Boosting Adversarial Attacks with Momentum
Yinpeng Dong\textsuperscript{1}, Fangzhou Liao\textsuperscript{1}, Tianyu Pang\textsuperscript{1}, Hang Su\textsuperscript{1}, Jun Zhu\textsuperscript{1}, Xiaolin Hu\textsuperscript{1}, Jianguo Li\textsuperscript{2}
\textsuperscript{1}Department of Computer Science and Technology, Tsinghua University, \textsuperscript{2}Intel Labs China

\section*{Introduction}

\begin{itemize}
  \item Adversarial examples are crafted by adding small, human-imperceptible noises to legitimate examples, but make a model output attacker-desired inaccurate predictions.
  \item Adversarial attacks:
    \begin{itemize}
      \item Identify the robustness of deep learning models.
      \item Provide more varied training data (i.e., adversarial training).
    \end{itemize}
\end{itemize}

\section*{Methodology}

\begin{itemize}
  \item Generating adversarial examples:
    \begin{itemize}
      \item Constrained optimization problem:
        \[
        \arg \max_{\mathbf{x}^*} f(\mathbf{x}^*, y) \quad \text{s.t.} \\| \mathbf{x}^* - \mathbf{x} \|_\infty \leq \epsilon
        \]
      \item Fast gradient sign method (FGSM, Goodfellow et al., 2015):
        \[
        \mathbf{x}^* = \mathbf{x} + \epsilon \cdot \text{sign}(\mathbf{g}(\mathbf{x}, y))
        \]
      \item Iterative fast gradient sign method (I-FGSM, Kurakin et al., 2016):
        \[
        \mathbf{x}_t = \mathbf{x}, \quad \mathbf{x}_{t+1} = \text{clip}(\mathbf{x}_t + \alpha \cdot \text{sign}(\mathbf{g}(\mathbf{x}_t, y)))
        \]
      \item Optimization-based method (Carlini and Wagner, 2017):
        \[
        \arg \min_{\mathbf{x}} \mathcal{L}_{\text{fgs}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, y)
        \]
    \end{itemize}
  \item Transferability:
    \begin{itemize}
      \item The adversarial examples generated for one model can also fool another model (Liu et al., 2017).
      \item Black-box attacks: how to generate more efficient adversarial examples for a black-box model (challenge).
    \end{itemize}
  \item Optimization with Momentum (Polyak, 1964):
    \begin{itemize}
      \item Accelerate gradient descent
      \item Escape from poor local minima and maxima
      \item Stabilize update directions of stochastic gradient descent
    \end{itemize}
\end{itemize}

\section*{Motivation}

\begin{itemize}
  \item The trade-off between the attack ability and transferability
    \begin{enumerate}
      \item FGSM: more transferable adversarial examples; low success rates for the white-box models. \textit{(Reason: linear assumption may not hold for large distortion; “underfit” the model.)}
      \item I-FGSM: high success rates for white-box models; poor transferability. \textit{(Reason: drop into poor local maxima; “overfit” the model.)}
    \end{enumerate}
\end{itemize}

\section*{Experiments}

\begin{itemize}
  \item Attacking a single model
    \begin{itemize}
      \item Table of success rates for different methods on different models.
    \end{itemize}
  \item Attacking an ensemble of models
    \begin{itemize}
      \item Table of success rates for different ensemble methods on different models.
    \end{itemize}
  \item Ablation studies
    \begin{itemize}
      \item Figures showing the impact of different factors on the success rates.
    \end{itemize}
\end{itemize}

\section*{Conclusion}

\begin{itemize}
  \item We propose a broad class of momentum-based iterative methods for generating more transferable adversarial examples.
  \item We propose to attack an ensemble of models whose logits are fused.
  \item Our method won the first places in both of the NIPS 2017 Non-targeted Adversarial Attack and Targeted Adversarial Attack competitions.
  \item Code available at: [link]
\end{itemize}