Towards Transferable Targeted Attack

Anonymous CVPR submission

Paper ID 4937

Appendix

A. Proof of Eq. 7

Denote o_i as the softmax input with C classes, p_i as corresponding output, then the derivative $\frac{\partial p_i}{\partial o_i}$ of the output $\mathbf{y}$ of the softmax function with respect to its input $\mathbf{o}$ can be calculated as:

$$\frac{\partial y_i}{\partial o_j} = \frac{\partial \frac{e^{o_c}}{\sum_i^C e^{o_i}}}{\partial o_j} \tag{1}$$

denote $\sum_i^C e^{o_i}$ as $\sum_C$. If $i = j$, we get:

$$\begin{aligned} \frac{\partial y_i}{\partial o_i} &= \frac{e^{o_i}\Sigma_C - e^{o_i}e^{o_i}}{\Sigma_C^2} \\ &= \frac{e^{o_i}}{\Sigma_C^2}(1 - \frac{e^{o_i}}{\Sigma_C}) \\ &= p_i(1 - p_i) \end{aligned} \tag{2}$$

else there has:

$$\frac{\partial y_i}{\partial o_j} = \frac{0 - e^{o_i}e^{o_j}}{\Sigma_C^2} = -\frac{e^{o_j}}{\Sigma_C}\frac{e^{o_i}}{\Sigma_C} = -p_i p_j \tag{3}$$

Let $\xi(y, p) = -y \cdot \log(p)$ denote cross-entropy error function, and then The derivative $\frac{\partial \xi}{z_i}$ of the loss function with respect to the softmax input z_i can be calculated as:

$$\begin{aligned} \frac{\partial \xi}{o_i} &= -\sum_{j=1}^{C} \frac{\partial y_j \log(p_j)}{\partial o_i} \\ &= -\sum_{j=1}^{C} y_j \frac{1}{p_i}\frac{\partial p_j}{\partial o_i} = -\frac{y_i}{p_i}\frac{\partial p_i}{\partial o_i} - \sum_{j\neq i}^{C} \frac{y_j}{p_j}\frac{\partial p_j}{\partial o_i} \\ &= -\frac{y_i}{p_i}p_i(1 - p_i) - \sum_{j\neq i}^{C} \frac{y_j}{p_j}(-p_i p_j) \\ &= -y_i(1 - p_i) + \sum_{j\neq i}^{C} y_j p_i = -y_i + p_i \sum_{j}^{C} y_j \\ &= p_i - y_i \end{aligned} \tag{4}$$

where $\sum_j^C y_j = 1$ as y is one hot label.

B. Additional Experiments

The results of ensemble networks (white-box setting) are shown in Section 4, and we will show more results of the proposed method on all the six models respectively (black-box setting for itself and white-box setting for the other five models) here. We adopt the same experimental setups with Section 4 including dataset, networks, and parameters. We also attack the adversarially trained models using our method. The results are shown in Table 1 for normally trained models which are hold-out, Table 2 for adversarially trained models. All of these results are conducted based on DI^2-FGSM and TI-FGSM. For black-box setting, we can always get the best performance, and for white-box setting, we can get the best performance on most of the situations.

In addition, extensive experiments about the parameter ϵ which means the maximum of the noise allowed to add are conducted to further evaluate the proposed method. In some sense, these results also give a seriers of basic references for targeted attack and non-targeted attack on the six models with different ϵ. The results are shown in Table 3 for ensemble models (white-box setting) with different ϵ, and Table 4 for hold-out models (black-box setting) with different ϵ. For black-box setting, we can always do the best. And for white-box setting, when ϵ is small, we can get higher performance than the baseline, and as ϵ gets bigger and bigger, the gap between baseline and our method becomes smaller and smaller because the success rates of attack is tending to 100% and there is little space for performance imporvement.

Attack	-Inc-v3	-Inc-v4	-IncRes-v2	-Res-50	-Res-101	-Res-152
MI-FGSM	11.9	9.0	8.4	16.3	20.1	19.6
DI2-FGSM	28.0	26.8	25.9	29.5	32.1	32.6
Ours	37.0	32.6	30.6	36.0	39.7	39.6
Ours+Trip	**38.3**	**36.6**	**32.0**	**38.5**	**41.2**	**40.6**
TI-FGSM	29.8	27.5	28.8	29.6	33.7	34.4
Ours	39.3	36.3	34.6	37.5	40.8	41.3
Ours+Trip	**39.5**	**36.6**	**35.1**	**39.3**	**43.0**	**42.9**

Table 1. The success rates (%) of adversarial attacks compared to MI-FGSM, DI^2-FGSM and TI-FGSM on six respective hold-out models—Inc-v3, Inc-v4, IncRes-v2, Res-50, Res-101, and Res-152, in other words, for black-box setting. The adversarial examples are crafted from ensemble networks using the six models except the hold-out one. The sign "−" indicates the hold-out network.

Model	Attack	Inc-v3 ens3	Inc-v3 ens4	IncRes-v2ens	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-101	Res-152
Inc-v3 ens3	DI^2-FGSM	0.7*	**66.0**	**37.2**	79.8	70.9	66.5	56.9	62.4	61.1
	Ours	1.1*	64.1	28.9	**92.4**	83.1	73.8	73.0	**77.6**	76.1
	Ours+Trip	**1.2***	65.9	26.2	91.8	**85.4**	**77.0**	**75.6**	76.6	**76.7**
	TI-FGSM	13.7*	55.7	**39.0**	88.5	79.5	76.1	65.3	69.1	67.2
	Ours	17.7*	**59.2**	27.4	94.0	**88.0**	81.2	78.0	**82.6**	80.4
	Ours+Trip	**18.1***	55.7	30.0	**94.1**	86.8	**83.2**	**79.4**	80.5	**81.5**
Inc-v3 ens4	DI^2-FGSM	64.8	1.1*	**39.8**	75.9	66.3	61.9	54.6	57.7	57.0
	Ours	66.8	**1.5***	27.7	90.7	79.5	73.0	71.2	74.9	74.7
	Ours+Trip	**67.0**	**1.5***	30.8	**92.5**	**81.7**	**74.5**	**72.4**	**75.1**	**75.4**
	TI-FGSM	58.4	10.4*	**37.9**	84.6	77.2	74.9	64.9	67.2	67.2
	Ours	62.1	12.9*	23.7	94.0	87.0	81.5	77.1	81.3	79.4
	Ours+Trip	**62.6**	**14.6***	27.0	**94.7**	**87.9**	**82.3**	**77.7**	**81.7**	**81.6**
IncRes-v2ens	DI^2-FGSM	61.2	**62.9**	0.5*	75.3	65.4	60.2	53.3	57.5	55.8
	Ours	**62.3**	61.3	1.0*	89.0	78.3	70.5	67.6	73.0	71.8
	Ours+Trip	60.4	62.7	**1.2***	**90.1**	**78.8**	**72.1**	**71.1**	**74.5**	**74.2**
	TI-FGSM	58.2	55.5	6.1*	83.8	77.7	73.8	65.9	69.9	67.4
	Ours	61.0	**57.0**	7.7*	**94.3**	**87.1**	80.0	76.1	78.8	**79.9**
	Ours+Trip	**61.9**	56.6	**8.4***	93.6	86.7	**80.9**	**76.9**	**81.3**	79.5

Table 2. The success rates (%) of adversarial attacks compared to DI^2-FGSM and TI-FGSM against three adversarially trained models—Inc-v3ens3, Inc-v3ens4, and IncRes-v2ens, and six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-50, Res-101, and Res-152. The adversarial examples are crafted from ensemble networks and for the three adversarially trained models in the first column respectively, the sign "$*$" indicates the black-box setting.

Ensemble		max-epslion												
Model	Attack	8	10	12	14	16	18	20	22	24	26	28	30	32
Inc-v3	DI2-FGSM	60.2	70.8	77.8	82.9	86.7	91.1	92.1	93.1	95.3	96.7	96.4	97.0	97.9
	Ours	70.6	79.3	85.9	**89.6**	**92.5**	93.6	94.8	**96.3**	95.9	96.6	**97.5**	**98.0**	97.9
	Ours+Trip	**72.3**	**81.0**	**86.0**	88.8	91.7	**94.0**	**94.9**	95.4	**96.8**	**97.0**	97.1	97.4	**98.0**
	TI-FGSM	60.4	73.2	78.6	85.7	89.7	92.4	93.9	**95.7**	**96.4**	**97.3**	97.2	**98.4**	**98.5**
	Ours	69.5	**79.2**	**85.6**	**89.9**	91.2	**93.4**	**94.7**	95.3	96.3	97.2	**97.4**	97.4	98.0
	Ours+Trip	**69.8**	78.2	83.7	88.4	**91.4**	91.7	93.8	95.4	95.8	96.4	**97.4**	97.1	96.8
Inc-v4	DI2-FGSM	63.5	74.0	79.8	85.5	87.9	92.1	93.4	94.3	94.8	96.5	96.5	97.7	**98.1**
	Ours	72.1	80.6	86.5	89.3	91.5	93.0	94.2	94.9	96.1	96.8	**97.5**	97.6	97.1
	Ours+Trip	**73.5**	**82.3**	**86.8**	**89.4**	**92.8**	**93.9**	**95.4**	**95.8**	**96.3**	**97.3**	96.9	**97.8**	97.6
	TI-FGSM	62.0	70.8	79.7	85.1	87.7	92.1	93.1	94.3	95.5	96.5	**96.9**	97.2	**97.7**
	Ours	69.5	78.3	83.6	**88.5**	90.6	92.5	93.6	**95.6**	**96.5**	96.0	96.7	**97.6**	97.2
	Ours+Trip	**70.9**	**80.9**	**86.8**	**88.5**	**91.7**	**93.0**	**94.9**	95.3	96.1	**96.8**	**96.9**	97.2	97.5
IncRes-v2	DI2-FGSM	49.2	62.2	69.5	76.5	81.9	84.1	87.8	89.4	90.9	92.1	93.4	94.2	**95.5**
	Ours	57.9	68.4	77.7	80.7	84.3	87.8	89.2	90.8	90.4	93.6	94.0	**94.7**	94.9
	Ours+Trip	**63.3**	**72.7**	**79.1**	**84.8**	**87.7**	**89.6**	**90.5**	**93.1**	**94.8**	**94.4**	**95.1**	**94.7**	95.3
	TI-FGSM	62.3	72.6	79.9	85.6	88.9	90.8	92.5	95.2	95.4	**96.8**	**97.0**	97.3	97.2
	Ours	69.3	78.2	85.1	87.4	89.9	93.0	**94.2**	95.2	95.6	**96.8**	96.4	**98.1**	**98.0**
	Ours+Trip	**73.4**	**81.3**	**86.4**	**89.8**	**91.4**	**93.6**	94.1	**95.4**	**96.1**	94.9	96.9	97.7	97.3
Res-50	DI2-FGSM	51.4	64.3	74.2	80.3	84.2	86.8	88.1	90.8	91.6	93.3	94.1	94.1	95.4
	Ours	62.6	72.5	78.5	85.1	87.0	88.2	92.3	93.2	93.5	93.7	94.8	95.7	95.6
	Ours+Trip	**65.5**	**76.9**	**82.1**	**85.5**	**88.8**	**90.2**	**93.5**	**93.8**	**94.7**	**94.9**	**96.2**	**96.8**	**95.9**
	TI-FGSM	50.6	63.5	69.7	77.7	82.0	85.0	88.6	89.8	91.9	93.3	93.6	95.0	95.3
	Ours	59.5	69.4	76.9	82.2	85.0	88.8	90.1	**92.2**	92.8	94.2	**94.4**	94.3	95.6
	Ours+Trip	**62.5**	**74.1**	**80.7**	**85.4**	**87.5**	**90.0**	**91.4**	91.4	**94.5**	**95.3**	94.2	**95.3**	**96.4**
Res-101	DI2-FGSM	53.4	64.8	75.4	80.9	83.4	87.0	88.9	91.6	93.9	94.7	94.8	95.7	95.7
	Ours	62.5	73.2	80.0	84.3	86.6	89.7	91.7	93.2	93.8	**94.9**	**95.9**	**95.9**	96.4
	Ours+Trip	**68.3**	**77.4**	**82.2**	**86.4**	**89.2**	**91.7**	**93.1**	**94.5**	**95.0**	94.8	95.8	96.8	**96.8**
	TI-FGSM	52.9	64.6	72.0	78.3	82.3	85.9	89.0	90.5	93.1	93.2	93.8	94.6	**96.4**
	Ours	60.9	72.0	76.5	82.3	86.0	88.3	89.4	92.8	93.3	93.7	94.4	95.4	94.9
	Ours+Trip	**66.3**	**74.4**	**79.7**	**83.9**	**87.7**	**89.3**	**92.1**	**93.2**	**94.5**	**95.0**	**94.9**	**95.9**	95.8
Res-152	DI2-FGSM	55.3	69.5	78.1	82.1	85.3	89.7	91.6	92.3	93.8	**95.4**	95.5	96.5	95.7
	Ours	64.8	75.9	82.0	86.9	88.9	91.7	91.8	93.8	**94.9**	94.9	95.8	**95.9**	**96.8**
	Ours+Trip	**70.2**	**79.3**	**84.5**	**89.0**	**91.4**	**92.0**	**94.2**	**94.7**	94.8	**95.4**	**96.2**	**95.9**	96.7
	TI-FGSM	54.3	66.8	75.1	80.0	84.8	87.6	90.3	91.1	94.0	94.4	**95.7**	95.1	96.3
	Ours	61.2	73.9	80.2	83.7	87.1	90.1	90.9	93.3	93.7	94.9	95.6	**96.2**	**96.8**
	Ours+Trip	**66.8**	**76.8**	**83.3**	**86.7**	**90.6**	**91.0**	**93.6**	**94.1**	**95.2**	**96.0**	**95.7**	95.7	96.3

Table 3. The success rates (%) of adversarial attacks compared to DI^2-FGSM and TI-FGSM against ensemble networks with different ϵ which are ensembled from the five networks except the current network, all the results are in white-box setting.

Hold-out		max-epslion												
Model	Attack	8	10	12	14	16	18	20	22	24	26	28	30	32
Inc-v3	DI2-FGSM	8.1	12.3	18.6	23.3	27.5	33.2	37.4	39.7	44.2	45.5	47.9	52.8	52.2
	Ours	**12.9**	19.2	**27.1**	31.5	37.7	41.4	46.1	47.8	51.7	54.5	**57.6**	**59.5**	**60.1**
	Ours+Trip	12.3	**20.6**	26.3	**32.8**	**38.4**	**43.7**	**47.0**	**49.9**	**53.8**	**55.9**	55.2	58.6	59.1
	TI-FGSM	12.1	18.9	24.0	31.9	36.0	40.3	46.5	49.4	53.5	56.3	58.0	59.7	61.0
	Ours	11.7	19.3	27.3	32.7	37.4	40.8	45.8	49.4	52.6	54.2	56.6	58.6	59.4
	Ours+Trip	**13.9**	**20.7**	**29.0**	**34.5**	**40.1**	**44.6**	**48.4**	**51.8**	**53.9**	**57.3**	**59.3**	**62.7**	**63.7**
Inc-v4	DI2-FGSM	8.1	11.7	17.9	20.8	26.1	30.4	35.9	38.1	41.5	42.5	46.0	50.6	50.8
	Ours	11.3	16.3	23.4	**30.3**	35.2	38.4	40.8	43.4	**48.2**	50.5	52.9	55.7	57.1
	Ours+Trip	**11.8**	**17.7**	**25.5**	30.1	**36.1**	**39.9**	**42.9**	**47.5**	47.2	**51.5**	**53.0**	**56.1**	**58.5**
	TI-FGSM	7.5	13.2	19.4	24.0	26.5	33.5	37.5	41.7	45.8	47.8	51.4	52.6	56.2
	Ours	11.9	17.8	24.9	31.1	35.2	39.6	43.6	48.1	**50.8**	53.0	56.3	58.2	60.7
	Ours+Trip	**12.7**	**19.2**	**25.3**	**33.6**	**36.6**	**41.0**	**45.4**	**48.9**	50.2	**54.4**	**57.1**	**60.1**	**62.2**
IncRes-v2	DI2-FGSM	7.1	11.6	18.0	22.5	27.4	32.0	35.5	37.1	42.9	42.9	46.9	50.2	49.7
	Ours	9.4	15.7	21.5	27.4	32.8	35.2	38.7	43.2	46.3	47.4	51.0	53.5	55.6
	Ours+Trip	**11.0**	**16.5**	**23.3**	**28.9**	**34.3**	**37.4**	**41.3**	**45.5**	**49.4**	**49.8**	**53.4**	**54.3**	**57.0**
	TI-FGSM	7.7	13.6	19.3	25.3	29.2	34.2	37.1	41.6	43.8	48.6	51.2	54.6	54.7
	Ours	10.9	17.4	**25.5**	28.9	34.4	38.0	41.6	46.4	48.2	52.2	53.8	55.8	57.7
	Ours+Trip	**11.3**	**18.1**	24.5	**31.7**	**36.7**	**39.7**	**42.9**	**48.5**	**50.6**	**52.6**	**56.3**	**56.7**	**58.2**
Res-50	DI2-FGSM	8.6	13.6	21.2	23.9	27.3	32.0	36.3	40.4	42.0	44.7	46.7	48.3	50.8
	Ours	14.2	**22.5**	28.1	32.6	37.1	40.9	44.6	47.8	49.9	53.1	56.0	**56.9**	57.7
	Ours+Trip	**15.6**	21.7	**28.7**	**34.4**	**37.3**	**42.7**	**45.1**	**48.8**	**50.9**	**53.3**	**56.2**	56.0	**58.2**
	TI-FGSM	9.1	14.9	21.3	25.5	31.8	35.9	36.5	40.2	44.6	46.9	49.2	52.4	52.8
	Ours	15.1	**22.3**	28.5	34.9	38.4	42.2	45.8	49.9	**52.1**	54.0	55.6	57.5	**59.7**
	Ours+Trip	**16.3**	21.7	**28.6**	**35.0**	**39.4**	**45.1**	**47.6**	**51.0**	51.6	**54.2**	**56.3**	**57.7**	58.9
Res-101	DI2-FGSM	8.9	14.7	21.4	27.1	31.7	36.8	40.3	44.5	47.3	50.0	53.8	55.1	57.6
	Ours	13.8	21.6	27.7	**35.0**	**41.6**	46.2	46.5	51.9	54.3	**57.8**	**59.7**	**60.2**	**62.4**
	Ours+Trip	**15.2**	**23.0**	**28.8**	**35.0**	40.4	**46.4**	**49.6**	**52.8**	**55.1**	56.0	58.3	**60.2**	61.3
	TI-FGSM	10.2	16.1	22.7	28.9	35.2	39.2	42.8	45.1	49.7	52.5	55.3	58.2	60.4
	Ours	15.1	**23.9**	30.2	36.6	41.6	**45.1**	48.4	**54.3**	55.3	58.4	59.7	**62.8**	63.2
	Ours+Trip	**16.4**	23.8	**30.5**	**37.5**	**41.7**	44.6	**51.2**	53.1	**56.2**	**59.5**	**62.0**	62.0	**64.9**
Res-152	DI2-FGSM	11.0	17.0	23.6	28.6	33.5	37.2	41.8	45.3	48.5	51.1	53.0	56.2	56.6
	Ours	16.0	21.6	27.8	**35.6**	39.9	45.3	47.5	50.4	**55.3**	55.5	57.3	**60.2**	**61.4**
	Ours+Trip	**16.8**	**23.8**	**30.4**	35.1	**41.0**	**45.7**	**48.7**	**51.3**	53.8	**56.9**	**58.2**	**60.2**	61.0
	TI-FGSM	11.0	17.7	24.2	28.6	35.7	39.3	44.2	47.1	51.4	55.0	57.8	57.7	60.1
	Ours	15.3	24.1	29.9	**36.4**	**42.2**	44.6	**50.7**	**54.0**	55.0	**57.3**	**61.0**	61.5	**63.8**
	Ours+Trip	**16.3**	**25.0**	**32.6**	36.0	42.1	**47.3**	50.0	52.3	**55.2**	56.9	57.6	**61.6**	63.1

Table 4. The success rates (%) of adversarial attacks compared to DI^2-FGSM and TI-FGSM against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-50, Res-101, and Res-152 with different ϵ. The adversarial examples are crafted from ensemble networks. All the results are in black-box setting.