APPENDIX
ON ADVERSARIAL ROBUSTNESS OF LARGE-SCALE AUDIO VISUAL LEARNING

Juncheng B Li, Shuhui Qu*, Xinjian Li, Po-Yao (Bernie) Huang, Florian Metze*

Carnegie Mellon University

1. PROOF OF THEOREM 1

Theorem 1 *There exists a sample $x_i \in \mathcal{D}$, and a unimodal sample-wise attack $\exists ||\delta_{A,i}||_p \leq \epsilon_A$ or $\exists ||\delta_{V,i}||_p \leq \epsilon_V$ that can break a multimodal fusion network $f(x_{V,i} \oplus x_{A,i}), y_i)$, changing its prediction label y_i.*

Here, $\mathcal{D}$ is the dataset, and $\epsilon_{A,i}$ and $\epsilon_{V,i}$ are the point-wise robustness threshold for each uni-modal of sample x_i. Therefore, as a conjecture, a unimodal attack can break a multimodal model, which we empirically verified the existence of such cases in our experiments.

Let's consider a binary classification task as an example for simplicity. Let $(x_{A,i}, x_{V,i})$ be a point with different prediction results for audio modality A and video modality V. Assume $\exists a, b$, such that $a^T g(x_{A,i}) = -s < 0$ and $b^T h(x_{V,i}) = t > 0$ where $s, t > 0$, and the correct label is -1. For the point-wise robustness threshold $\epsilon_{A,i}$ of this point, where an attack $\{\delta_{\epsilon_{A,i}} : ||\delta_{\epsilon_{A,i}}||_P \leq \epsilon_{A,i}\}$ changes the prediction label. By definition, we know $a^T g(x_{A,i} + \delta_{\epsilon_{A,i}}) \geq 0$ and $a^T g(x_{A,i} + \delta) \leq 0$ for all $0 \leq \delta \leq \epsilon_{A,i}$. If $s < t$, then the fused network predicted the wrong label even without any noise.

$$f(x_{A,i}, x_{V,i}) = (a, b)^T (g(x_{A,i}) \oplus h(x_{V,i})) \tag{1}$$
$$= a^T g(x_{A,i}) + b^T h(x_{V,i}) \tag{2}$$
$$= -s + t > 0 \tag{3}$$

Otherwise, in the case of $s \geq t$, by applying Intermediate Value Theorem to $g(x)$, there exists a point $0 \leq \delta \leq \epsilon_A$ such that $a^T g(x_{A,i} + \delta) = -t/2$:

$$\begin{aligned} f(x_{A,i} + \delta, x_{V,i}) &= (a, b)^T (g(x_{A,i} + \delta) \oplus h(x_{V,i})) \\ &= a^T g(x_{A,i} + \delta) + b^T h(x_{V,i}) \\ &= -\frac{t}{2} + t > 0 \end{aligned} \tag{4}$$

In both cases, we could find a noise $0 \leq \delta < \epsilon_{A,i}$ within the original unimodal robustness threshold to attack the multimodal network successfully. Vise versa for video. Thus, a unimodal attack can break a mulimodal model, which we also empirically verified the existence of such cases in our experiments. (see Table 2 of the main paper).

We postulate that such phenomenon is like the Mcgurk Effect, where multimodal fusion would further distort the already non-convex decision boundary (Figure 1 of the main paper), making the fused decision boundary very different than the original ones and unpredictable.

* equal contribution