UNIVERSITÉ DU LUXEMBOURG

LGV

Transferability from Large Geometric Vicinity

Presented by Martin Gubri

LGV: Boosting Adversarial Example Transferability from Large Geometric Vicinity

Martin Gubri[1], Maxime Cordy[1], Mike Papadakis[1], Yves Le Traon[1], and Koushik Sen[2]

[1] Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg, LU firstname.lastname@uni.lu
[2] University of California, Berkeley, CA, USA

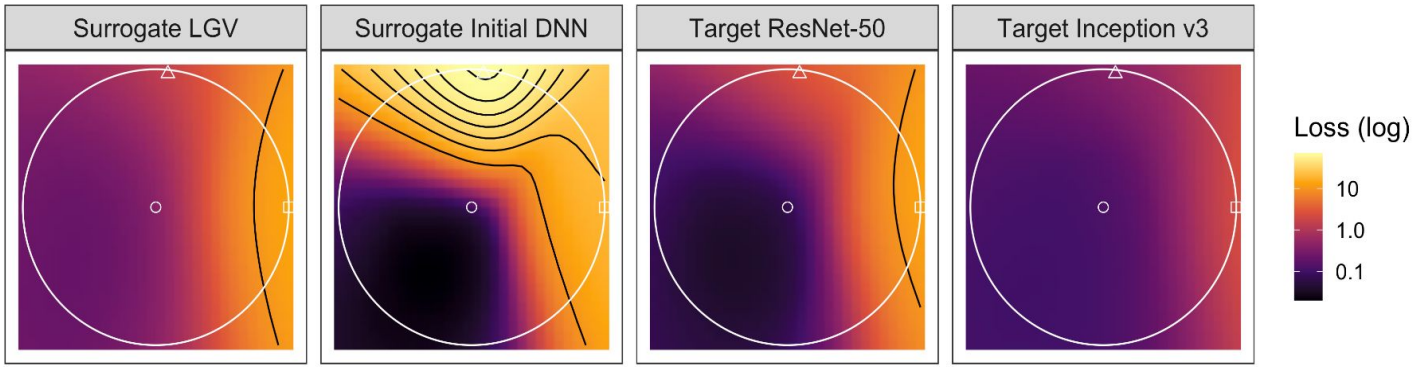
Abstract. We propose transferability from Large Geometric Vicinity (LGV), a new technique to increase the transferability of black-box adversarial attacks. LGV starts from a pretrained surrogate model and collects multiple weight sets from a few additional training epochs with a constant and high learning rate. LGV exploits two geometric properties that we relate to transferability. First, models that belong to a wider weight optimum are better surrogates. Second, we identify a subspace able to generate an effective surrogate ensemble among this wider optimum. Through extensive experiments, we show that LGV alone outperforms all (combinations of) four established test-time transformations by 1.8 to 59.9 percentage points. Our findings shed new light on the importance of the geometry of the weight space to explain the transferability of adversarial examples.

Accepted at ECCV 22

LGV in a nutshell

LGV is an adversarial attack that:

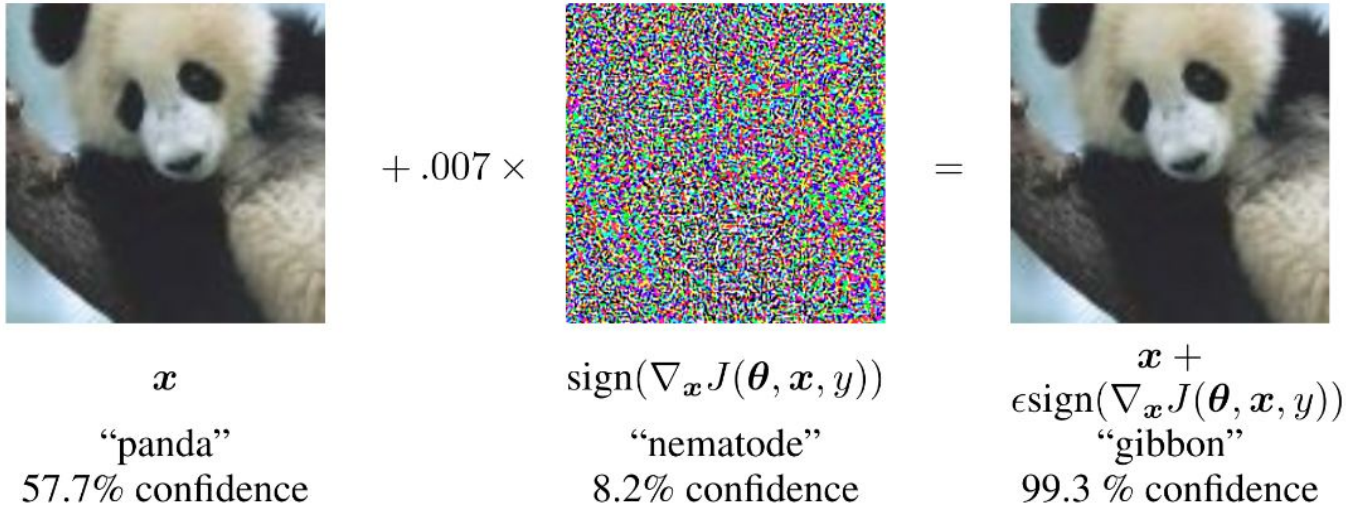
- improves the **transferability** (generalization) of adversarial examples,
- simply attacks models collected along the SGD trajectory with a high learning rate,
- beats (all combinations of) four state-of-the-art techniques,
- exploits the geometry of the weight space to find **flatter adversarial examples** in the feature space.



Background

Adversarial examples

Worst-case distributional shift.



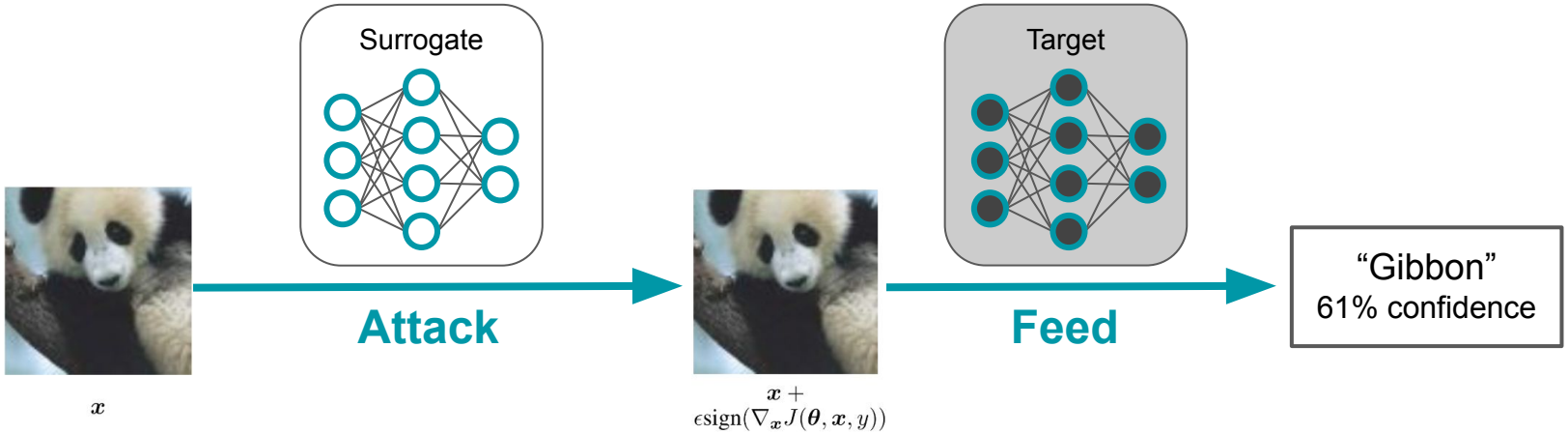
Goodfellow, I. et al. (2014). Explaining and Harnessing Adversarial Examples

Background

Transferability

An adversarial example against a model is likely to be also adversarial against another model.

Black-box attack



Motivation

Random directions in the **weight space** increase transferability.

$$\nabla_x \mathcal{L}(x'_k;\, y, w_0 + e_k) \text{ with } e_k \sim \mathcal{N}(\mathbf{0},\, \sigma^2 I_p)$$

DNN weights — Gaussian noise on weights

Table 1: Success rates of random directions (RD) in the weight and feature spaces under the L∞ attack. In %.

	Target							
Surrogate	RN50	RN152	RNX50	WRN50	DN201	VGG19	IncV1	IncV3
1 DNN (baseline)	45.3±2.4	29.6±0.9	28.8±0.2	31.5±1.6	17.5±0.6	16.6±0.9	10.4±0.5	5.3±1.0
+ **RD Weights**	**60.6**±1.5	**40.5**±3.0	**39.9**±0.2	**44.4**±3.2	**22.9**±0.8	**22.7**±0.5	**13.9**±0.2	**6.6**±0.7
+ RD Features	46.4±1.8	29.0±2.2	28.7±1.2	32.7±1.5	17.5±0.6	17.5±0.6	10.3±0.7	5.6±0.7

Random directions in the feature space do **not**.

$$\nabla_x \mathcal{L}(x'_k;\, y, w_0) + e'_k \text{ with } e'_k \sim \mathcal{N}(\mathbf{0},\, \sigma'^2 I_d)$$

Motivation

Random directions in the **weight space** increase transferability.

$$\nabla_x \mathcal{L}(x'_k;\, y, w_0 + e_k) \text{ with } e_k \sim \mathcal{N}(\mathbf{0},\, \sigma^2 I_p)$$

Equivalent to adding **feature** noise structured by local variations of input gradients in the weight space:

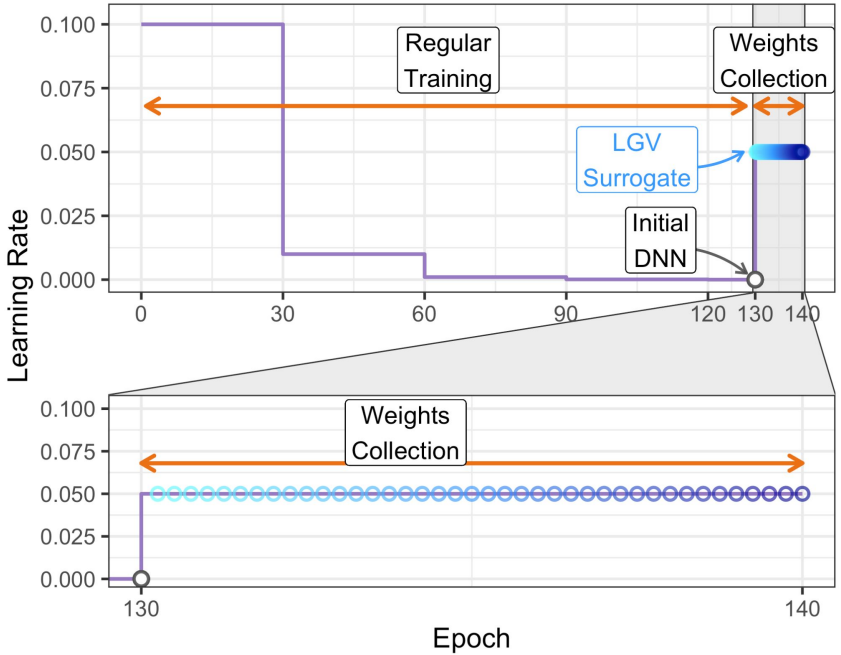
$$\mathcal{N}\left(\nabla_x \mathcal{L}(x'_k;\, y, w_0),\ \sigma^2\, \mathbf{J}_{\nabla_x \mathcal{L}(x'_k;\, y, \cdot)}(w_0)\, \mathbf{J}_{\nabla_x \mathcal{L}(x'_k;\, y, \cdot)}(w_0)^T\right)$$

The vicinity
in the weight space
is relevant for
transferability.

LGV: phase 1

Model Collection

Collect models during a few epochs with a **high learning rate**



Algorithm 1 LGV Weights Collection

Input: n_{epochs} number of epochs, K number of weights, η learning rate, γ momentum, w_0 pretrained weights, $\mathcal{D}$ training dataset

Output: $(w_1, \ldots, w_K)$ LGV weights

1: $w \leftarrow w_0$ ▷ Start from a regularly trained DNN
2: **for** $i \leftarrow 1$ **to** K **do**
3: $w \leftarrow \text{SGD}(w, \eta, \gamma, \mathcal{D}, \frac{n_{\text{epochs}}}{K})$ ▷ Perform $\frac{n_{\text{epochs}}}{K}$ of an epoch of SGD with η learning rate and γ momentum on $\mathcal{D}$
4: $w_i \leftarrow w$
5: **end for**

LGV: phase 2

Adversarial example crafting

Apply classical attack on one collected model per iteration

Algorithm 2 I-FGSM Attack on LGV

Input: (x, y) natural example, $(w_1, \ldots, w_K)$ LGV weights, n_{iter} number of iterations, ε p-norm perturbation, α step-size

Output: x_{adv} adversarial example

Shuffle $(w_1, \ldots, w_K)$ ▷ Shuffle weights

$x_{\text{adv}} \leftarrow x$

for $i \leftarrow 1$ **to** n_{iter} **do**

    $x_{\text{adv}} \leftarrow x_{\text{adv}} + \alpha \nabla_x \mathcal{L}(x_{\text{adv}}; y, w_{i \bmod K})$ ▷ Compute the input gradient of the loss of a randomly picked LGV model

    $x_{\text{adv}} \leftarrow \text{project}(x_{\text{adv}}, B_\varepsilon[x])$ ▷ Project in the p-norm ball centred on x of ε radius

    $x_{\text{adv}} \leftarrow \text{clip}(x_{\text{adv}}, 0, 1)$ ▷ Clip to pixel range values

end for

Evaluation

LGV alone beats all (combinations of) four state-of-the-art techniques.

Table 1: Success rates of state-of-the-art and LGV under the L∞ attack. Underline is best. "RD" stands for random directions in the weight space. In %.

	Target							
Surrogate	RN50	RN152	RNX50	WRN50	DN201	VGG19	IncV1	IncV3
Baselines (1 DNN)								
1 DNN	45.3±2.4	29.6±0.9	28.8±0.2	31.5±1.6	17.5±0.6	16.6±0.9	10.4±0.5	5.3±1.0
MI	53.0±2.2	36.3±1.5	34.7±0.4	38.1±2.0	22.0±0.1	21.1±0.3	13.9±0.4	7.3±0.8
GN	63.9±2.4	43.8±2.4	43.3±1.3	47.4±0.9	24.8±0.3	24.1±1.0	14.6±0.3	6.8±1.2
GN+MI	68.4±2.3	49.3±2.5	47.9±1.2	52.1±1.7	28.4±0.8	28.0±0.7	17.5±0.5	8.7±0.5
DI	75.0±0.2	56.4±1.9	59.6±1.5	61.6±2.4	41.6±1.1	39.7±0.9	27.7±1.0	15.2±1.0
DI+MI	81.2±0.3	63.8±1.9	67.6±0.9	68.9±1.5	49.3±0.7	46.7±0.4	33.0±1.0	19.4±0.9
SGM	64.4±0.8	49.1±3.1	48.9±0.6	51.7±2.8	30.7±0.9	33.6±1.3	22.5±1.5	10.7±0.9
SGM+MI	66.0±0.6	51.3±3.5	50.9±0.9	54.3±2.3	32.5±1.3	35.8±0.7	24.1±1.0	12.1±1.2
SGM+DI	76.8±0.5	62.3±2.7	63.6±1.7	65.3±1.4	45.5±0.9	49.9±0.8	36.0±0.7	19.2±1.7
SGM+DI+MI	80.9±0.7	66.9±2.5	68.7±1.2	70.0±1.7	50.9±0.6	56.0±1.4	42.1±1.4	23.6±1.6
Our techniques								
RD	60.6±1.5	40.5±3.0	39.9±0.2	44.4±3.2	22.9±0.8	22.7±0.5	13.9±0.2	6.6±0.7
LGV-SWA	84.9±1.2	63.9±3.7	62.1±0.4	61.1±2.9	44.2±0.4	42.4±1.3	31.5±0.8	12.2±0.8
LGV-SWA+RD	90.2±0.5	71.7±3.4	69.9±1.2	69.1±3.3	49.9±1.0	47.4±2.0	34.9±0.3	13.5±0.9
LGV (ours)	95.4±0.1	85.5±2.3	83.7±1.2	82.1±2.4	69.3±1.0	67.8±1.2	58.1±0.8	25.3±1.9

Usage

Implemented in the *torchattacks* library

```
[ ]  from torchattacks import LGV, BIM

     atk = LGV(base_model, trainloader, lr=0.05, epochs=10, nb_models_epoch=4,
               wd=1e-4, attack_class=BIM, eps=4/255, alpha=4/255/10, steps=50)
     # atk.load_models(list_models)  # uncomment to load our publicly available pretrained models
     atk.collect_models()  # or collect models yourself
     atk.save_models('models/lgv')

     report_success_rate(atk)
```

```
     Phase 2: craft adversarial examples with BIM
     Success rate of LGV-BIM: 97.6%
```

Possible to combine with any other attack
Demo notebook available

How to explain the success of LGV?

RQ Why do weights from a vicinity help to attack a model from another vicinity?

2 keys:

1. LGV produces **flatter** adversarial examples.
2. The LGV subspace embeds **geometric properties** relevant for transferability.

I - The surrogate-target misalignment hypothesis

Background about flatness for (natural) generalization

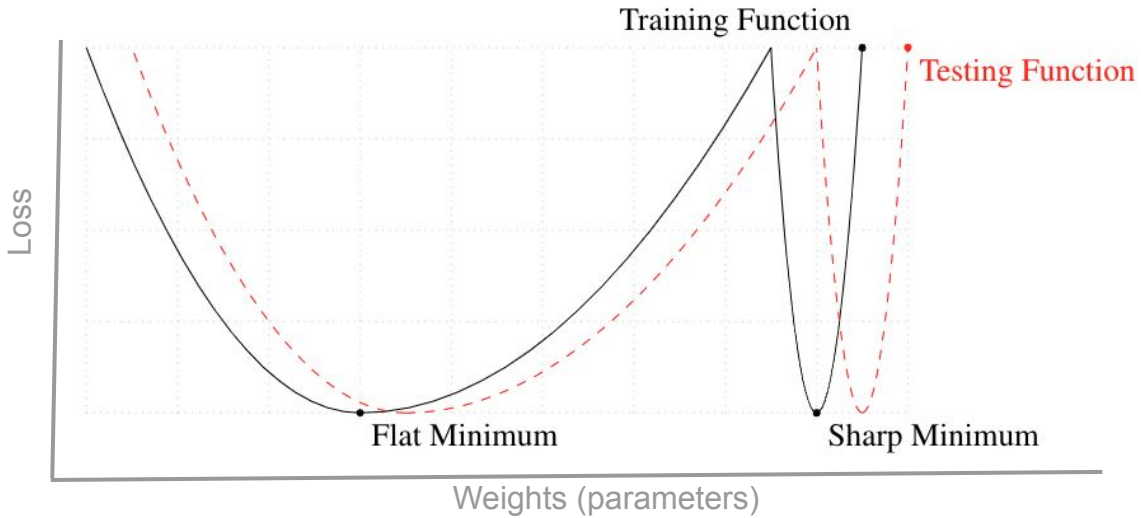
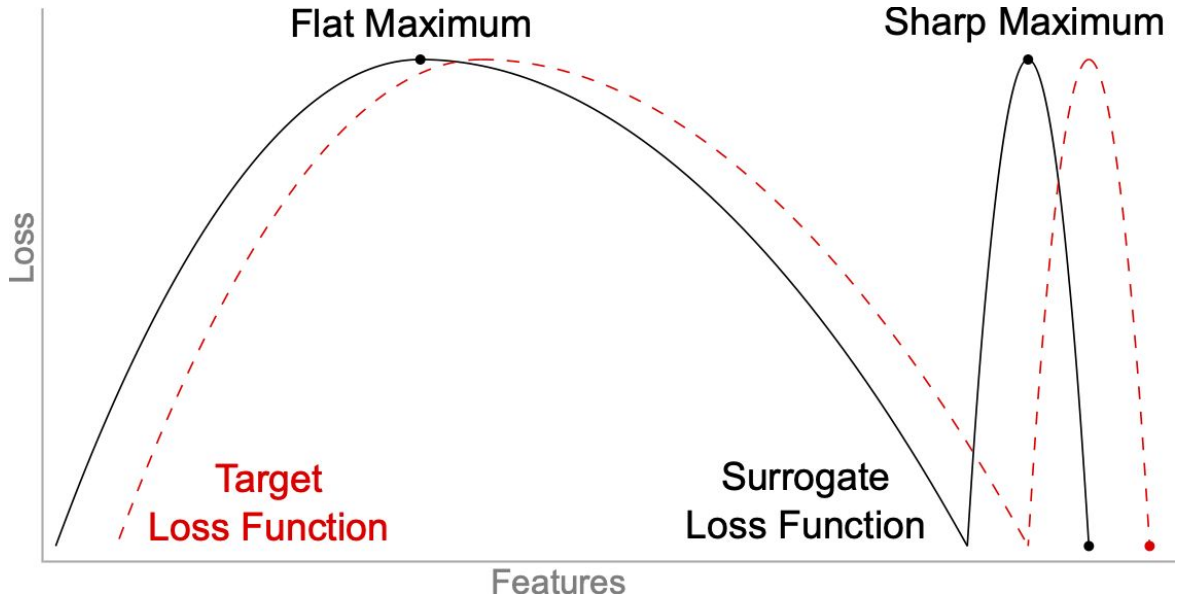


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

Keskar, N. S., et al. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR 2017*

I - The surrogate-target misalignment hypothesis

Flatter adversarial examples may be more robust to misalignment between surrogate and target.



LGV → Flatness in the **weight** space

LGV collects models in flatter regions of the weight space…

Hessian-based sharpness metrics

Table 2: Sharpness metrics in the weight space, i.e., the largest eigenvalue and the rank of the Hessian, computed on three types of surrogate and 10,000 training examples.

	Hessian	
Model	Max EV	Trace
1 DNN	558 ±57	16258 ±725
LGV indiv.	168 ±127	4295 ±517
LGV-SWA	30 ±1	1837 ±70

Random rays

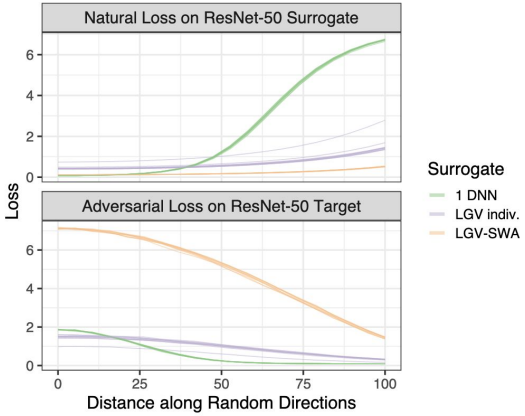


Fig. 3: L_∞ attack crafted on surrogate with natural loss (*up*), evaluated on target (*down*) with respect to the 2-norm distance along 10 random directions in the weight space from the LGV-SWA solution (*orange*), random LGV weights (*purple*), and the initial DNN (*green*).

Interpolation in the weight space

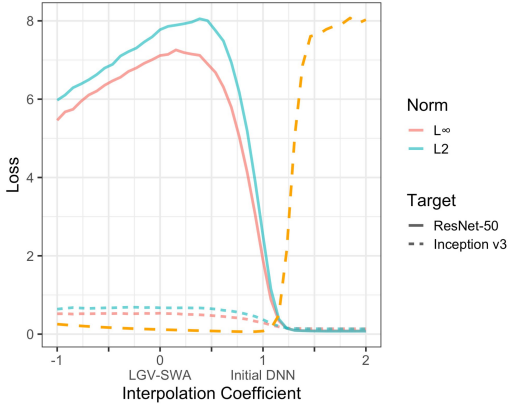


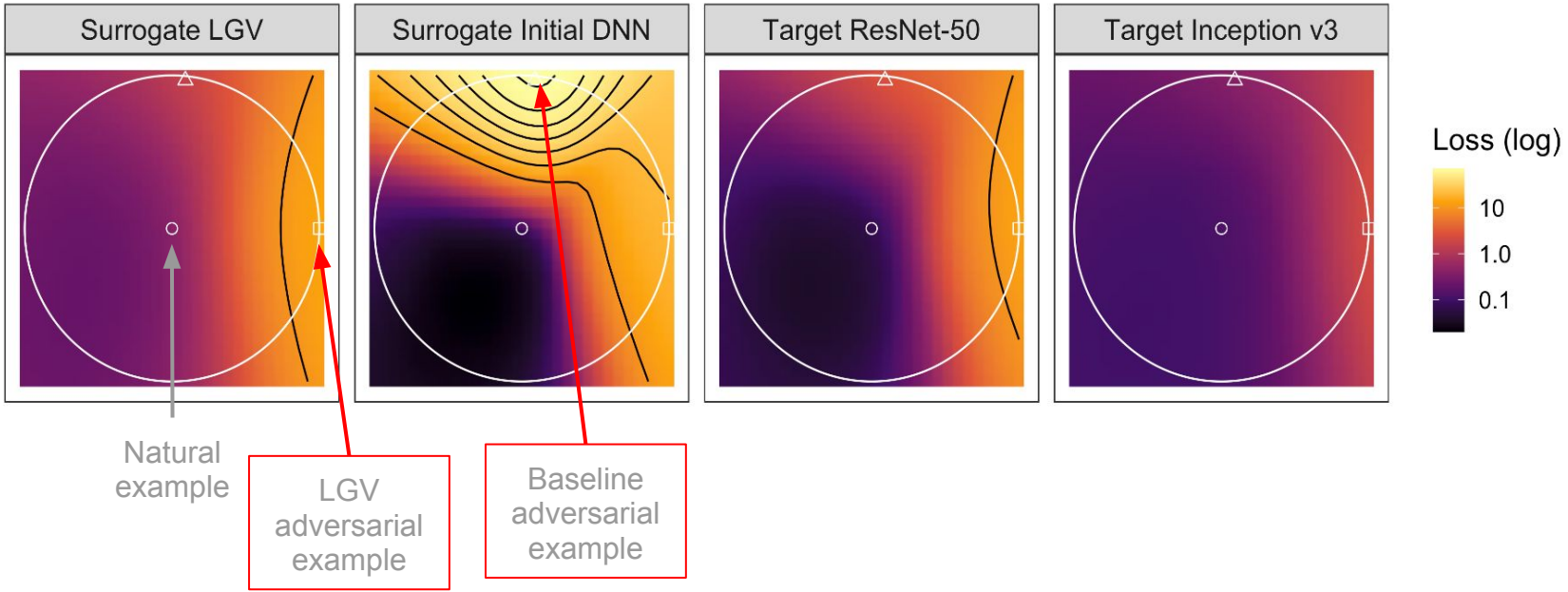
Fig. 9: Adversarial target loss (*plain*) and surrogate natural loss (*orange dashed*) with respect to the interpolation coefficient α between the LGV-SWA solution and the initial model.

Among others

Optimal lengths of random and LGV deviations vectors

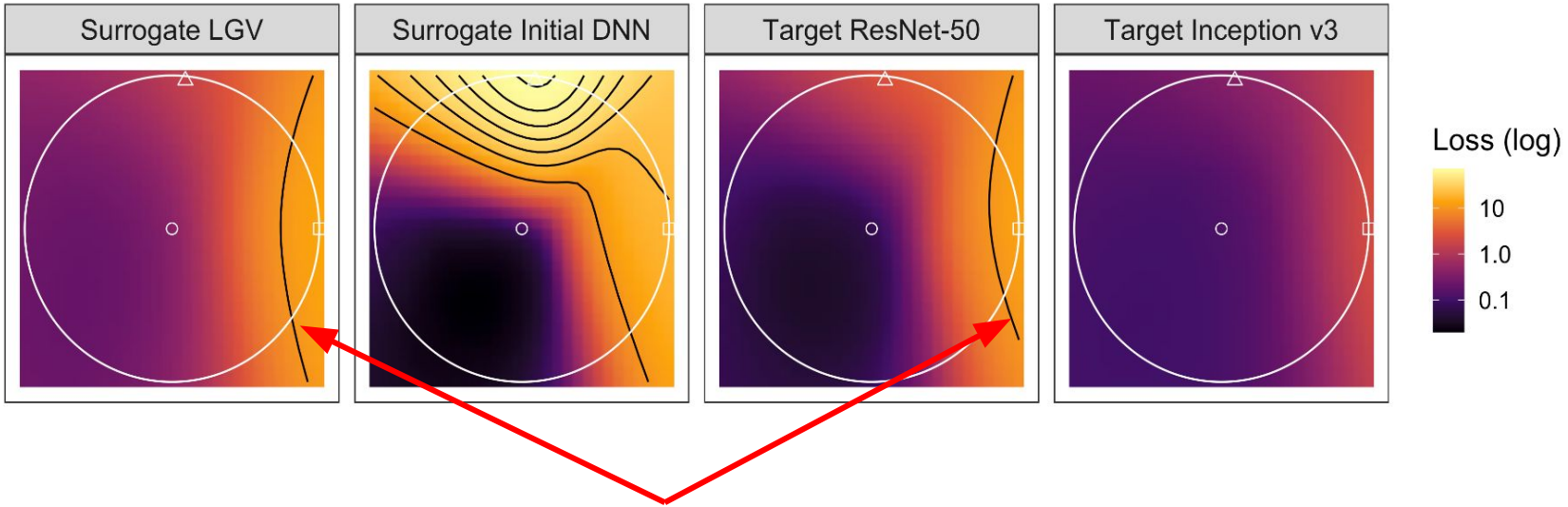
LGV → Flatness in the **feature** space

…as a result, LGV produces adversarial examples flatter in the feature space.



The surrogate-target (mis)alignment

LGV appears particularly well aligned with the target

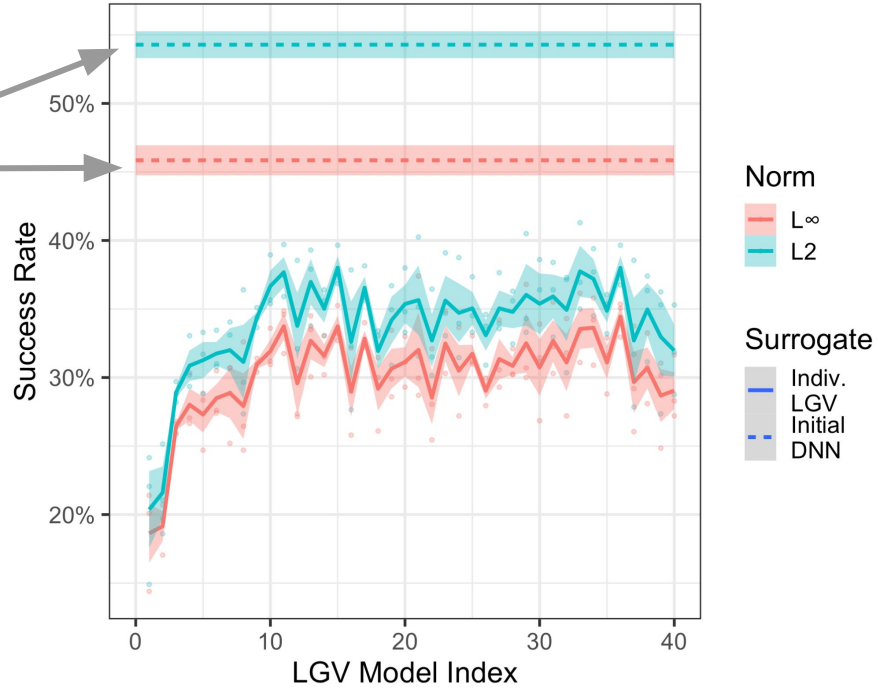


Loss contours have similar shape which appear shifted

Flatness is not enough

However individual LGV models do not succeed on their own…

1 DNN baseline



II - LGV subspace properties

We consider the weight subspace defined by deviations of LGV weights from their average,

$$\mathcal{S} = \{w \,|\, w = w_{\mathrm{SWA}} + \mathbf{P}z\}\,, \tag{4}$$

where w_{SWA} is called the shift vector, $\mathbf{P} = (w_1 - w_{\mathrm{SWA}}, \ldots, w_K - w_{\mathrm{SWA}})^\intercal$ is the projection matrix of LGV weights deviations from their mean, and $z \in \mathbb{R}^K$.

The subspace $\mathcal{S}$ is:

1. Densely related to transferability, i.e., **useful,**
2. Composed of directions whose relative importance correlates with geometrical properties, i.e., **its geometry is relevant,**
3. Useful when shifted to other solutions, i.e., **its geometry captures generic properties.**

Background - SGD Subspace

Despite the high dimensionality of the weight space, SGD updates are concentrated in a tiny subspace

---

GRADIENT DESCENT HAPPENS IN A TINY SUBSPACE

Guy Gur-Ari[*]
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ 08540, USA
guyg@ias.edu

Daniel A. Roberts[*]
Facebook AI Research
New York, NY 10003, USA
danr@fb.com

Ethan Dyer
Johns Hopkins University
Baltimore, MD 21218, USA
edyer4@jhu.edu

ABSTRACT

We show that in a variety of large-scale deep learning scenarios the gradient dynamically converges to a very small subspace after a short period of training. The subspace is spanned by a few top eigenvectors of the Hessian (equal to the number of classes in the dataset), and is mostly preserved over long periods of training. A simple argument then suggests that gradient descent may happen mostly in this subspace. We give an example of this effect in a solvable model of classification, and we comment on possible implications for optimization and learning.

A) A subspace **useful** for transferability

The LGV subspace is significantly better than a random subspace
→ Specific relation to transferability.

Random directions in full weight space vs. LGV

Table 9: Transfer success rate of random directions sampled in LGV deviations subspace.

Norm	Surrogate	Target							
		RN50	RN152	RNX50	WRN50	DN201	VGG19	IncV1	IncV3
L∞	LGV	95.5±0.1	85.5±2.1	83.6±1.1	82.2±2.4	69.6±1.0	67.8±0.9	58.4±0.6	25.6±1.7
L∞	LGV-SWA + RD in $\mathcal{S}$	96.0±0.2	85.6±2.5	83.6±0.6	82.1±2.8	68.6±1.1	65.7±1.5	54.5±0.9	23.5±0.4
L∞	LGV-SWA + RD	90.4±0.3	71.9±3.4	70.0±1.2	69.2±3.4	50.0±1.0	47.4±1.9	34.9±0.4	13.4±0.7
L2	LGV	96.3±0.1	90.1±1.0	88.8±0.4	87.5±1.6	79.8±1.1	78.1±1.6	71.9±0.6	43.1±0.6
L2	LGV-SWA + RD in $\mathcal{S}$	96.6±0.3	90.1±1.4	88.7±0.5	87.3±2.0	77.6±1.0	75.6±1.5	67.4±1.9	37.4±0.4
L2	LGV-SWA + RD	91.9±0.6	78.2±2.9	76.2±1.3	75.4±2.5	58.1±0.3	55.8±1.6	42.7±0.6	20.0±0.6

A) A subspace **useful** for transferability

Sampling random directions in the subspace have results close to LGV.

→ Densely related to transferability

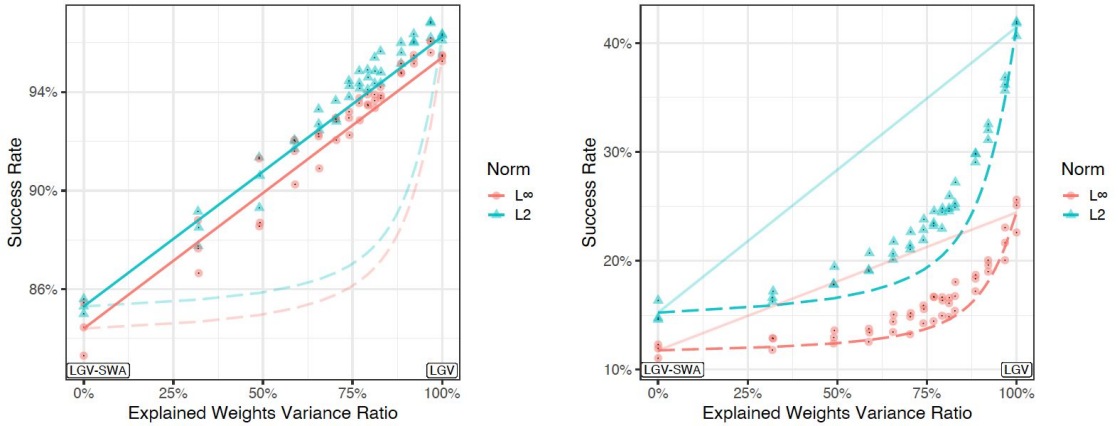
Random directions in LGV subspace vs. LGV

Table 9: Transfer success rate of random directions sampled in LGV deviations subspace.

Norm	Surrogate	Target							
		RN50	RN152	RNX50	WRN50	DN201	VGG19	IncV1	IncV3
L∞	LGV	$95.5_{\pm 0.1}$	$85.5_{\pm 2.1}$	$83.6_{\pm 1.1}$	$82.2_{\pm 2.4}$	$69.6_{\pm 1.0}$	$67.8_{\pm 0.9}$	$58.4_{\pm 0.6}$	$25.6_{\pm 1.7}$
L∞	LGV-SWA + RD in $\mathcal{S}$	$96.0_{\pm 0.2}$	$85.6_{\pm 2.5}$	$83.6_{\pm 0.6}$	$82.1_{\pm 2.8}$	$68.6_{\pm 1.1}$	$65.7_{\pm 1.5}$	$54.5_{\pm 0.9}$	$23.5_{\pm 0.4}$
L∞	LGV-SWA + RD	$90.4_{\pm 0.3}$	$71.9_{\pm 3.4}$	$70.0_{\pm 1.2}$	$69.2_{\pm 3.4}$	$50.0_{\pm 1.0}$	$47.4_{\pm 1.9}$	$34.9_{\pm 0.4}$	$13.4_{\pm 0.7}$
L2	LGV	$96.3_{\pm 0.1}$	$90.1_{\pm 1.0}$	$88.8_{\pm 0.4}$	$87.5_{\pm 1.6}$	$79.8_{\pm 1.1}$	$78.1_{\pm 1.6}$	$71.9_{\pm 0.6}$	$43.1_{\pm 0.6}$
L2	LGV-SWA + RD in $\mathcal{S}$	$96.6_{\pm 0.3}$	$90.1_{\pm 1.4}$	$88.7_{\pm 0.5}$	$87.3_{\pm 2.0}$	$77.6_{\pm 1.0}$	$75.6_{\pm 1.5}$	$67.4_{\pm 1.9}$	$37.4_{\pm 0.4}$
L2	LGV-SWA + RD	$91.9_{\pm 0.6}$	$78.2_{\pm 2.9}$	$76.2_{\pm 1.3}$	$75.4_{\pm 2.5}$	$58.1_{\pm 0.3}$	$55.8_{\pm 1.6}$	$42.7_{\pm 0.6}$	$20.0_{\pm 0.6}$

B) Relevance of **Geometry**

The subspace is composed of directions whose relative importance depends on the functional similarity between surrogate and target.



(a) ResNet-50 Target

(b) Inception v3 Target

Fig. 5: Success rate of the LGV surrogate projected on an increasing number of dimensions with the corresponding ratio of explained variance in the weight space. Hypothetical average cases of proportionality to variance (*solid*) and equal contributions of all subspace dimensions (*dashed*). Scales not shared.

C) **Generic** Geometry Properties

LGV deviations can be shifted in the weight space and significantly outperform random directions.

Table 10: Transfer success rate of LGV deviations shifted to other independent solutions, for target architectures in the ResNet family.

Norm	Surrogate	Target			
		RN50	RN152	RNX50	WRN50
L∞	LGV-SWA + (LGV' - LGV-SWA')	$94.3_{\pm 0.5}$	$81.5_{\pm 2.3}$	$79.1_{\pm 1.4}$	$78.1_{\pm 2.4}$
L∞	LGV-SWA + RD	$90.4_{\pm 0.3}$	$71.9_{\pm 3.4}$	$70.0_{\pm 1.2}$	$69.2_{\pm 3.4}$
L∞	LGV (ours)	$95.4_{\pm 0.1}$	$85.3_{\pm 2.1}$	$83.7_{\pm 1.1}$	$82.1_{\pm 2.5}$
L∞	1 DNN + γ (LGV' - LGV-SWA')	$73.3_{\pm 2.0}$	$52.8_{\pm 2.9}$	$52.6_{\pm 1.6}$	$56.6_{\pm 2.8}$
L∞	1 DNN + RD	$60.8_{\pm 1.6}$	$40.8_{\pm 2.7}$	$40.2_{\pm 0.3}$	$44.8_{\pm 2.7}$
L2	LGV-SWA + (LGV' - LGV-SWA')	$95.2_{\pm 0.5}$	$86.1_{\pm 1.9}$	$84.2_{\pm 1.0}$	$82.7_{\pm 1.6}$
L2	LGV-SWA + RD	$92.0_{\pm 0.5}$	$77.9_{\pm 3.0}$	$76.2_{\pm 1.4}$	$75.2_{\pm 2.8}$
L2	LGV (ours)	$96.3_{\pm 0.1}$	$90.2_{\pm 1.1}$	$88.6_{\pm 0.6}$	$87.6_{\pm 1.7}$
L2	1 DNN + γ (LGV' - LGV-SWA')	$84.2_{\pm 0.8}$	$68.7_{\pm 2.6}$	$70.0_{\pm 1.3}$	$72.4_{\pm 1.5}$
L2	1 DNN + RD	$74.6_{\pm 0.5}$	$55.8_{\pm 3.1}$	$56.1_{\pm 0.6}$	$59.9_{\pm 3.2}$

Conclusion

LGV is simple yet effective to enhance black-box attack.

Overall, the improved transferability of LGV comes from the **geometry** of the subspace formed by LGV weights in a **flatter** region of the loss.

📄 Paper

🐍 Code

🪧 Poster