data is available in the early hours of a disaster. Both studies confirm that the deep learning based method significantly outperformed the conventional machine learning methods, such as Logistic Regression, Support Vector Machines and Random Forest.

(a) Raining Ash and No Rest: Firefighters Struggle to Contain California Wildfires.

(b) Earthquake leaves hundreds dead, crews combing through rubble in #Mexico.

Fig. 1. Illustration of crisis-related data collected from twitter.

Adversarial Training: Deep Neural Networks (DNNs) have achieved remarkable success in various tasks, e.g., image classification, natural language processing, and speech recognition. However, DNN models are vulnerable to adversarial examples. Small perturbations that are imperceptible to humans will cause the DNN models to generate incorrect results when adding into the input. Such issue poses security concerns regarding the uses of DNN models in security-sensitive applications. To improve the robustness of DNN models, adversarial attack and defence techniques are widely explored recently and have become an influential direction for deep learning research. Various defensive techniques against adversarial examples for deep neural networks in the image domain have been proposed [17], [18]. Recent studies [19], [20] also shown the potential of adversarial attacks and defence in the texts domain. After crafting adversarial examples, adversarial training process that injects such examples into training data could help increase DNN robustness [17]. In this paper, we explore the adversarial training technique for multi-modal data and apply it to the robust classification of crisis-related tweets.

III. Multi-modal Adversarial Training

We first present the architecture for multi-modal neural network (MMN), and its visual-CNN and text-CNN components that extract information from multiple modalities into a multi-modal representation for tweet classification. Then, we explain the adversarial training process, which retrains the MMN with crafted adversarial examples for images and text for more robust crisis-related tweets classification.

A. Multi-modal neural network

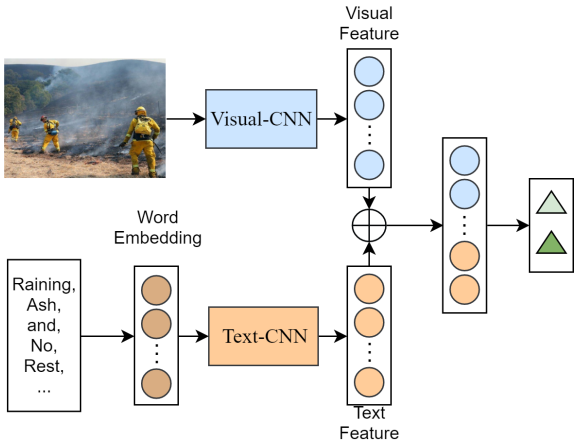


Fig. 2. Multi-modal neural network with both image and text data as input

Majority of past studies only focus on textual content for crisis-related data classification. In this study, we consider data of two different modalities: image and text. Image data is usually represented as a 2 or 3-dimensional matrix with real values, while textual tweets are represented as word sequences. The multi-modal neural network aims to learn feature from both types of data and fuse into a unified representation as shown in Figure 2, where the Visual Convolutional Neural Network (Visual-CNN) is for image input processing and Text Convolutional Neural Network (Text-CNN) is for text input precessing (shown in Figure 3).

Convolutional Neural Network (CNN) has been widely employed in many different types of supervised tasks, e.g. image classification [8]–[10], and natural language processing [11], [12]. The Visual Convolutional Neural Network (Visual-CNN) component aims to extract visual features from images of the tweets. The Visual-CNN consists of two sets of convolutional and max pooling layers, followed by a flatten layer and a fully-connected layer. It should be noted that the performance of visual feature extractor may further be improved with higher

quality image input and pre-trained CNN models, e.g., VGG [8], ResNet [9] and DenseNet [10].

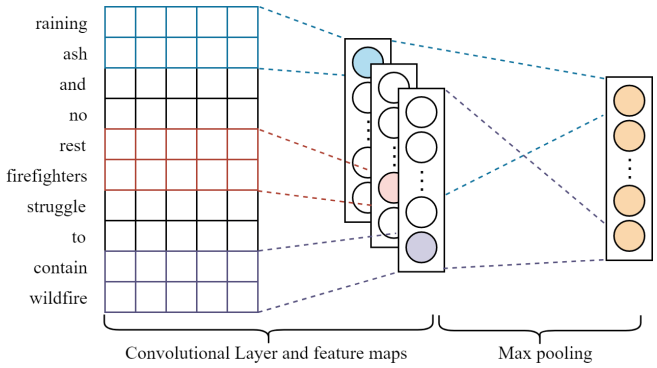


Fig. 3. Text Convolutional Neural Network (Text-CNN) architecture

The Text Convolutional Neural Network (Text-CNN) component attempts to extract an effective representation for the short social media texts. The input to the Text-CNN is a sequence of words, each of which is represented as a word vector. They are initialised with the glove vectors pre-trained on 2 billion twitter posts [21]. The architecture of Text-CNN component is shown in Figure 3. Similar to the work proposed in [11], it consists of a convolutional layer and a max pooling layer. In the convolutional layer, convolutional filters iteratively take the contiguous word vectors from a sequence of words and calculate the output feature map. A max-pooling layer is then used to extract the maximum values of each feature map and form a text representation. Then, the extracted visual and text representations are concatenated into a multi-modal feature representation, which is further used for detecting and classifying crisis-related tweets.

B. Adversarial Training

To improve the robustness of deep neural network against adversarial examples, adversarial training techniques are widely investigated in recent years. By adding small perturbations η to original input x, the adversarial example x' aims to fool the classifier f so that $f(x') \neq y$. Such small perturbations are imperceptible to humans eye but may cause deep learning models' misclassification. A robust deep neural network model should not change its output for these small perturbations in its input. Therefore, the adversarial training process utilise such adversarial examples as an augmented dataset to increase model robustness [17]. Fast Gradient Sign Method (FGSM) [18], shown in Equation 1, is one of the most popular methods used to calculate perturbations and generate adversarial examples.

$$\eta = \epsilon sign(\nabla J(\theta, x, y)) \tag{1}$$

where θ is the parameters of a model, x and y are the input and associated labels, $J(\theta, x, y)$ is the cost of the model. For the multi-modal twitter dataset used in this study, x includes both image input x_i and text input x_t, and thus the associated perturbations are η_i and η_t.

Because the input space of image data is continuous, image adversarial examples can be directly created with $x'_i = x_i + \eta_i$. While the input space of texts is discrete, a perturbed word vector may not represents any word. We cannot directly set the word vector to specific real values in the word embedding space. Therefore, FGSM could not be directly applied to texts to generate adversarial examples. Inspired by work in [19], we craft adversarial tweets by replacing original words with new words that have the largest projection length in the direction of perturbation η_t. To keep semantic similarity, adversarial words are selected from a candidate set C, where only n most similar words in the word embedding space are included for each word in a tweet. The process of word replacement is shown in Equation 2.

$$w' = \underset{w' \in C}{\arg\max}(w' - w) \cdot \eta_t \tag{2}$$

where w' is the adversarial word for w, C is the candidate set, and η_t represents the perturbation for word w calculated by Equation 1. Iteratively, replacement for each word in a sentence can be determined and finally form an adversarial example for tweet that may be misclassified by deep learning models. In addition, as we may not want all word to be replaced in a tweet, a threshold for $(w' - w) \cdot \eta_t$ can be set so as to update only the most impactful words for the model's prediction results.

After generating adversarial examples for image and text, the multi-modal neural network can be retrained with the augmented dataset for better robustness against adversarial examples. The whole multi-modal adversarial training process is shown in III-B.

Algorithm 1 Multi-modal Adversarial Training Algorithm

1: Input: labelled multi-modal input (x_i, x_t, y); MMN classifier f
2: Train MMN classifier f with (x_i, x_t, y)
3: Obtain perturbation (η_i, η_t) for (x_i, x_t) with Equation 1.
4: Create Image adversarial x'_i with $x'_i = x_i + \eta_i$
5: Create Text adversarial x'_t by iteratively update words with Equation 2.
6: Data augmentation with: $(x_i^{aug}, x_t^{aug}, y^{aug}) = data\_aug((x_i, x_t, y), (x'_i, x'_t, y))$
7: retrain MMN classifier f with augmented dataset $(x_i^{aug}, x_t^{aug}, y^{aug})$

IV. Experiments and Evaluation

A. Dataset

We used the multi-modal crisis-related twitter dataset published in [6], in which 16,097 tweets were collected during seven disasters, specifically California wildfires, Hurricane Harvey, Hurricane Irma, Hurricane Maria, Iraq-Iran earthquake, Mexico earthquake, and Sri Lanka floods. We select four datasets for performance evaluation, which are California wildfires, Hurricane Harvey, Mexico earthquake and Sri Lanka floods.

Our goal is to categorise images and text data simultaneously into one of the two classes: (1) **Informative** represents the tweet or image is useful for humanitarian aid. The examples of informative include cautions and advice, infrastructure and utility damage, injured or dead people, affected individuals, missing or found people, etc. The objective is to providing assistance to people who need help in order to save lives, reduce suffering, and rebuild affected communities. (2) **Not informative** represents the tweet or image is not useful for humanitarian aid, e.g., images showing banners, logos, and cartoons. In the original dataset, image and text are labelled as informative and not informative separately for each tweet. We combine them together so as to extract all informative tweets that have either an informative image or informative text.

B. Experiments

Each input sample to the model consists of an image and a textual message (in the form of a sequence of word vectors). As the dimensions of each image could be different, we resized each image into 64×64 and preprocess values into range 0 to 1. For textual messages, we represented each word with a glove word vector of 100 dimensions. Most twitter messages are short in length, so we only considered the first 15 words in each tweet and result in 15×100 dimensions.

We performed a grid search to determine the best parameters for the proposed MMAT. In the Text-CNN, the window size of filter was set to 2, and the dimension of the hidden units in both Visual-CNN and Text-CNN was set to 64. For generating adversarial examples, θ in FGSM is set to 0.1 and the threshold for crafting adversarial text is set to 0.5. The number of batch size was 32; the dropout rate was set to 0.5; the Adam optimiser with early stopping was used to avoid overfitting. These experiments were run using Keras 2.1.5, Tensorflow 1.3, python 3.6, and Windows 10 on a laptop with a i7-6700HQ CPU, 8GB RAM and GTX-970M GPU.

We implemented some baseline models that process either data of single modality or data of multiple modalities for performance comparison.

1) Baseline Models with Data of Single Modality:

- **Visual-CNN** and **Text-CNN** are explained in Section III-A, and shown in Figure 2 and Figure 3. CNN is used separately to extract features from image data and text data as described in Section III-A. A fully connected layer with the hidden size of 64 and a sigmoid output layer were added at the top of these two models to generate the final prediction.
- **Visual-CNN**$_{adv}$ and **Text-CNN**$_{adv}$ are the adversarial training versions of Visual-CNN and Text-CNN. After training Visual-CNN and Text-CNN with the original dataset, adversarial examples for image and text are crafted separately with the equations in 1 and 2. Such adversarial examples are then added into training data to retrain Visual-CNN and Text-CNN and named Visual-CNN$_{adv}$ and Text-CNN$_{adv}$.

2) Baseline Models with Data of Multiple Modalities:

- **Multi-modal neural network (MMN)** uses CNN to extract visual and textual features from image and text data separately, and concatenated the feature vectors for prediction. Detailed architecture of MMN is explained in III-A, and shown in Figure 2. The architecture of MMN has been employed in various different types of supervised applications, e.g., audio-visual speech enhancement [16], fake news detection [15] and Image-sentence ranking [22]. In this study, we used the MMN method as a baseline to evaluate the performance of adversarial training technique for the multi-modal dataset.

C. Evaluation

We assess classification performance with four twitter datasets of different natural disasters, i.e., California wildfire, Harvey Hurricane, Mexico Earthquake and Srilanka floods. The evaluation results of the proposed MMAT and the baseline models are shown in Table I.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Methods	Visual		Text		Multi-modal	
	CNN	CNN$_{adv}$	CNN	CNN$_{adv}$	MMN	MMAT
California Wildfire						
original	86.16	85.53	86.79	87.42	**88.68**	88.05
adversarial	45.91	82.39	76.73	84.28	67.30	**88.05**
Harvey Hurricane						
original	80.67	80.22	88.09	88.09	**89.21**	87.42
adversarial	35.51	78.43	73.26	**86.07**	73.26	84.04
Mexico Earthquake						
original	72.66	75.54	82.73	84.89	84.89	**86.33**
adversarial	30.94	71.94	71.94	78.42	58.71	**79.86**
Srilanka Floods						
original	67.96	66.02	92.23	92.23	**93.20**	**93.20**
adversarial	36.89	66.02	80.85	**89.32**	79.61	**89.32**

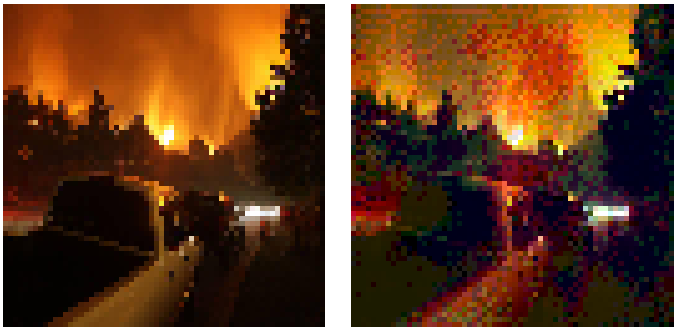
Table I shows that models (i.e. Visual-CNN, and Text-CNN) which process images and text separately can detect and classify tweets successfully with a certain degree. In general, as image data usually contain much redundant information and sometimes not related to the topic, the Visual-CNN model produced the lowest accuracy among all the methods in these four datasets. On the contrary, text data contains more obvious features, e.g. crisis-related keywords, which help extract more informative representations and result in better accuracy. One notable observation from the table is that the overall performance of multi-modal models, i.e., MMN and MMAT, that process multi-modal data simultaneously is better than the models with data of single modality. This observation shows that exploiting complementary data of multiple modalities helps extract more comprehensive knowledge and improve classification performance.

Another notable observation from the table is that the adversarial training technique can significantly improve mod-

els' classification accuracy against adversarial examples range from 6.48% to 42.92%. For text data, the adversarial examples cannot be directly crafted with perturbations calculated from FGSM. Thus, the improvement of text adversarial training is lower than the improvement of image adversarial training. Among the three adversarial models that retrained with adversarial examples, the proposed MMAT outperforms Visual-CNN$_{adv}$ and Text-CNN$_{adv}$ with the highest accuracy in California wildfire dataset 88.05%, Mexico earthquake dataset 79.86%, Srilanka floods dataset 89.32%, and with slightly lower accuracy in Harvey Hurricane dataset 84.04%. This confirms the effectiveness of multi-modal adversarial training that the crafted adversarial examples for image and text could be used simultaneously to help improve the robustness of DNN models with multiple modalities.

D. Case Study

Two tweets extracted from the California wildfire dataset and their adversarial examples are shown as below. MMN model is vulnerable to these multi-modal adversarial examples, and may generate incorrect prediction results.



(a) ICYMI: Why California Wildfires Are Infernos In October.

(b) ICYMI: Why California Wildfires Are Infernos Was October.

Fig. 4. adversarial example of an informative tweet

(a) Fake News: NO Illegal Muslim From Iran Arrested For Starting California Wildfire.

(b) Fake Reports: NO Criminals Christians Of Iraq Arrested For Starting California Wildfire.

Fig. 5. adversarial example of a not informative tweet

Figure 4a shows a tweet that describes the situation of California wildfire and thus is labelled as informative. In Figure 5a, although the tweet has the keyword "California wildfire", it's related to fake news and should be classified as not informative. The MMN could correct classify these two tweets as informative and not informative with 89.1% and 97.2% confidence separately. Adversarial examples of these two tweets, as shown in Figure 4b and Figure 5b, are generated with the equations in 1 and 2. Although the modified images have some visible perturbation noise and the crafted sentence includes some word replacements that may not follow grammar rules, they still have the same semantic meaning and should be classified as the same. However, the MMN misclassified them as not informative and informative with 63.5% and 90.1% confidence separately. These two examples illustrate multi-modal networks are vulnerable to adversarial examples, and thus an adversarial training process is necessary for improving the model's robustness.

V. Conclusion and Future Work

This paper presents an application of crisis-related data classification on social media. To extract valuable knowledge from tweets, a multi-modal network is applied to learn features from both images and text that can complement each other. As the nature of social media, image and text data posted by ordinary people are usually noisy, inconsistent and may not follow grammar rules. We proposed a multi-modal adversarial training framework to improve the robustness of crisis-related tweets classification. The evaluation results clearly showed the advantages of the proposed MMAT over other models with data of single and multiple modalities.

In the current work, we only focus on extracting informative tweets from not informative ones. We plan to extend the current model to support multi-class or multi-label classification, which would provide users with more detailed information, e.g., infrastructure damage, injured people, affected individuals, etc. For textual adversarial examples, generating high-quality sentence that follows grammar rules could be another future research direction. Moreover, we plan to further refine the proposed model and extend to other smart city applications, e.g., transportation, energy, and environmental protection.

Acknowledgment

This research is funded by the Research Development Fund at Xi'an Jiaotong-Liverpool University, contract number RDF-16-01-34.

References

[1] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962, 2015.

[2] Sina Dabiri and Kevin Heaslip. Developing a twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, 118:425–439, 2019.

[3] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902*, 2016.

[4] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[5] J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, et al. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1):1–22, 2015.

[6] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 2017.

[13] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[14] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[15] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857. ACM, 2018.

[16] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

[20] Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. Interpretable adversarial perturbation in input embedding space for text. *arXiv preprint arXiv:1805.02917*, 2018.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.