Input Reaction

Model Top-1 Prediction + Input Attribution

$$\begin{array}{c} \text{CI} \\ \text{HCI} + \\ \text{CI} \xrightarrow{\text{Al}} \text{CI} \\ \text{ua: 0.17} \\ \text{ua: 0.21} \\ \text{ua: 0.33} \\ \end{array} + \begin{array}{c} \text{ua: 0.024} \\ \text{0.031} \\ \text{ua: 0.19} \\ \end{array} + \begin{array}{c} \text{ua: 0.024} \\ \text{0.031} \\ \text{ua: 0.19} \\ \end{array}$$

ua = uniform attribution

(b) Adversarial Examples

HCI +
$$\begin{pmatrix} CI \\ CI \end{pmatrix} + \begin{pmatrix} CI \\ AI \end{pmatrix} \begin{pmatrix} CI \\ CI \end{pmatrix} + \begin{pmatrix} CI \\ CI \end{pmatrix} \begin{pmatrix} CI \\ CI \end{pmatrix} + \begin{pmatrix} CI \\ CI \end{pmatrix} \begin{pmatrix}$$

b) Adversarial Examples

HCI +
$$CI \stackrel{AI}{\longrightarrow} CI$$
 + $CI \stackrel{AI}{\longrightarrow} CI$ +

b) Adversarial Examples

HCI +
$$\frac{\text{CI}}{\text{CI}}$$
 + $\frac{\text{N}}{\text{N}}$ $\frac{\text{O}}{\text{CI}}$ + $\frac{\text{N}}{\text{N}}$ $\frac{\text{O}}{\text{O}}$ + $\frac{\text{O}}{\text{O}}$ $\frac{\text{$