Adversarial Attacks Against LipNet: End-to-End Sentence Level Lipreading

Mahir Jethanandani and Derek Tang
University of California, Berkeley

***Abstract*—Visual adversarial attacks inspired by Carlini-Wagner targeted audiovisual attacks can fool the state-of-the-art Google DeepMind LipNet model to subtitle anything with over 99% similarity. We explore several methods of visual adversarial attacks, including the vanilla fast gradient sign method (FGSM), the L_∞ iterative fast gradient sign method, and the L_2 modified Carlini-Wagner attacks. The feasibility of these attacks raise privacy and false information threats, as video transcriptions are used to recommend and inform people worldwide and on social media.**

***Index Terms*—Carlini-Wagner attacks, fast gradient sign method, LipNet.**

I. Introduction

THE purely visual sibling to the Carlini-Wagner audiovisual attacks, the attacks explored against LipNet intend to prove the threat of subtitle manipulation and video transcriptions. Using the deep learning subtasks of subtitling (discussed here) and video description (soon to be discussed), social media and advertising companies rely on computer interpretability of exabytes of video data that crosses their platforms every day. Adversarial attacks intended to fool computer interpretability can spread misinformation and target vulnerable audiences. Since Carlini-Wagner audiovisual attacks exploit audio-dependent classifiers, this project focuses on attacks exploiting video-to-text classifiers.

There are numerous examples where fooling state-of-the-art captioning programs reliant on visual observation can go wrong. For example, take a video of a presidential candidate speaking. An opponent can intentionally edit the video to misquote or decontextualize the candidate's original words, and then target voter subpopulations by targeting subtitles and filling them with keywords. In another case, videos commonly muted (as they are by default on Facebook) rely on captioning to attract user attention. Upon automating captioning, minimal perturbations to original content can render dangerous captions. A presidential campaign ad that opens with 'My dear fellow citizens...' can be captioned to anything the attacker chooses.

Lipreading is an essential speech-related deep learning problem, as it aims to gather an understanding of facial movement and its relation to human speech. Human ability to lipread is tough to accomplish, as already observed in psychological study. Its application to bettering the lives of individuals hard of hearing, no matter the scale of hearing loss, became the source of inspiration for the project [9]. As mentioned in the original LipNet paper [1], automating lipreading has applications in improving hearing aids, silent dictation in noisy places, biometric identification, and caption processing. Adversarial attacks and their respective defenses are critical problems to raise as lipreading technology advances, and as adversarial machine learning learns alongside their defenses.

II. Background

A. Related Work

Previous work surrounding adversarial attacks on deep neural networks has largely focused on the image domain. Many different adversarial attack methods have been shown to be tremendously effective on common image prediction datasets such as CIFAR-10, MNIST and ImageNet [6].

Video-specific adversarial attacks include "A Surprising Density of Illusionable Natural Speech" [20]; "Adversarial perturbations against real-time video classification systems" [21]; "Targeted nonlinear adversarial perturbations in images and videos" [22]; "Black-box Adversarial Attacks on Video Recognition Models" [23]; and "Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems" [24]. All pieces were helpful in wrestling the temporal and spectral dynamics of adversarial video attacks, which differ from less complex image adversarial attacks.

As opposed to image data, video inputs include a temporal component in conjunction with spatial components, which increases the complexity of the problem. In this paper, we aim to show that the attack methods that have been shown to be successful in the image domain can be equally effective when applied to videos. However, as far as we know, there are few documented examples of white-box adversarial attacks on end-to-end video data and none relating to the problem of lipreading for captioning purposes.

B. Data

The LipNet model is trained on the standardized GRID corpus dataset, which is the state-of-the-art public dataset for lipreading. The corpus also provides audio data, alongside varying video quality for 34 speakers. Each speaker says every set of predefined sentences. Similar lipreading approaches include audio input, but the LipNet model focuses on the visual-only task. This creates room for future audio- and visual-based adversarial attacks. The GRID corpus follows a highly-specific sentence structure, sticking to similar sentences with 4-6 words. The GRID corpus is available through the University of Sheffield Research Fund [10].

DOI 10.1109/SPW50608.2020.00020

Fig. 1: An example of a LipNet video.

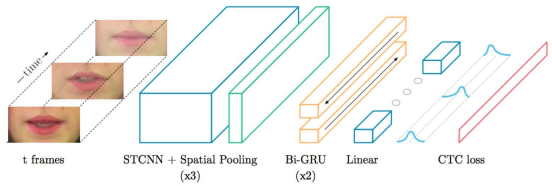

Fig. 2: The LipNet architecture

C. LipNet

The LipNet model receives video of various individuals speaking (Figure 1), and attempts to identify the English phoneme spoken. LipNet presents a uniquely interesting problem to attack as it supports end-to-end prediction, meaning that it analyzes across the entire input video at once instead of going frame-by-frame. The design of the model is as follows: a three second video is encoded a 75 (# of frames) x 100 (width) x 50 (height) x 3 (# of channels) array which is initially passed through a spatiotemporal CNN. The features extracted from the pass are then input to a bidirectional gated recurrent units (GRUs), whose output is fed through a linear layer and a softmax layer. Finally, Connectionist temporal classification (CTC) loss is used to find the most likely alignment from the output of the model [2]. The model complexity allows LipNet to achieve much higher prediction accuracy when compared to other non-end-to-end lipreading models.

The attacks we present focus on the gradient computation of the loss function. The gradient with respect to the loss is taken at this final layer. The fast gradient sign method (FGSM) attacks make special use of the computed gradient.

Third-party implementations of LipNet, unaffiliated with Google DeepMind's approach, are available online. The best implemented model is done in Keras/Tensorflow [11].

D. CTC Loss

Due to the sequential nature of the input data, the choice of loss function must be carefully considered. In order to deal with the fact that the phrases in the videos have no particular temporal alignment and are variable in length, the LipNet model employs the CTC (Connectionist Temporal Classification) loss function which handles both these problems.

For a sequence of frames, the CTC loss procures an alignment between the input and an output by estimating viseme[1] class probabilities at each time-step. A sequence π is generated from taking the probability of each viseme at every frame $f(x_i)^j$ is the probability of viseme j at time frame x_i. Each character in π is either a valid viseme, a space, or a special token ϵ, where ϵ is used to represent a break from one viseme to the next. For a label l, there are many sequences that reduce to l where a reduction involves removing duplicate visemes and special tokens. As an example using the English alphabet, the sequence $h\ h\ e\ \epsilon\ l\ \epsilon\ l\ l\ \epsilon\ o$ reduces to $hello$. A valid alignment for a label l with respect to an output y is a sequence π that is the same length as y and reduces to l. Then, the probability of an alignment π given the softmax output $f(x) = y$ is:

$$Pr(\pi|y) = \prod y^i_{\pi^i}$$

The probability of a label l with respect to an output y is the sum of the probability of alignment π given y across all valid alignments, which can be written as:

$$Pr(l|y) = \sum \prod y^i_{\pi^i}$$

Finally, the negative log-likelihood of the target label is used to train the loss function for the model:

$$CTC - loss(f(x), l) = -logPr(f(x)|l)$$

Since the CTC loss function has a well-defined gradient, it allows us to explore gradient-based adversarial attacks on sequential classification tasks. A more detailed explanation of CTC loss and phrase recovery can be found at [19].

III. Threat Model

Given an input video sample x, our end goal is to generate an adversarial example x' such that the output of the network $F(x')$ matches some target label t that is distinct from the original true label l. At the same time, we attempt to minimize the size of the perturbation necessary to reach x', so that the constructed adversarial example x' is indistinguishable from the original sample x to the human eye. We primarily focus on white-box attacks, where the adversary has access to the model parameters. Specifically, we use a set of pre-trained weights that are included with the LipNet model implementation [11]. We explore two avenues of generating our adversarial examples. Firstly, we examine the applicability of existing software on the model, and secondly, we turn to custom attack generation, loosely following the methods outlined by Carlini and Wagner [6].

IV. Gradient Based Attacks

Szegedy et al. [3] show that the problem of generating a minimally perturbed adversarial example can be distilled down to a relatively simple optimization problem:

[1]A viseme is a generic facial image that can be used to describe a particular sound.

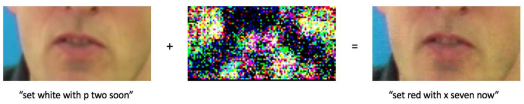

Fig. 3: Vanilla Fast Gradient Sign Method example.

$$\min_{\delta} \|\delta\|_p \text{s.t.} \quad F(x+\delta) = tx + \delta \in [0,1]^n \tag{1}$$

where δ represents our perturbation, t is our target label, $x + \delta = x'$ is our adversarial example, and $\|\delta\|_p$ is our choice of distance metric. This minimization problem solves for a perturbation δ that generates an adversarial example that is as close to the original input as possible subject to the constraints that the networks prediction of the adversarial example $x + \delta$ is the target label and is a feasible example (pixel values are mapped between 0 and 1). Carlini and Wagner [6] show that with a few small tweaks to this problem formulation, a good local solution can be found by using standard optimization algorithms (stochastic gradient descent, RMSProp, Adam, etc.) The choice of distance metric is flexible and should be taken into consideration by a potential adversary. In this paper we will investigate the L_∞ norm and the L_2 norm as distance metrics and outline the pros and cons of each choice.

A. Fast Gradient Sign Method

Before attempting a targeted attack, we first demonstrate that it is possible to generate untargeted adversarial examples with minimal perturbation using a gradient-based attack. Goodfellow et al. [8] shows that the Fast Gradient Sign Method is essentially a one-step linear approximation to maximizing the loss with respect to the true label by adding a clipped gradient of the loss to the original input. To implement this attack on the LipNet model, we use Keras backend functions to extract the gradient of the CTC loss with respect to the sample input video x and apply the one-step attack below to generate our adversarial example x':

$$x' = x + \epsilon * sign[\nabla_x L(x, y_{true})] \tag{2}$$

where L is the CTC loss function and y_{true} is the true label. Using an ϵ value of .025, we are able to successfully perturb the sample input – which LipNet accurately predicts as saying 'set white with p two soon' to generate an adversarial input that LipNet now predicts to say 'set red with x seven now'. Limiting the perturbation of each pixel to ϵ is equivalent to using the L_∞ norm as the distance metric for our attack. Thus, the per pixel difference between the original video and our adversarial example is at most 2.5%. We have extracted a random frame from the original video and the adversarial video to show the difference between the two in Figure 3.

The benefit of using Fast Gradient Sign Method is that it is able to quickly generate an adversarial attack that results in an output that is relatively distinct from the original label while being almost undetectable to the human eye. However, the limitations on FGSM are clear when we try to generate targeted attacks, which almost always fail with small ϵ. Although it has been shown that single step FGSM can generate targeted attacks in the image domain, its ineffectiveness on video input suggests that the LipNet model does not share the property of being locally linear, perhaps due to the use of CTC loss for sequential data.

B. L_∞ Attack: Iterative FGSM

While generating fast, untargeted attacks can be useful for certain objectives, an adversary would optimally like to trick a deep network into predicting a certain target phrase. For example, in a criminal case where a neural net for lipreading is used as evidence, an adversary may want to undermine the authority of the network by predicting a phrase that may exonerate the convict in question.

In order to create targeted attacks on LipNet, we move away from one-step attacks to iterative methods. Luckily, our FGSM attack from before can easily be adapted to an iterative attack [12]. We now focus on solving the minimization problem from above using the L_∞ norm as our distance metric. Since it is difficult to take the gradient of the L_∞ norm, we approximate the solution with iterative FGSM. At each iteration, we perform FGSM from before with two changes. First, instead of adding the gradient with respect to the true label, we subtract the gradient with respect to our target label. Second, we clip the gradient by $\frac{\epsilon}{\alpha}$ where α is the number of iterations. Madry et. al. [12] shows that this method is equivalent to performing gradient descent on the objective and projecting each intermediate step onto the L_∞ norm. Our update rule is as follows:

$$x_{k+1} = x_k - \frac{\epsilon}{\alpha} * sign[\nabla_x L(x, t)] \tag{3}$$

Carlini and Wagner [4] show that iterative white-box targeted attacks are much more effective than single step attacks, which is corroborated by our results. On top of being better at reaching a target label, iterative FGSM does not perturb the original video any more than one-step FGSM, as the overall perturbation is bounded by the same ϵ value (.025). We are able to successfully construct an adversarial example for any target phrase in the GRID dataset within 30 iterations using this method. The effectiveness of iterative FGSM at generating targeted attacks is clear; however, this method is still an approximate solution to our minimization problem insofar as the examples produced are not minimally perturbed. While the size of the perturbation can be controlled by shrinking ϵ, we find that if ϵ is too small (below .015), iterative FGSM can take an increasingly large number of iterations to converge to a solution, if at all.

Thus, if our objective is to create a efficient targeted adversarial attack with minimal perturbation, we would like to more precisely solve the minimization problem from above.

C. L_2 Attack: Modified C-W Attack

Following the work of Carlini and Wagner [6], we formulate our minimization problem that works to find an optimal delta using the L_2 norm distance metric.

$$\min_{\delta} \|\delta\|_2 \text{s.t.} \quad F(x+\delta) = tx + \delta \in [0,1]^n \tag{4}$$

There are a few nice properties of the L_2 norm in the context of our problem. Since the L_2 norm is differentiable, we can now nicely take the gradient over the entire objective function with respect to delta. Additionally, the L_2 norm over the entire perturbation sums over all the frames, which intuitively means that frames that are not as relevant to the solutions will see little to no perturbation, whereas frames that matter more to the loss function see relatively more perturbation. As an aside, we note that this may allow for key frame analysis to see which frames are the most important for classification. Finally, if the goal is to prevent the human eye from distinguishing between the original video and an adversarial example, using the L_2 norm may be advantageous as the human eye is too slow to pick out individual frames in a video.

Carlini and Wagner show that the minimization problem can be recast as follows:

$$\min_{\delta} \|\delta\|_2 + \lambda * L(x+\delta) \quad \text{s.t.} \quad x + \delta \in [0,1]^n \tag{5}$$

where we incorporate our constraint $F(x+\delta) = t$ into our objective function by adding the CTC loss function multiplied by some hyperparameter λ. We can now directly run an optimization algorithm by taking the gradient across the entire objective function, and checking that the second constraint is not violated (we can simply bound our adversarial example to have values between 0 and 1). To solve this minimization problem, we perform gradient descent on delta with our update rule at each iteration being:

$$\delta_{k+1} = \delta_k - \eta * (\lambda \dot{\nabla}_\delta L(x+\delta, t) + 2\delta) \tag{6}$$

where we initialize δ_0 to a random Gaussian centered at 0 with a variance of 0.02.

From this update rule, we see that our hyperparameter λ controls the relative importance of being our attack being adversarial vs. being minimally perturbed. Based on our research, we find that larger values of lambda allow us to find a solution faster at the cost of the size of the perturbation being bigger. After experimenting with multiple values of λ from 0.001 to 100 on a log scale, we find that $\lambda = .1$ is best for constructing successful adversarial examples while keeping the perturbation unnoticeable.

While running gradient descent with a learning rate $\eta = 0.1$, we observe that our solution oscillates after 50+ iterations. To solve this problem, we employ a learning rate decay of .1 after a every 30 iterations while simultaneously increasing λ by a scale of 10. We find that introducing learning rate decay helps with convergence when the loss is sufficiently small, but oscillating. By increasing λ at the same time, we can put more importance on minimizing the loss with respect to the target versus keeping our perturbation small. The reason we can do this while avoiding an increase in the size of our perturbation is twofold. First, enough iterations of gradient descent have occurred to where we have a perturbation δ with a small L_2 norm that is already fairly close to a solution. Second, since we have also lowered our learning rate, any further updates to δ will be relatively small and will not drastically increase the size of the perturbation. Additionally, we stop decay from dropping our learning rate under .001.

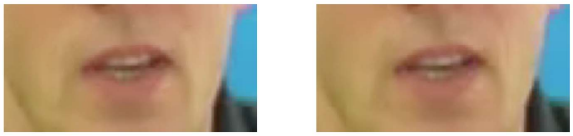

Fig. 4: Side-by-side comparison of the original image and the L_2 attack.

Figure 4 contains a side by side comparison of two frames extracted from the original sample and the adversarial one. Note that they are essentially identical to the human eye.

V. Results

With both our L_2 and L_∞ norm attacks, we are able to successfully find an adversarial example for any target phrase that follows the vocabulary of the GRID dataset. While this is a rather remarkable conclusion, it is not too surprising that these attack methods are able to transfer to the LipNet model with high success rates. We evaluate our two attacks on two criteria: 1. The average pixel change per frame and 2. The amount of time taken to find a solution. We measure these criteria across 10 different target examples each attack. For our L_∞ attack, we find that the smallest ϵ that we achieve efficient and consistent solutions with is .025. This means that the average perturbation per frame is bounded by a 2.5% change and in practice is often closer to a 1% change due to the gradient for certain pixels being 0, making our adversarial examples at least 97.5% similar to the original video. Convergence to a solution is relatively fast, taking at most 30 iterations and around 30 minutes on an AWS p2.xlarge GPU instance. The p2.xlarge GPU instance is the cheapest of GPUs available on AWS, hence the slow convergence time.

For our L_2 attack, our average pixel change per frame is measured to be $.33\%$, and the maximally perturbed frame has an average pixel change of 1%, making our adversarial examples over 99% similar to the original sample. However, the increase in similarity comes at the cost of an increase in the number of iterations required to find a solution. Depending on the target phrase, convergence generally takes between 40 and 120 iterations of gradient descent.

We encourage the reader to check out the links to the original sample video and an adversarial example that our L_2 attack generates[2]. The original video's true label is 'set white with p two soon' and the successfully attacked target label is 'lay red by q zero please'. The success of our attacks show that the video domain is equally susceptible to adversarial attack methods that have been explored in other domains, and we

[2]Original video sample: 'set white with p two soon' and adversarial video sample: 'lay red by q zero please'.

hope that our work encourages building robustness to attacks into deep neural networks.

VI. Conclusion

The vanilla fast gradient sign method, the L_∞ iterative fast gradient sign method, and the L_2 modified Carlini-Wagner attacks were highly effective in fooling the trained LipNet model. Both targeted and untargeted attacks were successful, and converged in reasonable time. The attacks bring attention to the power of adversarial attacks against video-based neural networks, which closely match the effectiveness of audio attacks presented by Carlini and Wagner [4]. The LipNet adversarial attacks presented appear to be the first of their kind, and join the family of adversarial attacks against video-related tasks.

VII. Future Considerations

In this paper we have demonstrated the effectiveness of white-box attacks on video classification problems. Previous work has shown that an interesting property of adversarial attacks is transferability, the idea that adversarial examples trained on one network will work on a different network, which is possible with black-box attacks. We believe that our work can be used to implement black-box attacks on video classification systems in the future.

One consideration includes reducing the perturbation δ magnitude, which can be achieved by running more iterations of gradient descent with a lower learning rate. Additionally, this process could be sped up by using better optimizers or by incorporating techniques such as momentum.

With every deep learning problem, a more general problem scope can benefit the real-world application of the task at hand. The sentence structures and the corpus are fixed and very refined, rendering the LipNet model's scope of lip reading very limited. Adversarial attacks against LipNet can only advance in complexity as LipNet does, so what remains are the types and effectiveness of attacks on the state-of-the-art LipNet models today.

Finally, in terms of model architecture, a breakdown of phoneme vulnerability could improve the optimization task solved by the adversarial agent. Should the agent be informed of the original sentence and be capable of detecting each used phoneme's presence in the sentence, the model can use the existing priors or the expected perturbation size given the phoneme. Visualization methods like a confusion matrix of phonemes, which measure the difference in perturbations for two targeted classes, could aid adversarial creators in designing future adversarial methods.

Acknowledgments

We would like to thank <Anonymous> and <Anonymous> for their mentorship, feedback, and debugging help throughout this project.

References

[1] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). *Lipnet: End-to-end sentence-level lipreading*. arXiv preprint arXiv:1611.01599.

[2] Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006, June). *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376). ACM.

[3] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. arXiv preprint arXiv:1312.6199.

[4] Carlini, N., & Wagner, D. (2018, May). *Audio adversarial examples: Targeted attacks on speech-to-text*. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 1-7). IEEE.

[5] Yuan, X., He, P., & Li, X. A. (2018). *Adaptive adversarial attack on scene text recognition*. arXiv preprint arXiv:1807.03326.

[6] Carlini, N., & Wagner, D. (2017, May). *Towards evaluating the robustness of neural networks*. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE.

[7] Yuan, X., He, P., Zhu, Q., & Li, X. (2019). *Adversarial examples: Attacks and defenses for deep learning*. IEEE transactions on neural networks and learning systems.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). *Generative adversarial nets*. In Advances in Neural Information Processing Systems.

[9] Gelder, B. D., Vroomen, J., & Van der Heide, L. (1991). *Face recognition and lipreading in autism*. European Journal of Cognitive Psychology, 3(1), 69-86.

[10] Barker, J., Cooke, M., Cunningham, S., & Shao, X. (2013). *The GRID audiovisual sentence corpus*.

[11] Rizkiarm (2017). *Keras implementation of 'LipNet: End-to-End Sentence-level Lipreading'*.

[12] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). *Towards deep learning models resistant to adversarial attacks*. arXiv preprint arXiv:1706.06083.

[13] Rauber, J., Brendel, W., & Bethge, M. (2017). *Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models*. arXiv preprint arXiv:1707.04131, 5.

[14] Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017, May). *Towards estimating the upper bound of visual-speech recognition: The visual lipreading feasibility database*. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 208-215). IEEE.

[15] Hassanat, A. B. (2014). *Visual words for automatic lipreading*. arXiv preprint arXiv:1409.6689.

[16] Easton, R. D., & Basala, M. (1982). *Perceptual dominance during lipreading*. Perception & Psychophysics, 32(6), 562-570.

[17] Wei, X., Zhu, J., & Su, H. (2018). *Sparse adversarial perturbations for videos*. arXiv preprint arXiv:1803.02536.

[18] Meng, L., Zhao, B., Chang, B., Huang, G., Tung, F., & Sigal, L. (2018). *Where and When to Look? Spatio-temporal Attention for Action Recognition in Videos*. arXiv preprint arXiv:1810.04511.

[19] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*. In Proceedings of the 23rd international conference on Machine learning, pages 369-376. ACM, 2006.

[20] Guan, M. Y., & Valiant, G. (2019). *A Surprising Density of Illusionable Natural Speech*. arXiv preprint arXiv:1906.01040.

[21] Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Chowdhury, A. K. R., & Swami, A. (2018). *Adversarial perturbations against real-time video classification systems*. arXiv preprint arXiv:1807.00458.

[22] Rey-de-Castro, R., & Rabitz, H. (2018). *Targeted nonlinear adversarial perturbations in images and videos*. arXiv preprint arXiv:1809.00958.

[23] Jiang, L., Ma, X., Chen, S., Bailey, J., & Jiang, Y. G. (2019). *Black-box Adversarial Attacks on Video Recognition Models*. arXiv preprint arXiv:1904.05181.

[24] Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Roy-Chowdhury, A. K., Swami, A. (2019). *Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems*. In NDSS.