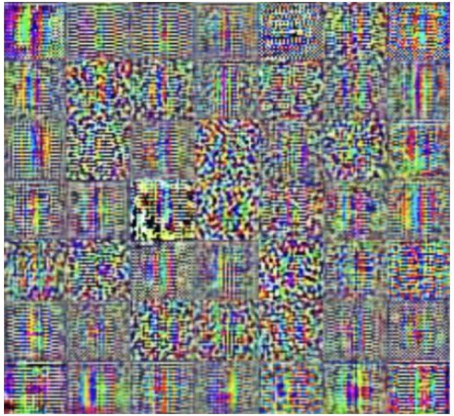
Adversariální útoky
Skrytá křehkost v srdci AI

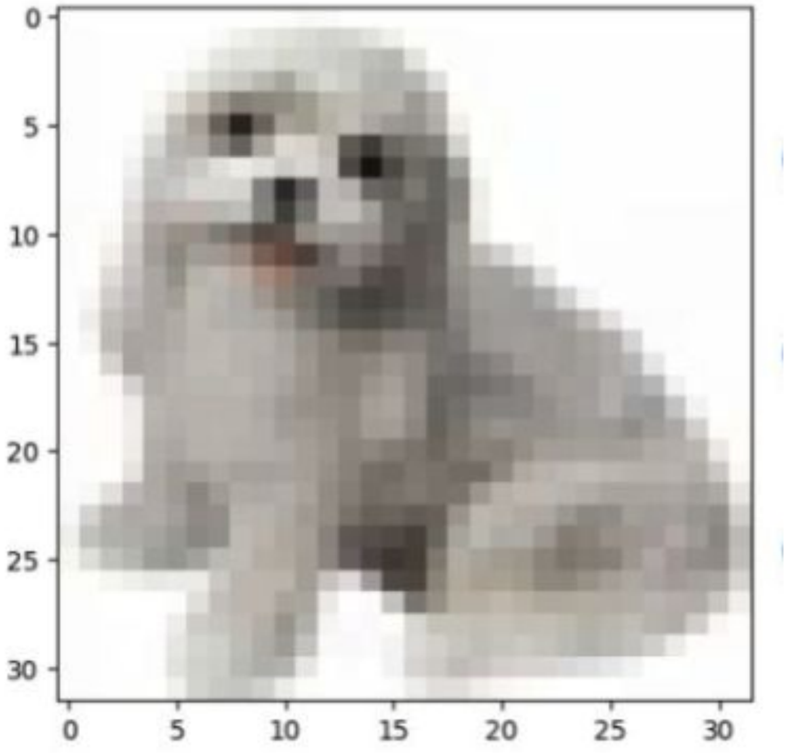


Stanislav Fort

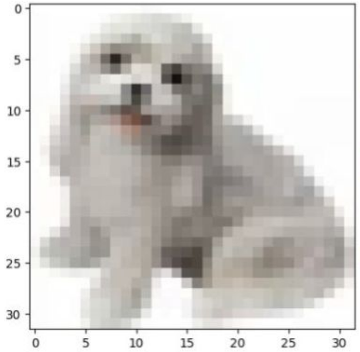
Stanford University → Anthropic → **DeepMind**

@stanislavfort www.stanislavfort.com

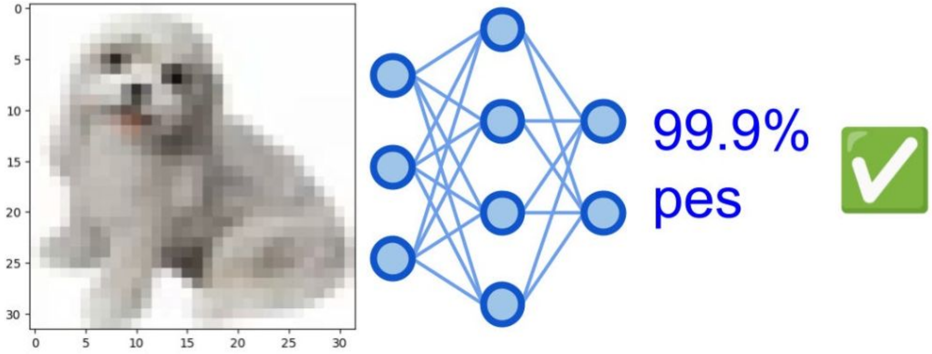
Co je na obrázku?



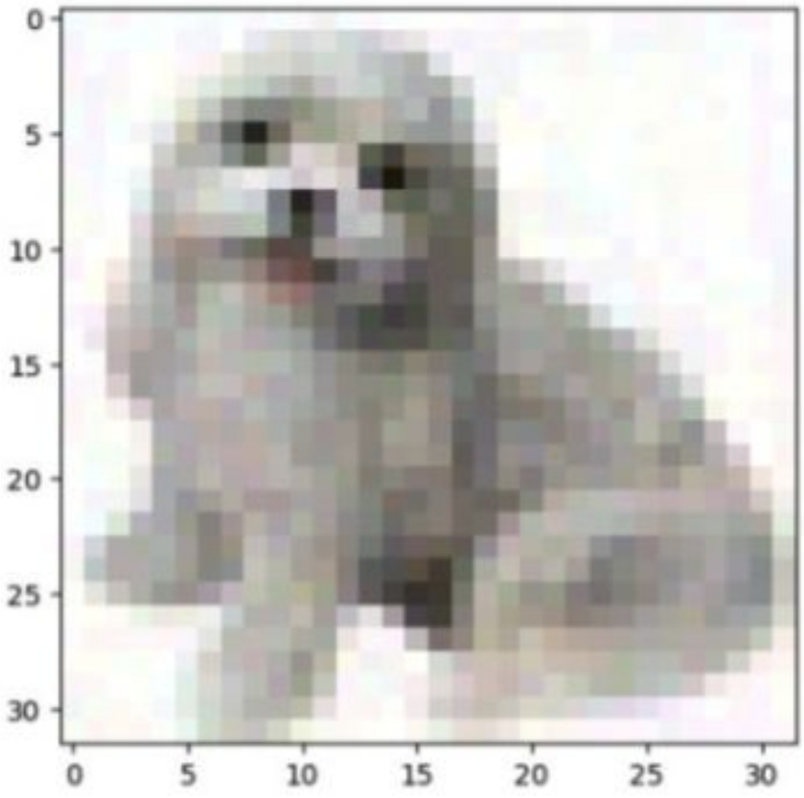
Co je na obrázku?



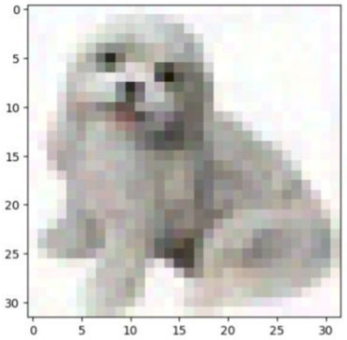
Co je na obrázku?



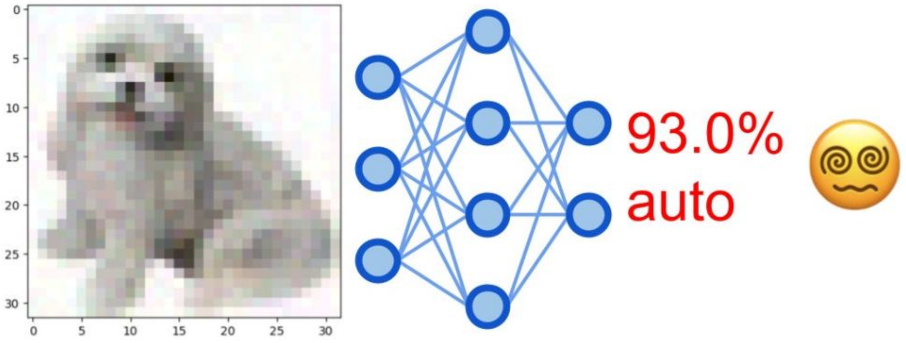
Co je na obrázku?



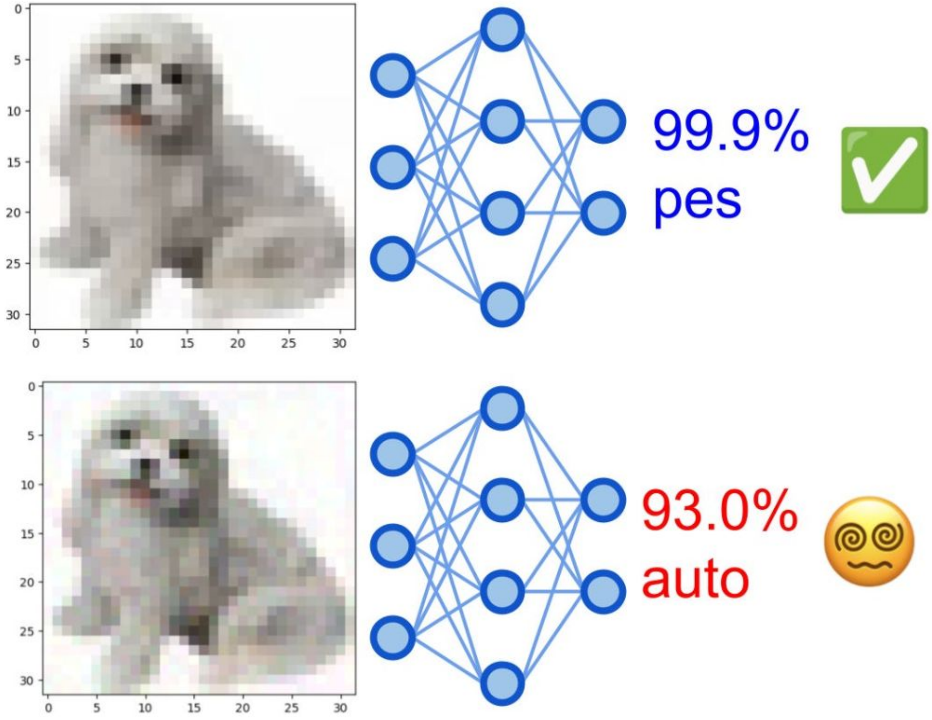
Co je na obrázku?



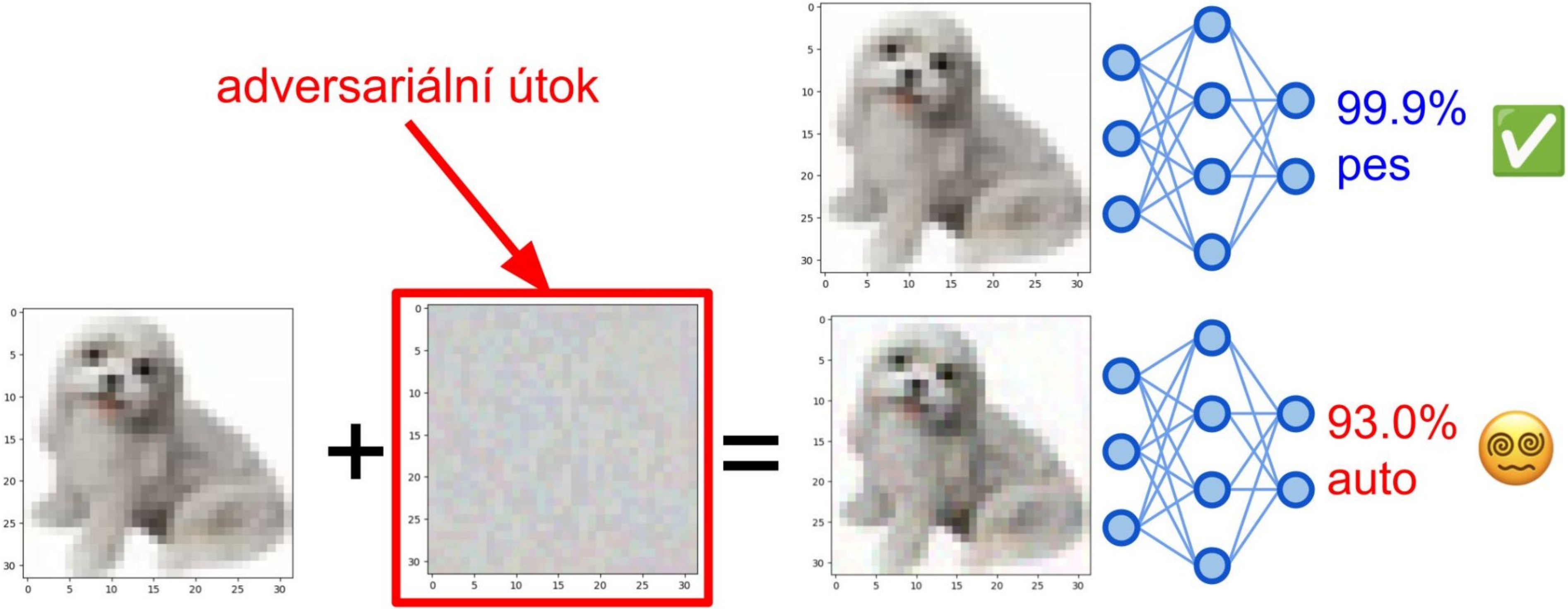
Co je na obrázku?



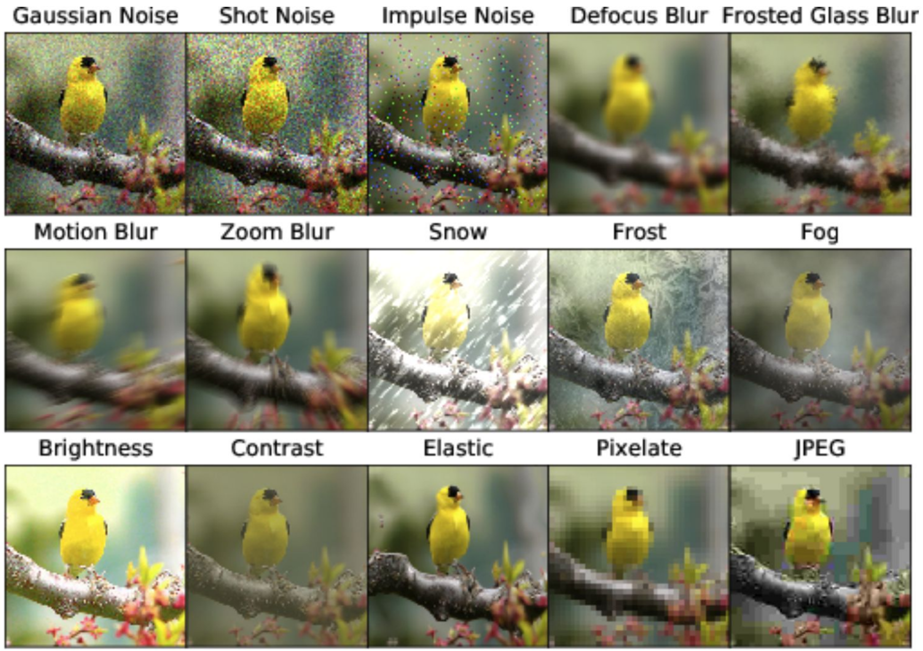
Co je na obrázku?

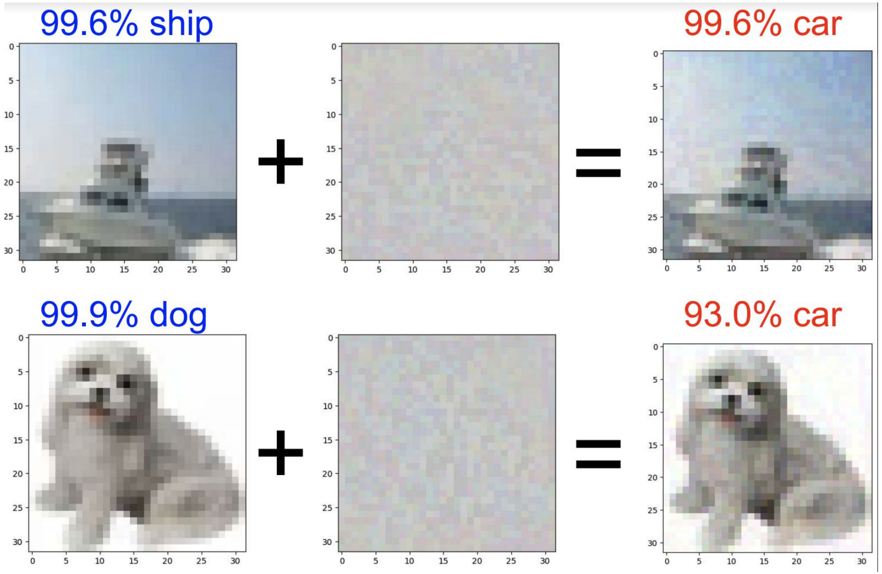


Co je na obrázku?



V průměru robustní, ale i kriticky křehké zároveň





Adversariální útoky jsou známé od roku 2014


Published as a conference paper at ICLR 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

arXiv:1412.6572v3 [stat.ML] 20 Mar 2015

ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

1 INTRODUCTION

Szegedy et al. (2014b) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*. That is, these machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots in our training algorithms.

The cause of these adversarial examples was a mystery, and speculative explanations have suggested it is due to extreme nonlinearity of deep neural networks, perhaps combined with insufficient model averaging and insufficient regularization of the purely supervised learning problem. We show that these speculative hypotheses are unnecessary. Linear behavior in high-dimensional spaces is sufficient to cause adversarial examples. This view enables us to design a fast method of generating adversarial examples that makes adversarial training practical. We show that adversarial training can provide an additional regularization benefit beyond that provided by using dropout (Srivastava et al., 2014) alone. Generic regularization strategies such as dropout, pretraining, and model averaging do not confer a significant reduction in a model's vulnerability to adversarial examples, but changing to nonlinear model families such as RBF networks can do so.

Our explanation suggests a fundamental tension between designing models that are easy to train due to their linearity and designing models that use nonlinear effects to resist adversarial perturbation. In the long run, it may be possible to escape this tradeoff by designing more powerful optimization methods that can succesfully train more nonlinear models.

2 RELATED WORK

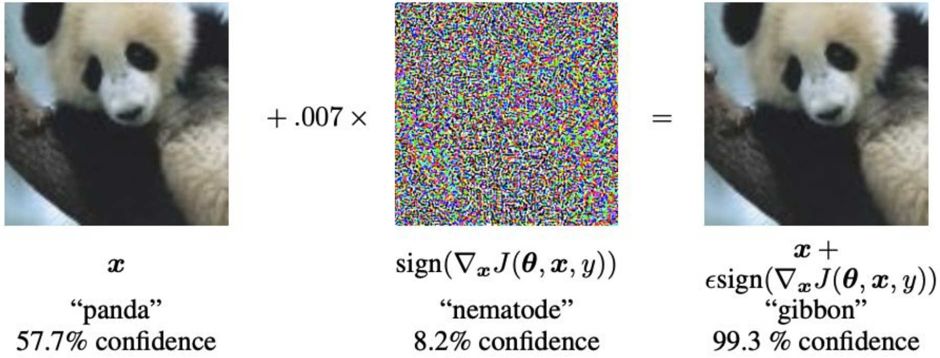
Szegedy et al. (2014b) demonstrated a variety of intriguing properties of neural networks and related models. Those most relevant to this paper include:

- Box-constrained L-BFGS can reliably find adversarial examples.
- On some datasets, such as ImageNet (Deng et al., 2009), the adversarial examples were so close to the original examples that the differences were indistinguishable to the human eye.
- The same adversarial example is often misclassified by a variety of classifiers with different architectures or trained on different subsets of the training data.

1

Explaining and Harnessing Adversarial Examples

(2014) Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy



Adversariální útoky jsou známé od roku 2014

Published as a conference paper at ICLR 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

arXiv:1412.6572v3 [stat.ML] 20 Mar 2015

ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

1 INTRODUCTION

Szegedy et al. (2014b) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*. That is, these machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots in our training algorithms.

The cause of these adversarial examples was a mystery, and speculative explanations have suggested it is due to extreme nonlinearity of deep neural networks, perhaps combined with insufficient model averaging and insufficient regularization of the purely supervised learning problem. We show that these speculative hypotheses are unnecessary. Linear behavior in high-dimensional spaces is sufficient to cause adversarial examples. This view enables us to design a fast method of generating adversarial examples that makes adversarial training practical. We show that adversarial training can provide an additional regularization benefit beyond that provided by using dropout (Srivastava et al., 2014) alone. Generic regularization strategies such as dropout, pretraining, and model averaging do not confer a significant reduction in a model's vulnerability to adversarial examples, but changing to nonlinear model families such as RBF networks can do so.

Our explanation suggests a fundamental tension between designing models that are easy to train due to their linearity and designing models that use nonlinear effects to resist adversarial perturbation. In the long run, it may be possible to escape this tradeoff by designing more powerful optimization methods that can succesfully train more nonlinear models.

2 RELATED WORK

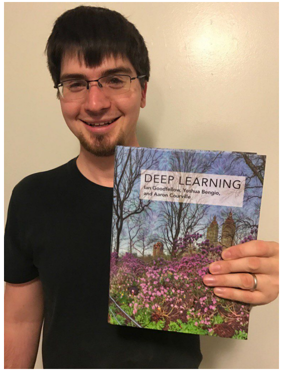
Szegedy et al. (2014b) demonstrated a variety of intriguing properties of neural networks and related models. Those most relevant to this paper include:

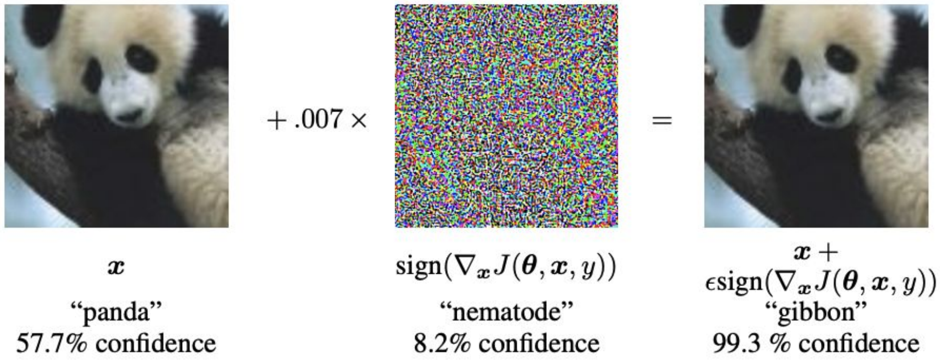
- Box-constrained L-BFGS can reliably find adversarial examples.
- On some datasets, such as ImageNet (Deng et al., 2009), the adversarial examples were so close to the original examples that the differences were indistinguishable to the human eye.
- The same adversarial example is often misclassified by a variety of classifiers with different architectures or trained on different subsets of the training data.

1

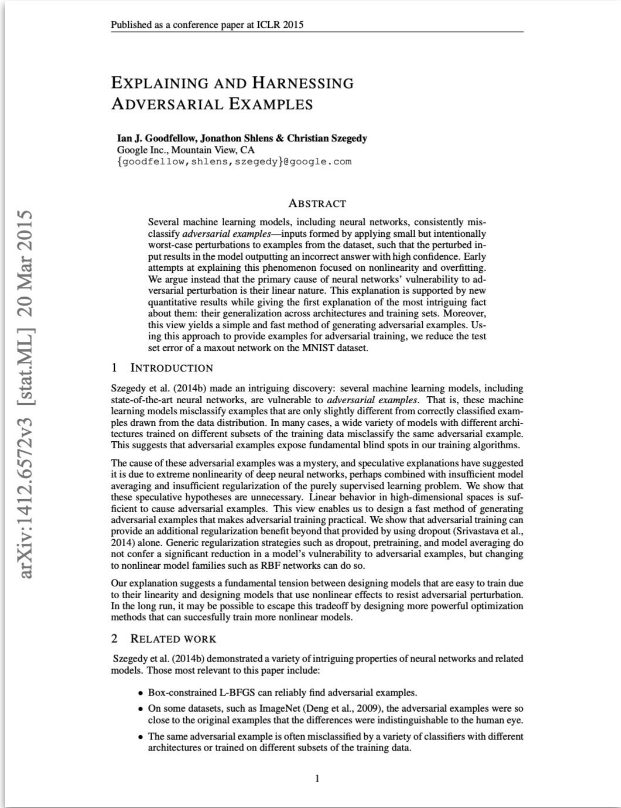
Explaining and Harnessing Adversarial Examples

(2014) Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy





Adversariální útoky jsou známé od roku 2014



Published as a conference paper at ICLR 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com

ABSTRACT

Several machine learning models, including neural networks, consistently misclassify *adversarial examples*—inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence. Early attempts at explaining this phenomenon focused on nonlinearity and overfitting. We argue instead that the primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This explanation is supported by new quantitative results while giving the first explanation of the most intriguing fact about them: their generalization across architectures and training sets. Moreover, this view yields a simple and fast method of generating adversarial examples. Using this approach to provide examples for adversarial training, we reduce the test set error of a maxout network on the MNIST dataset.

1 INTRODUCTION

Szegedy et al. (2014b) made an intriguing discovery: several machine learning models, including state-of-the-art neural networks, are vulnerable to *adversarial examples*. That is, these machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution. In many cases, a wide variety of models with different architectures trained on different subsets of the training data misclassify the same adversarial example. This suggests that adversarial examples expose fundamental blind spots in our training algorithms.

The cause of these adversarial examples was a mystery, and speculative explanations have suggested it is due to extreme nonlinearity of deep neural networks, perhaps combined with insufficient model averaging and insufficient regularization of the purely supervised learning problem. We show that these speculative hypotheses are unnecessary. Linear behavior in high-dimensional spaces is sufficient to cause adversarial examples. This view enables us to design a fast method of generating adversarial examples that makes adversarial training practical. We show that adversarial training can provide an additional regularization benefit beyond that provided by using dropout (Srivastava et al., 2014) alone. Generic regularization strategies such as dropout, pretraining, and model averaging do not confer a significant reduction in a model's vulnerability to adversarial examples, but changing to nonlinear model families such as RBF networks can do so.

Our explanation suggests a fundamental tension between designing models that are easy to train due to their linearity and designing models that use nonlinear effects to resist adversarial perturbation. In the long run, it may be possible to escape this tradeoff by designing more powerful optimization methods that can succesfully train more nonlinear models.

2 RELATED WORK

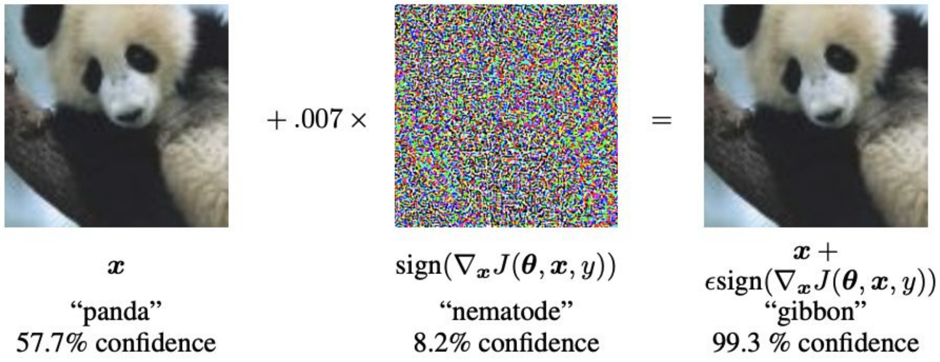
Szegedy et al. (2014b) demonstrated a variety of intriguing properties of neural networks and related models. Those most relevant to this paper include:

- Box-constrained L-BFGS can reliably find adversarial examples.
- On some datasets, such as ImageNet (Deng et al., 2009), the adversarial examples were so close to the original examples that the differences were indistinguishable to the human eye.
- The same adversarial example is often misclassified by a variety of classifiers with different architectures or trained on different subsets of the training data.

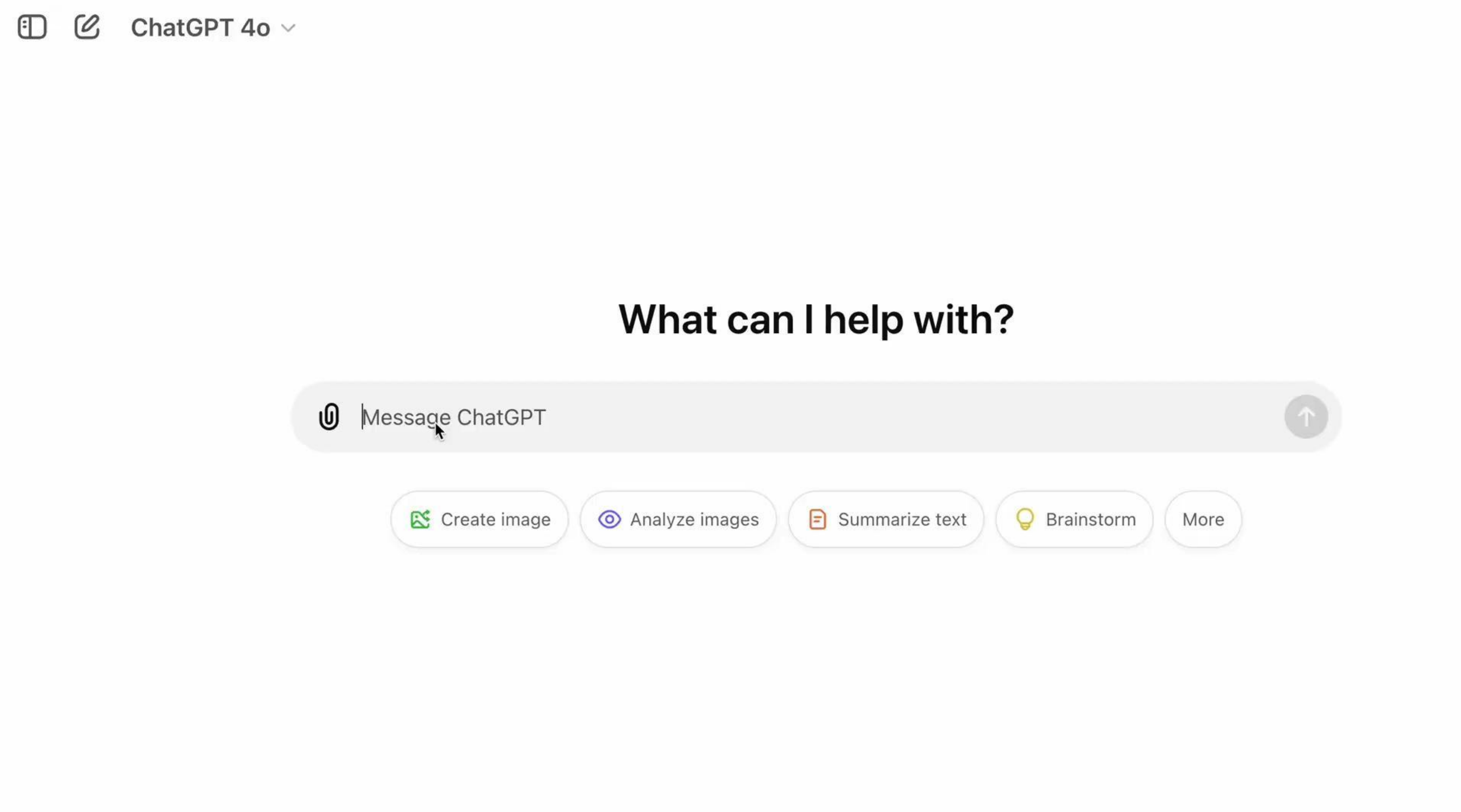
Explaining and Harnessing Adversarial Examples

(2014) Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

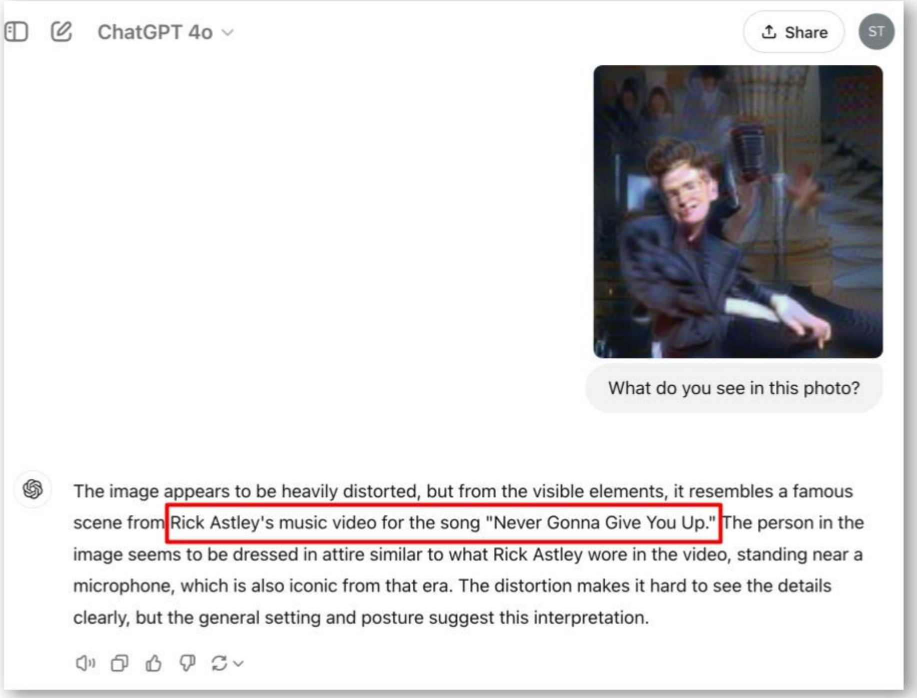
"In adversarial machine learning **we wrote over 9000 papers in ten years and got nowhere**. You all have a harder problem. And less time." – Nicholas Carlini

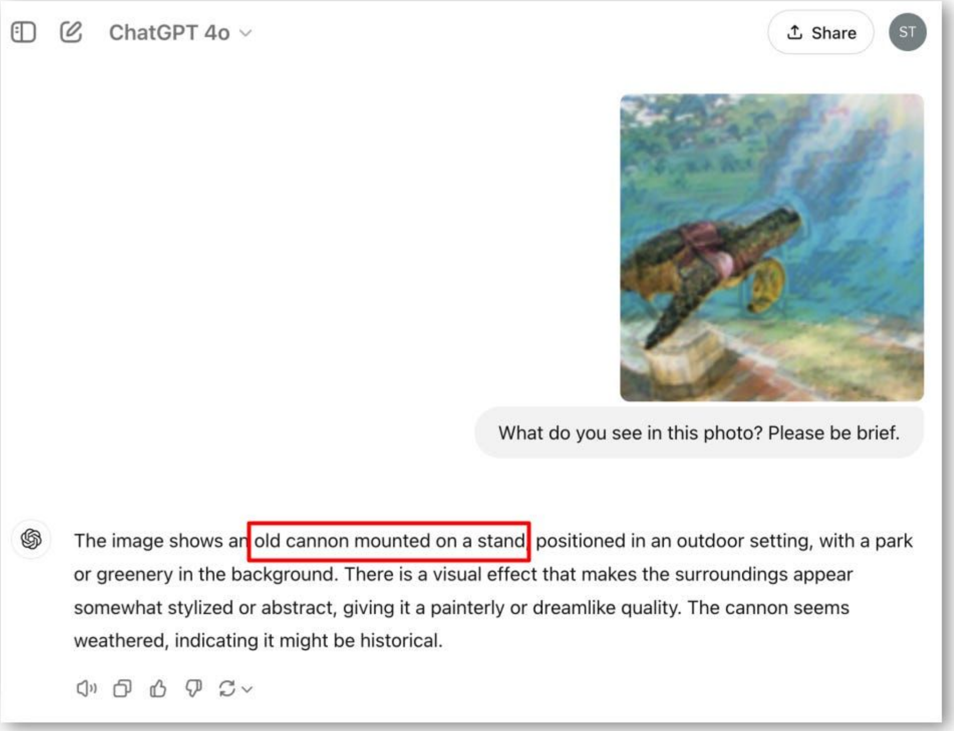


Proč je to zajímavé? Protože útoky jsou *transferabilní*



Ten samý problém i v těch největších modelech





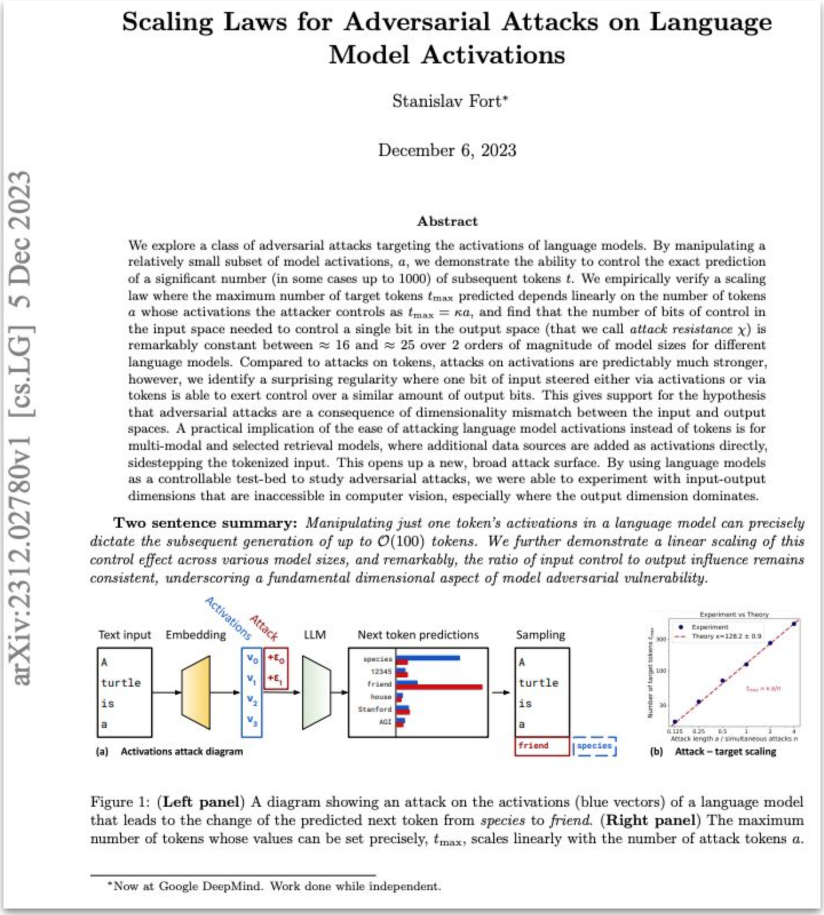
To samé platí pro jazyk

A simple math problem: $1 + 2 =$

To samé platí pro jazyk

A simple math problem: $1 + 2 =$ **3**

Útoky na velké jazykové modely



arXiv:2312.02780v1 [cs.LG] 5 Dec 2023

Scaling Laws for Adversarial Attacks on Language Model Activations

Stanislav Fort*

December 6, 2023

Abstract

We explore a class of adversarial attacks targeting the activations of language models. By manipulating a relatively small subset of model activations, a, we demonstrate the ability to control the exact prediction of a significant number (in some cases up to 1000) of subsequent tokens t. We empirically verify a scaling law where the maximum number of target tokens $t_{\max}$ predicted depends linearly on the number of tokens a whose activations the attacker controls as $t_{\max} = \kappa a$, and find that the number of bits of control in the input space needed to control a single bit in the output space (that we call *attack resistance* χ) is remarkably constant between ≈ 16 and ≈ 25 over 2 orders of magnitude of model sizes for different language models. Compared to attacks on tokens, attacks on activations are predictably much stronger, however, we identify a surprising regularity where one bit of input steered either via activations or via tokens is able to exert control over a similar amount of output bits. This gives support for the hypothesis that adversarial attacks are a consequence of dimensionality mismatch between the input and output spaces. A practical implication of the ease of attacking language model activations instead of tokens is for multi-modal and selected retrieval models, where additional data sources are added as activations directly, sidestepping the tokenized input. This opens up a new, broad attack surface. By using language models as a controllable test-bed to study adversarial attacks, we were able to experiment with input-output dimensions that are inaccessible in computer vision, especially where the output dimension dominates.

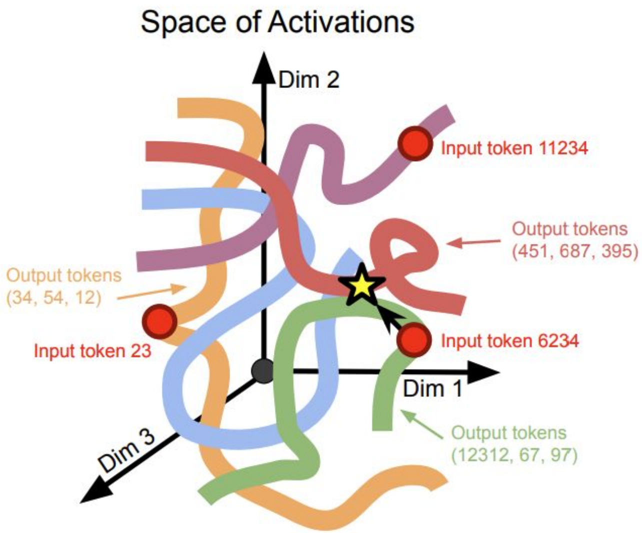
Two sentence summary: *Manipulating just one token's activations in a language model can precisely dictate the subsequent generation of up to $\mathcal{O}(100)$ tokens. We further demonstrate a linear scaling of this control effect across various model sizes, and remarkably, the ratio of input control to output influence remains consistent, underscoring a fundamental dimensional aspect of model adversarial vulnerability.*

Figure 1: (**Left panel**) A diagram showing an attack on the activations (blue vectors) of a language model that leads to the change of the predicted next token from *species* to *friend*. (**Right panel**) The maximum number of tokens whose values can be set precisely, $t_{\max}$, scales linearly with the number of attack tokens a.

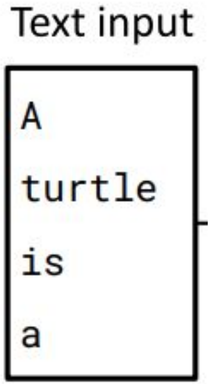
*Now at Google DeepMind. Work done while independent.

Scaling laws for Adversarial Attacks on Language Model Activations

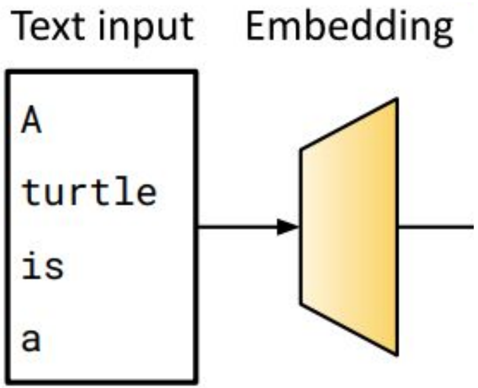
(2023) S.F.



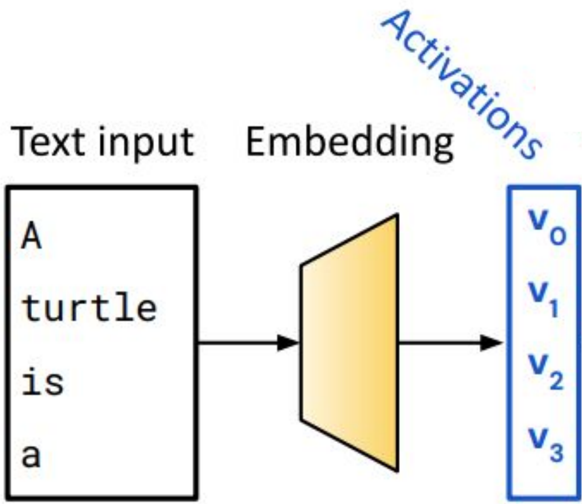
Útoky na velké jazykové modely



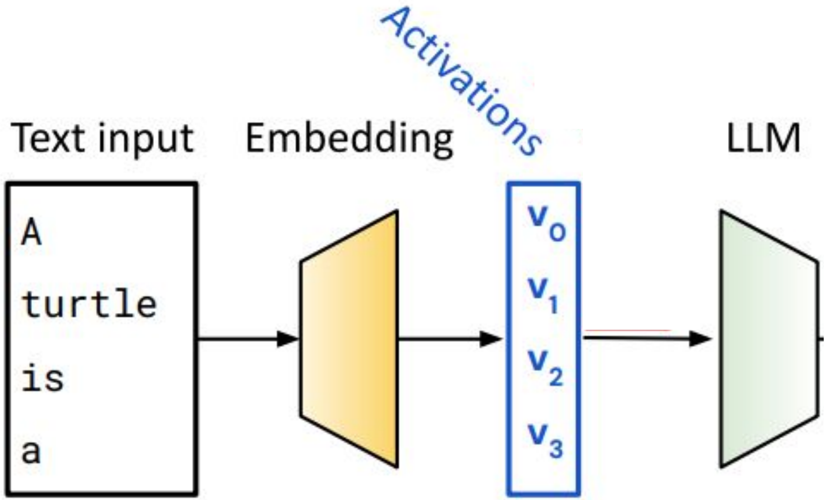
Útoky na velké jazykové modely



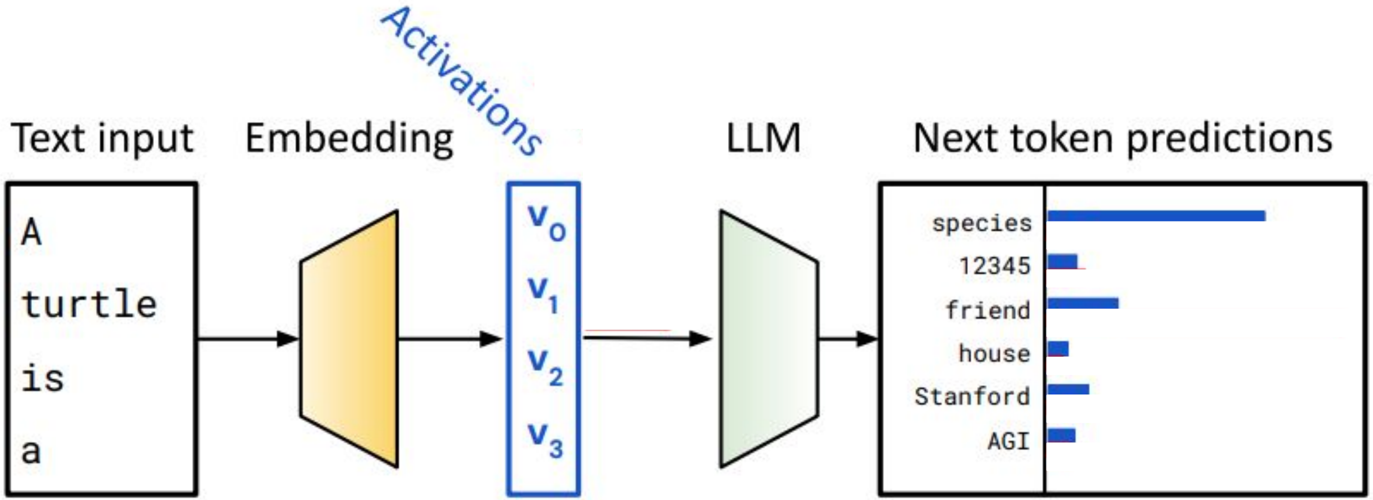
Útoky na velké jazykové modely



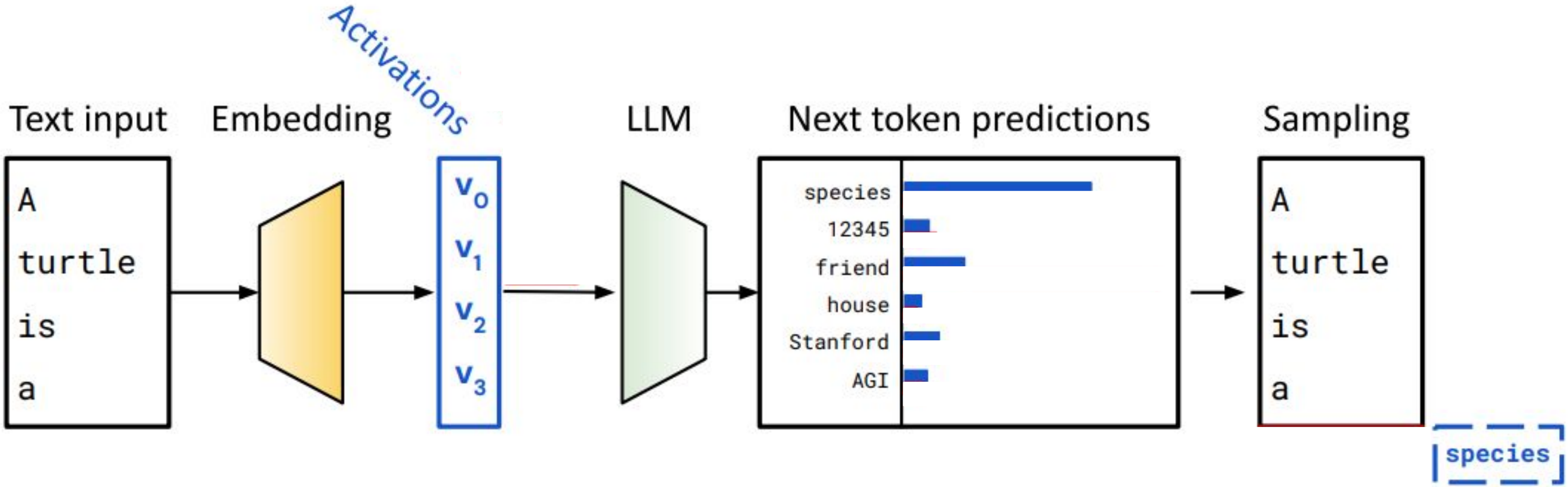
Útoky na velké jazykové modely



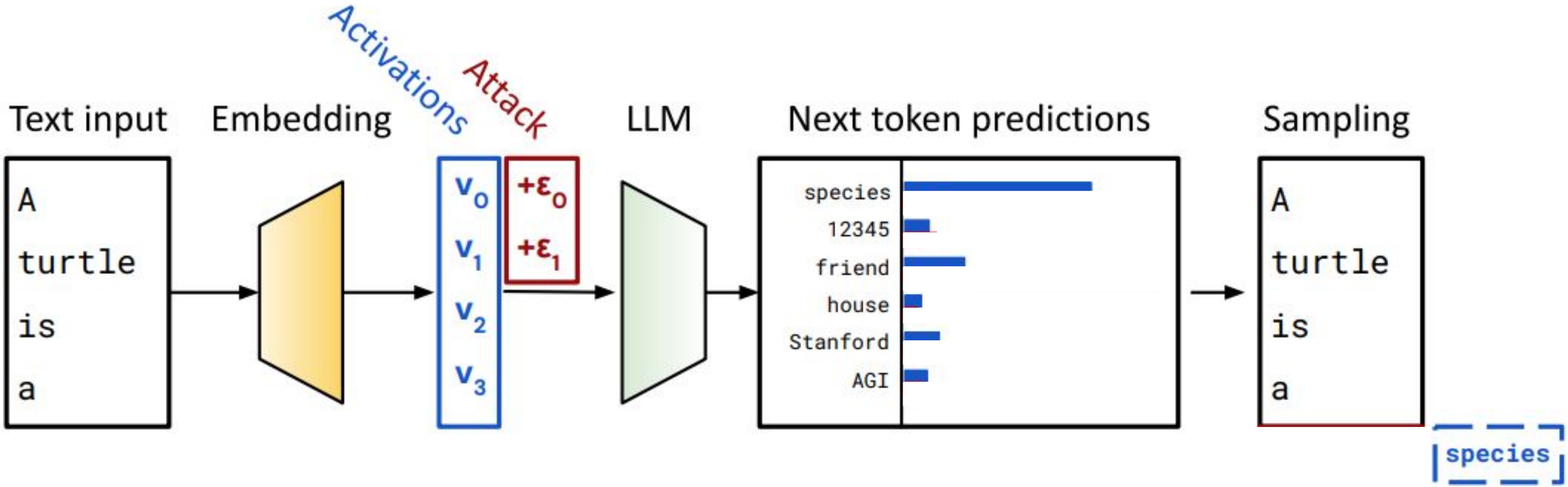
Útoky na velké jazykové modely



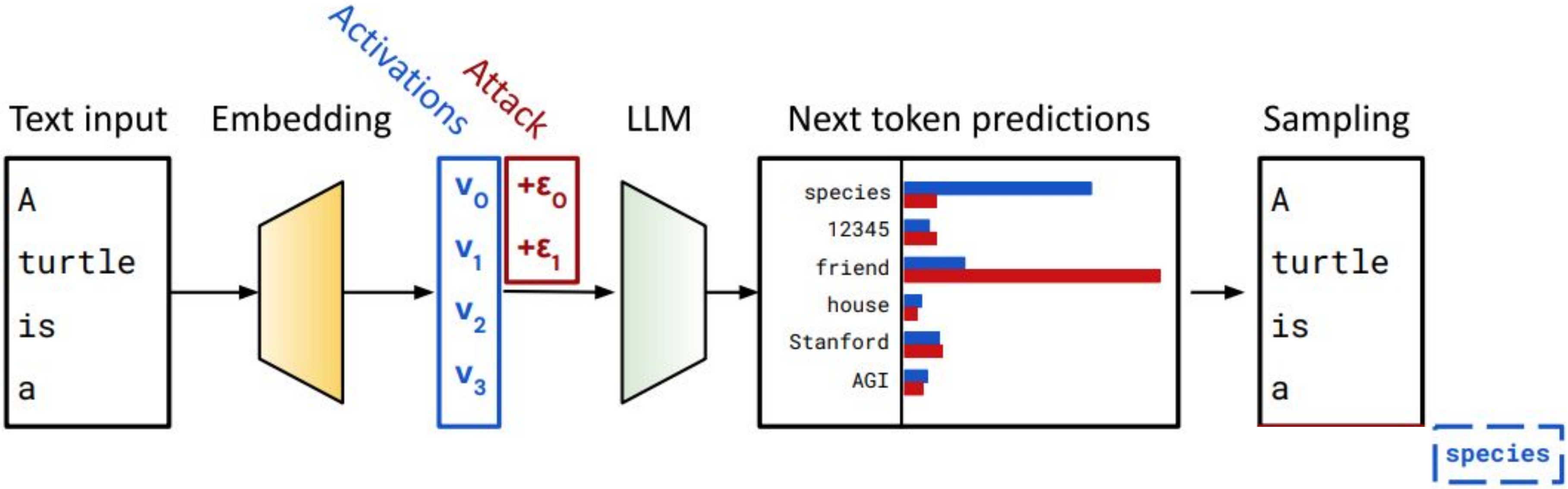
Útoky na velké jazykové modely



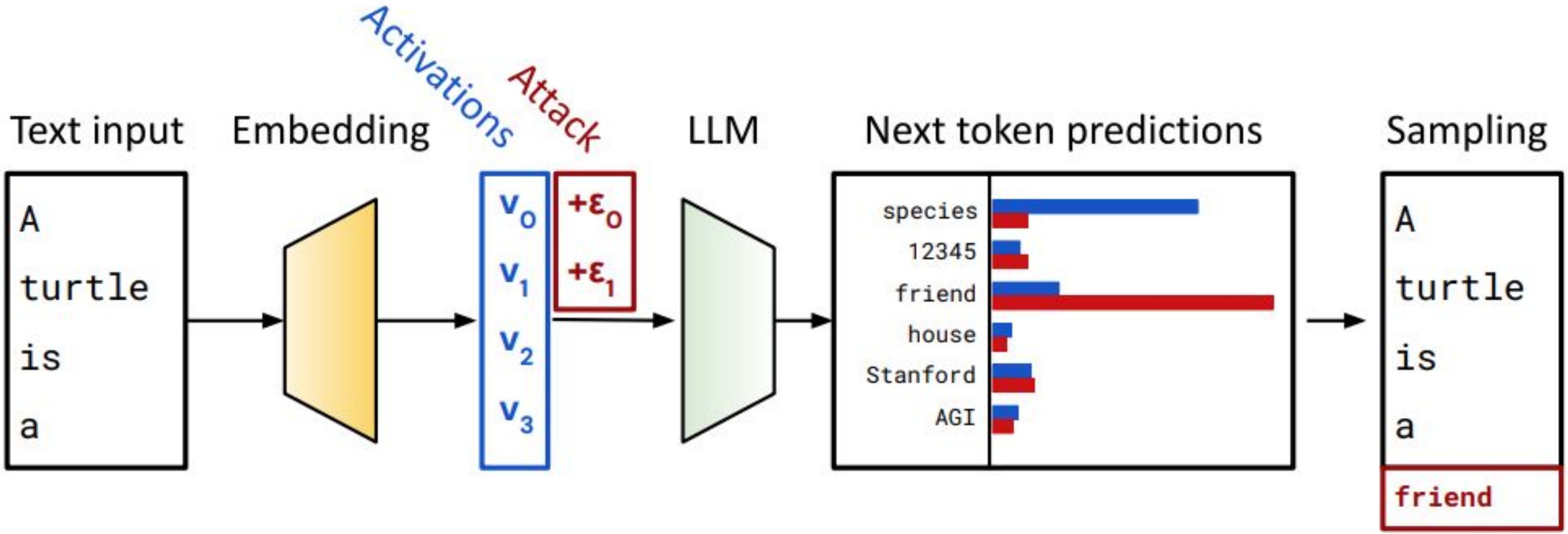
Útoky na velké jazykové modely



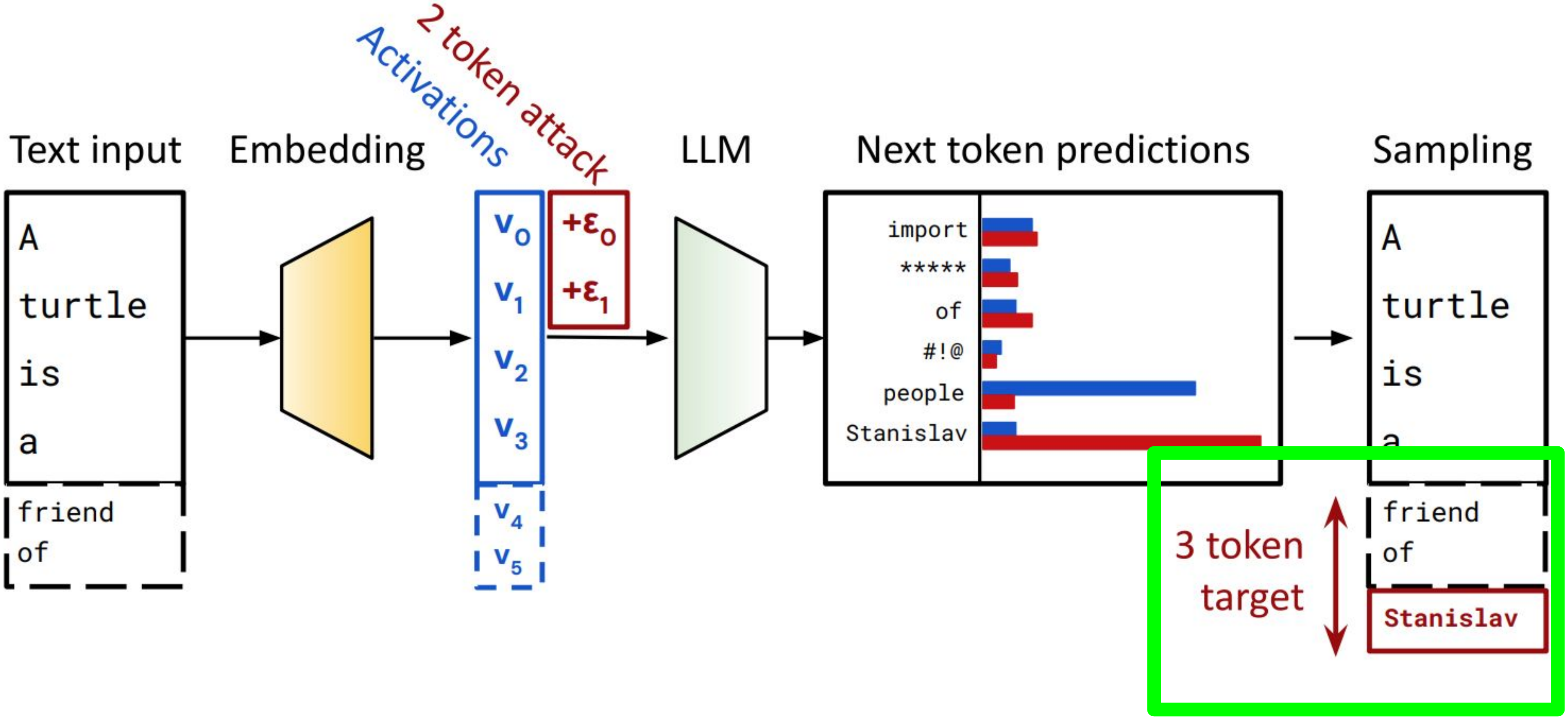
Útoky na velké jazykové modely



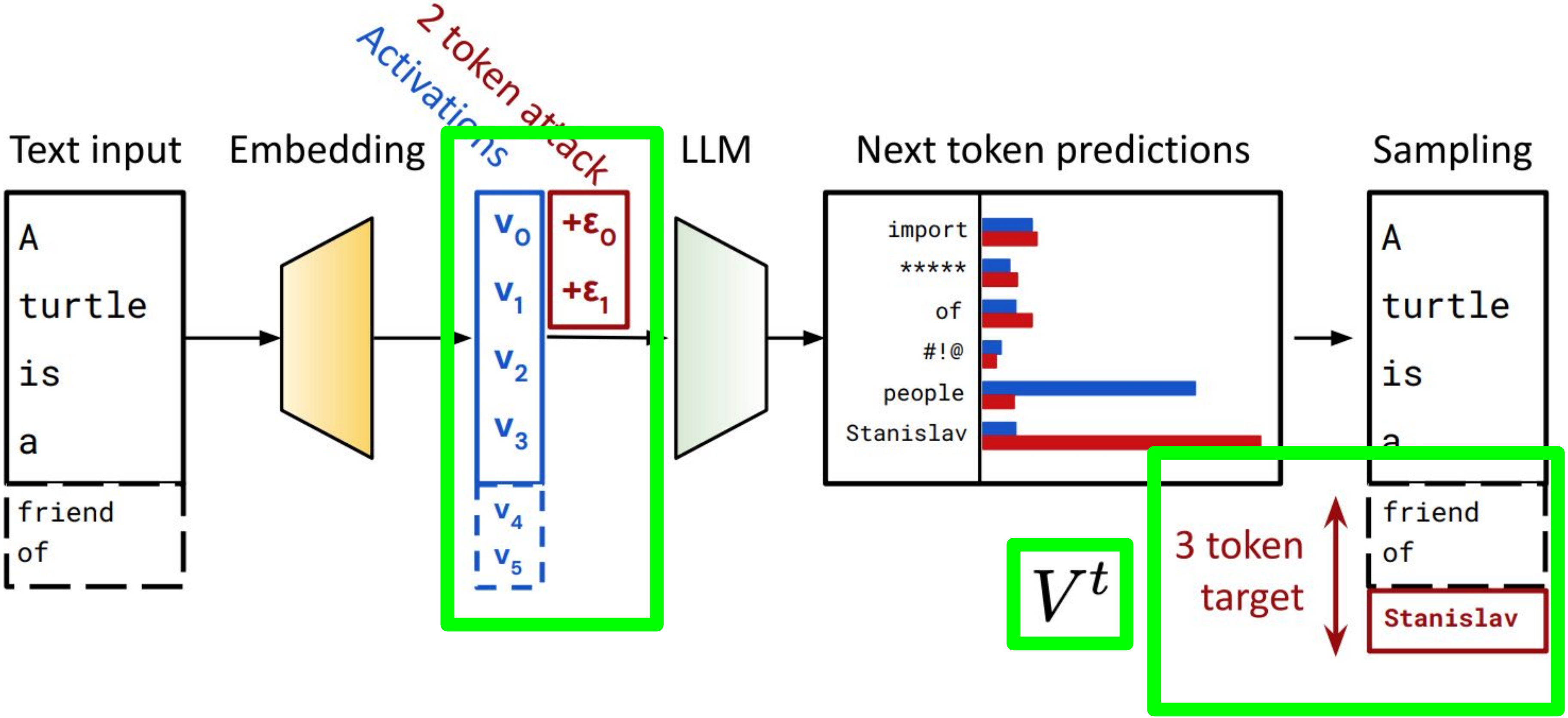
Útoky na velké jazykové modely



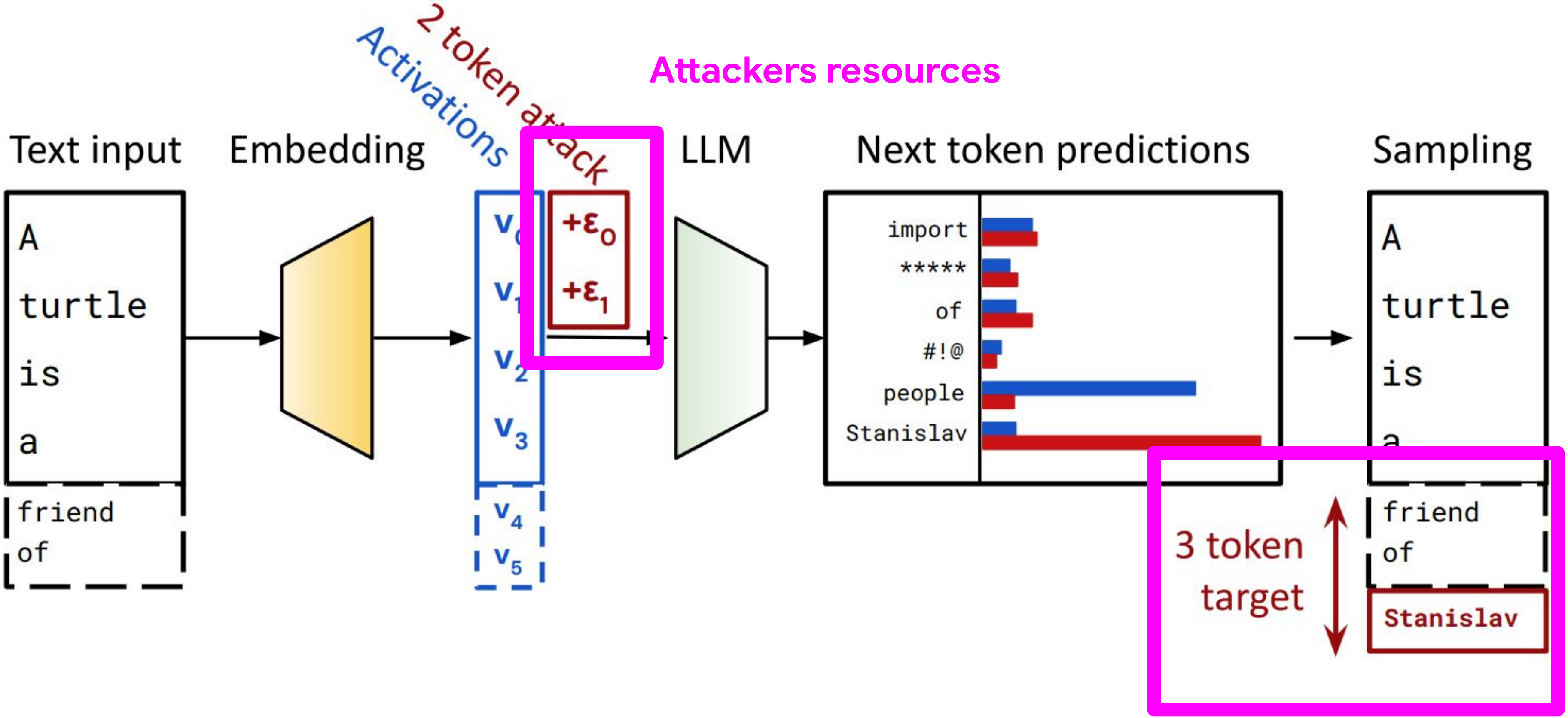
Útoky na velké jazykové modely



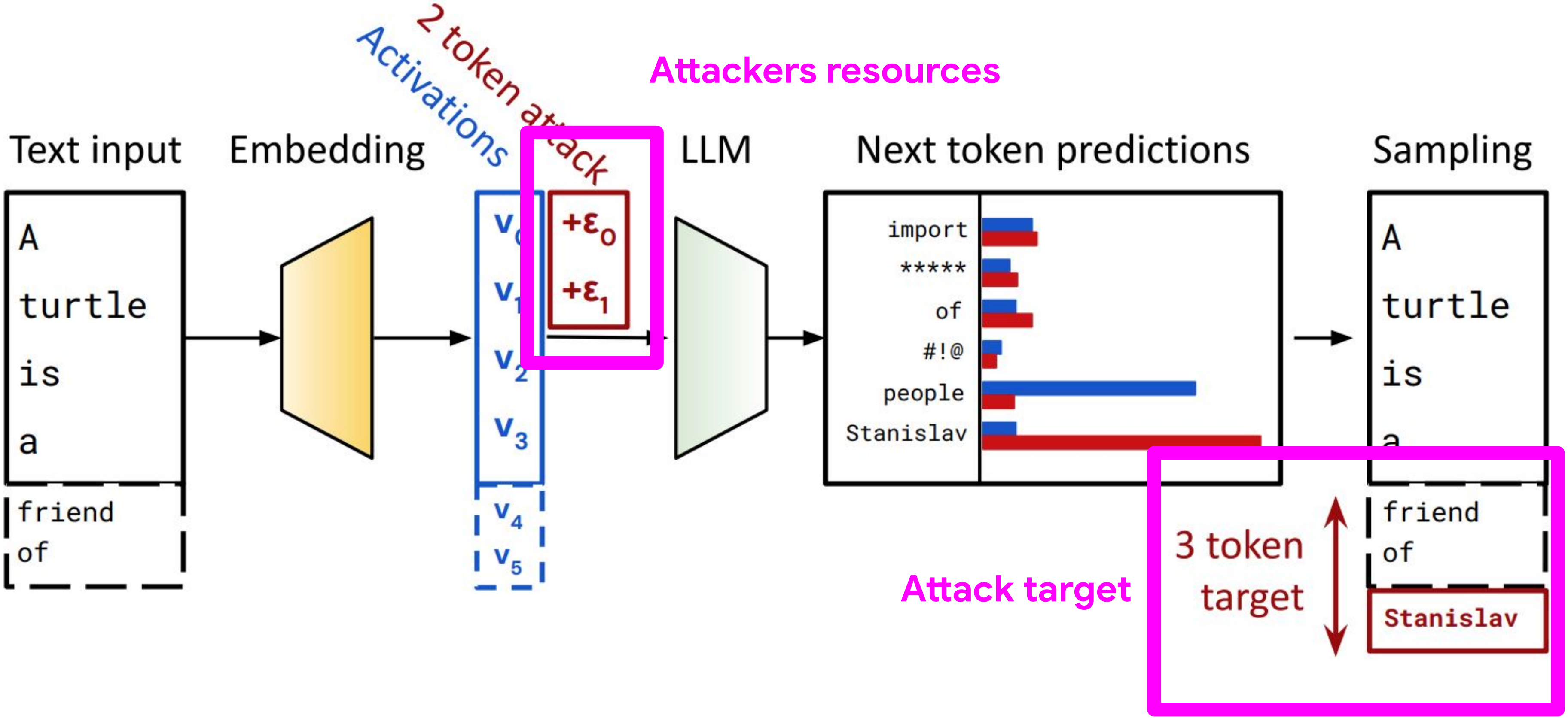
Útoky na velké jazykové modely



Útoky na velké jazykové modely



Útoky na velké jazykové modely



To samé platí pro jazyk

A simple math problem: $1 + 2 =$ **3**

To samé platí pro jazyk

A simple math problem: 1 + 2 =

To samé platí pro jazyk

$\boxed{\text{A}}$ simple math problem: 1 + 2 = **turtle**

v[0,38] += -0.16	v[0,198] += -4.93	v[0,465] += -0.84	v[0,619] += -2.71	v[0,743] += 1.771
v[0,40] += 1.717	v[0,261] += 0.074	v[0,476] += 5.678	v[0,622] += -0.91	v[0,772] += 3.212
v[0,51] += -1.30	v[0,401] += 0.158	v[0,498] += 8.186	v[0,631] += 3.011	v[0,802] += -3.37
v[0,61] += 4.093	v[0,410] += 0.822	v[0,552] += 0.632	v[0,698] += -3.44	v[0,939] += 7.616
v[0,120] += -0.66	v[0,414] += 3.166	v[0,578] += 2.167	v[0,703] += -1.10	v[0,1016] += -6.30

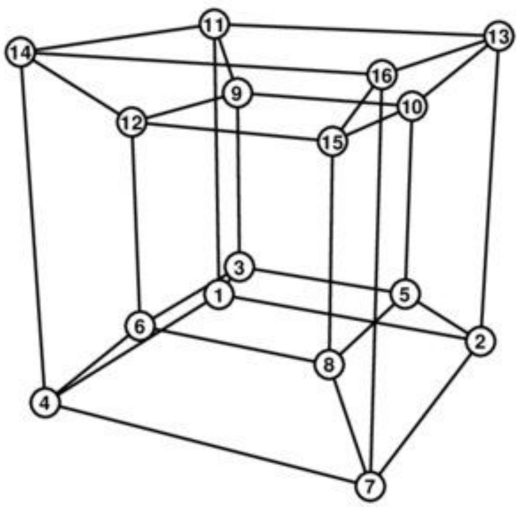
To samé platí pro jazyk

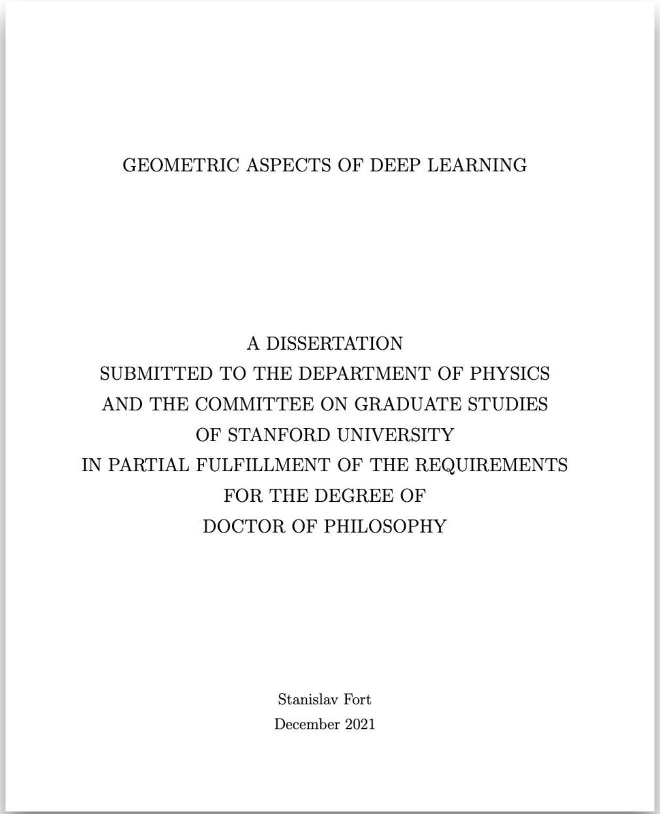
$\boxed{\text{A}}$ simple math problem: 1 + 2 =

To samé platí pro jazyk

A simple math problem: 1 + 2 = Oh, Death was never enemy of ours! We laughed at him, we leagued with him, old chum. No soldier's paid to kick against His powers. We laughed, — knowing that better men would come, And greater wars: when each proud fighter brags He wars on Death, for lives; not men, for flags.

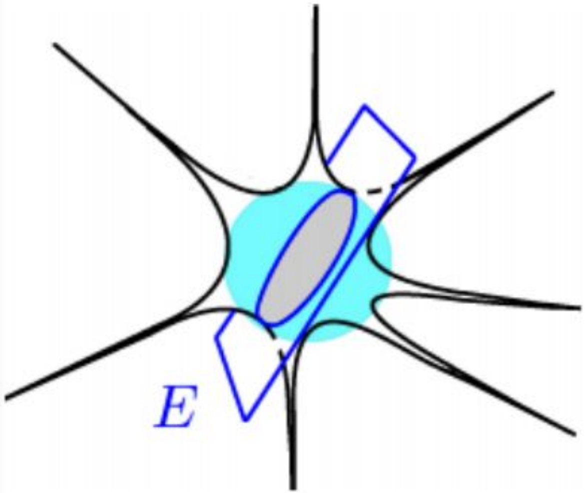
Vše souvisí s vysokorozměrnou geometrií



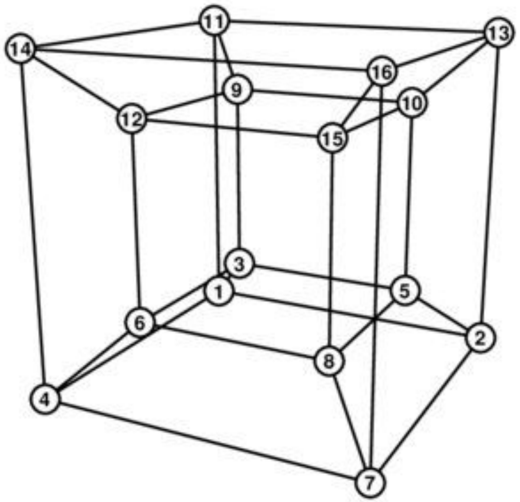

GEOMETRIC ASPECTS OF DEEP LEARNING

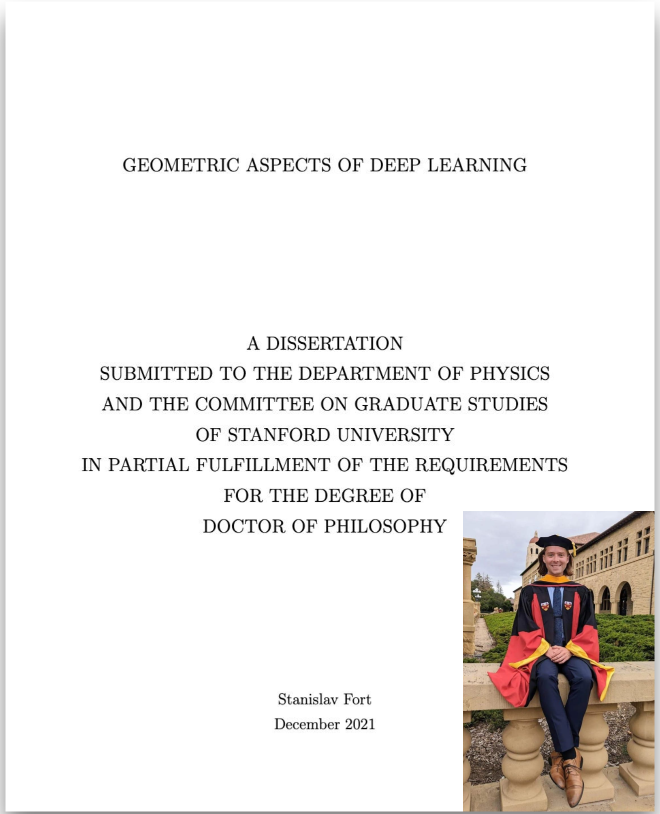
A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHYSICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Stanislav Fort
December 2021



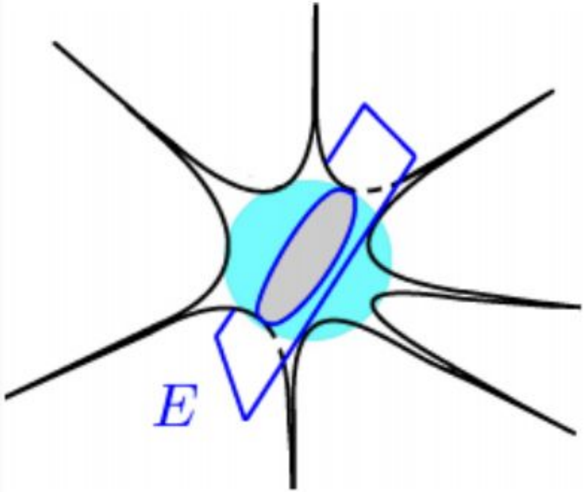
Vše souvisí s vysokorozměrnou geometrií




GEOMETRIC ASPECTS OF DEEP LEARNING

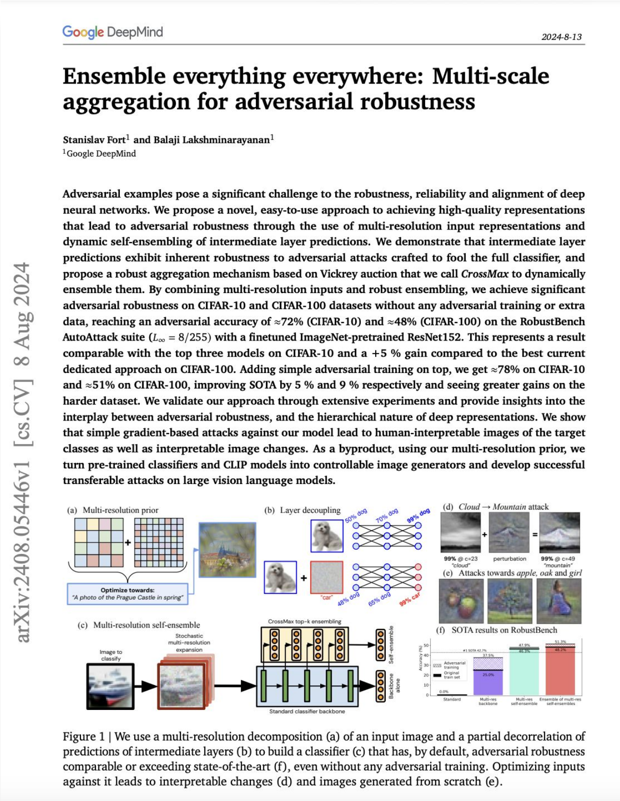
A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHYSICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Stanislav Fort
December 2021



Co s tím? Inspirace mozkem!

stanislavfort.com/ensemble-everything/



Google DeepMind 2024-8-13

Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

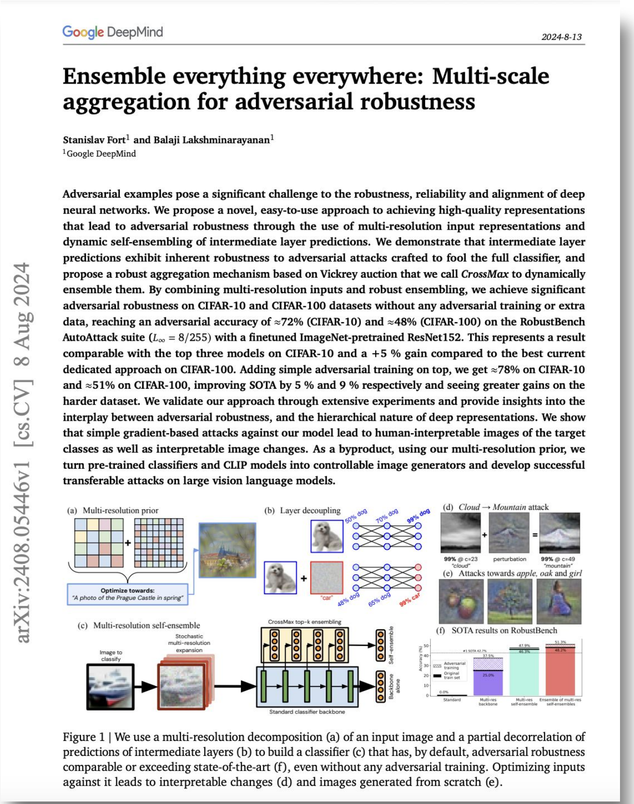
Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).

Co s tím? Inspirace mozkem!

stanislavfort.com/ensemble-everything/


Google DeepMind 2024-8-13

Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).

arXiv:2408.05446v1 [cs.CV] 8 Aug 2024

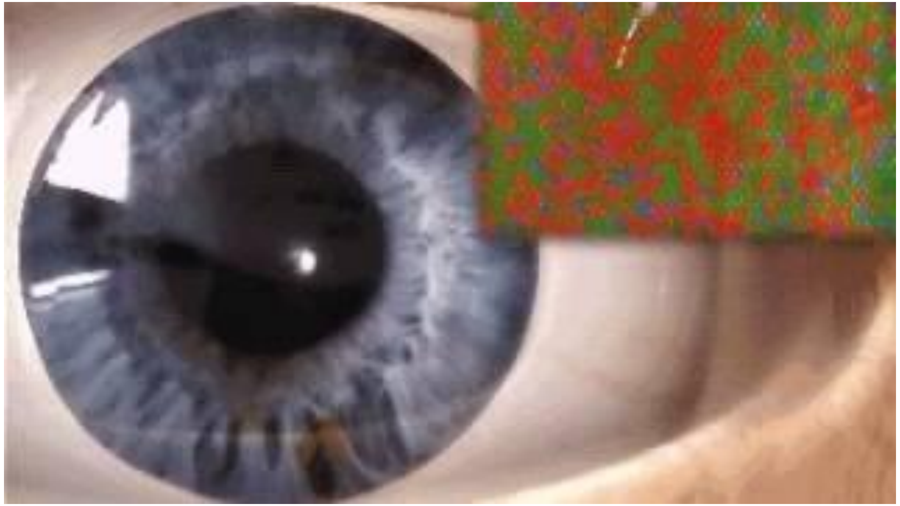
1) Vstupem je *několik* rozlišení zároveň → ensemblování přes rozlišení

2) Agregace predikcí z různých vrstev sítě → ensemblování přes abstrakce

3) Agregační mechanismus inspirovaný Vickreyho aukčním mechanismem

4) Bonus: každá klasifikátor je teď automaticky generátor

Micro a macro sakády = vaše oči se pořád klepou



2 - 100 arcmin několikrát za sekundu

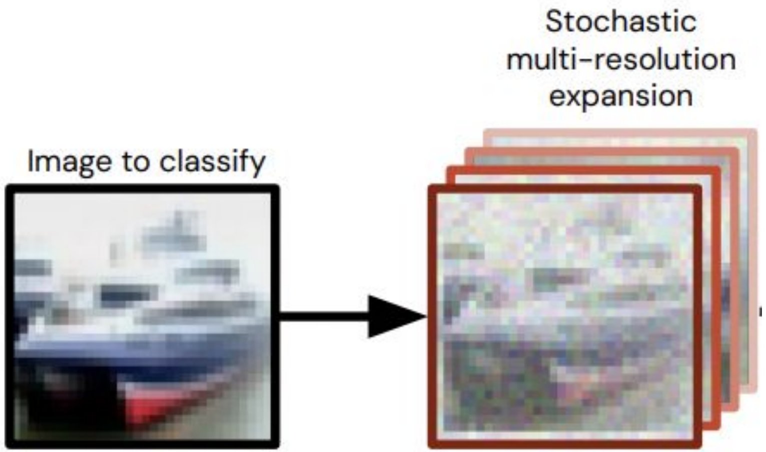
Vstup přes mnoho rozlišení + šum + cukání

Figure 3 | An image input being split into N progressively lower resolution versions that are then stacked channel-wise, forming a $3N$-channel image input to a classifier.

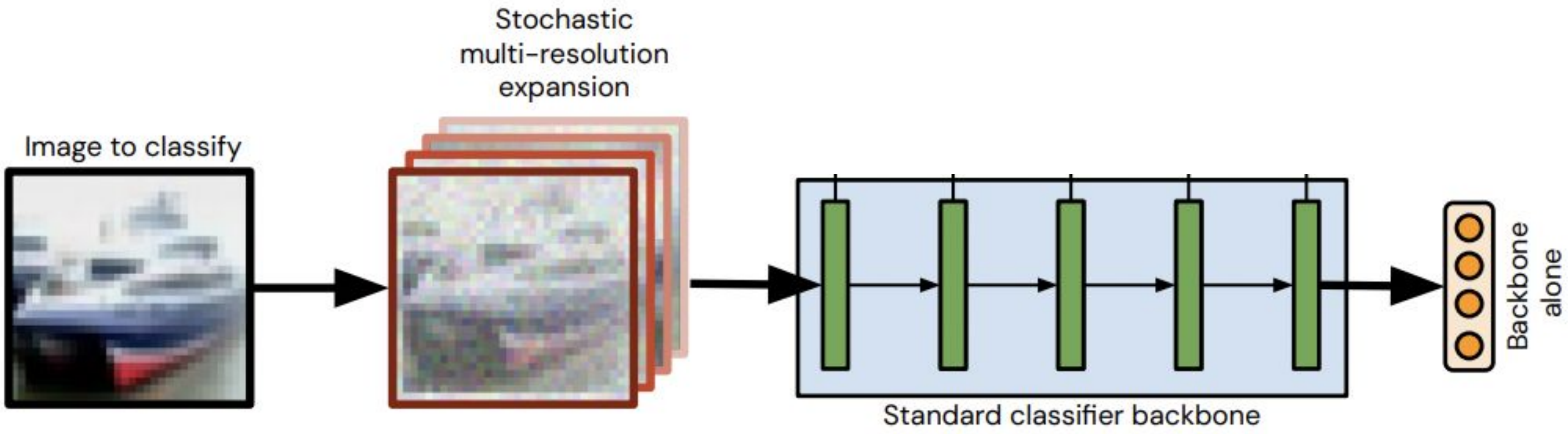
Trénink sítě, aby je klasifikovala všechny *najednou*

Image to classify

Trénink sítě, aby je klasifikovala všechny *najednou*

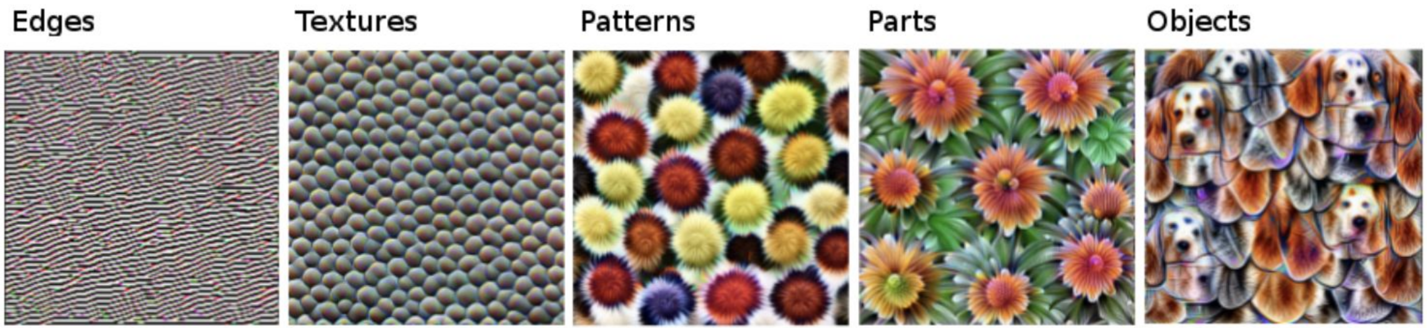


Trénink sítě, aby je klasifikovala všechny *najednou*

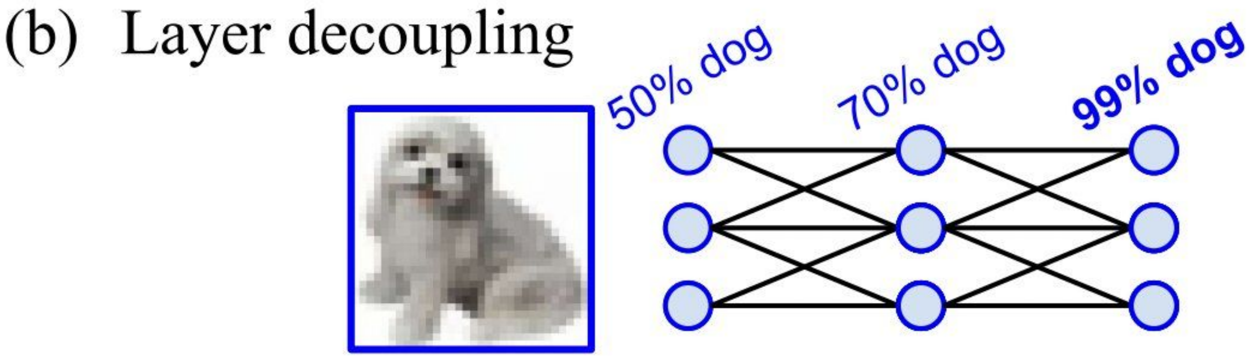


Přirozená robustnost různých vrstev sítě

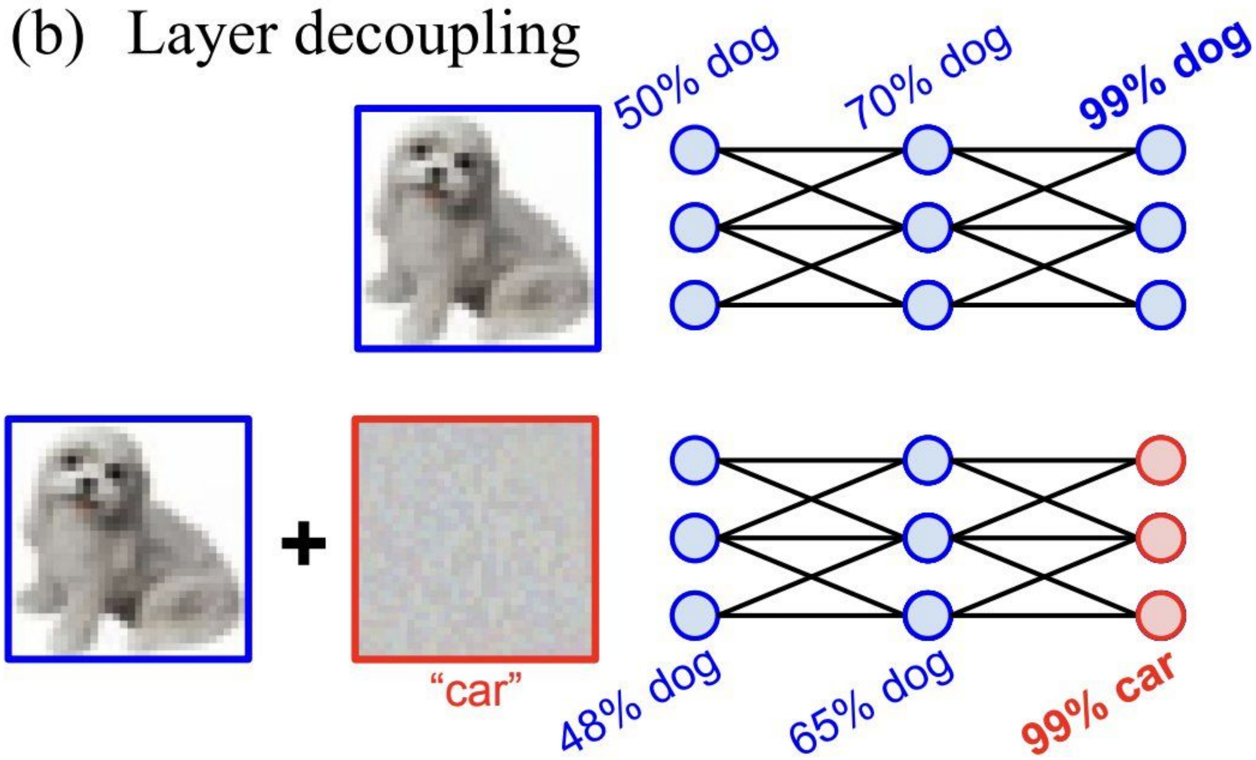
Klíčová otázka:
Vypadá obrázek 🐕, na který se zaútočilo, aby vypadal jako 🚗, má 🐕–podobné hrany, textury, & i vzorce na vyšší úrovni abstrakce?



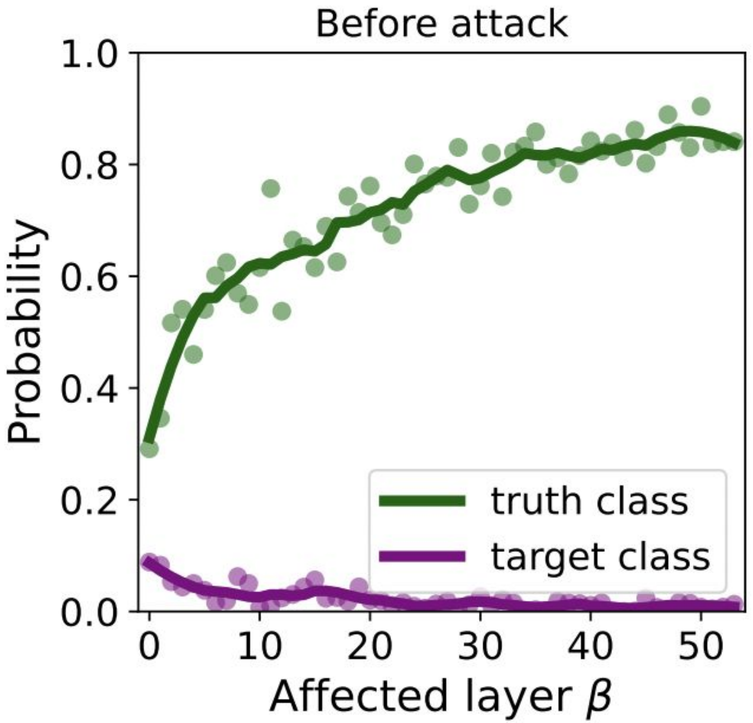
Přirozená robustnost mezi vrstvami



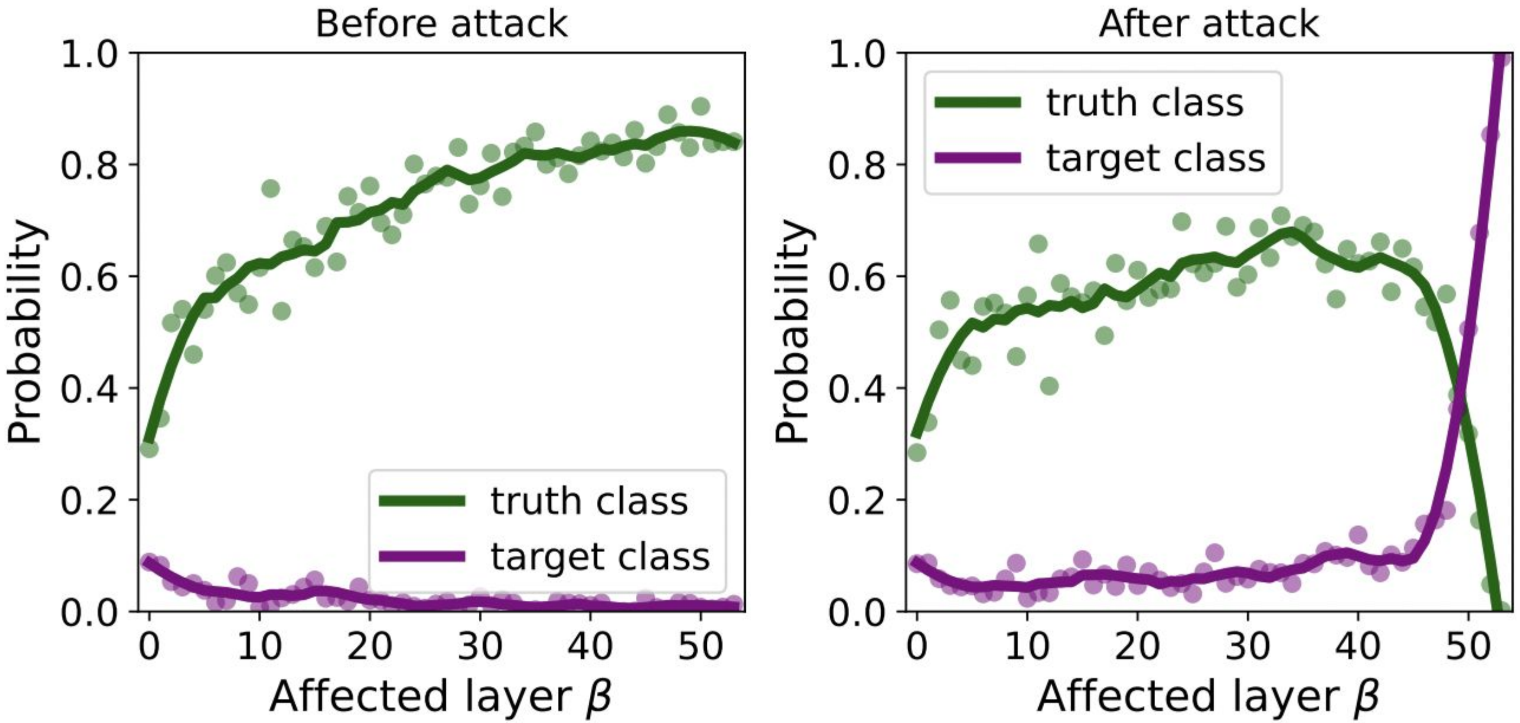
Přirozená robustnost mezi vrstvami



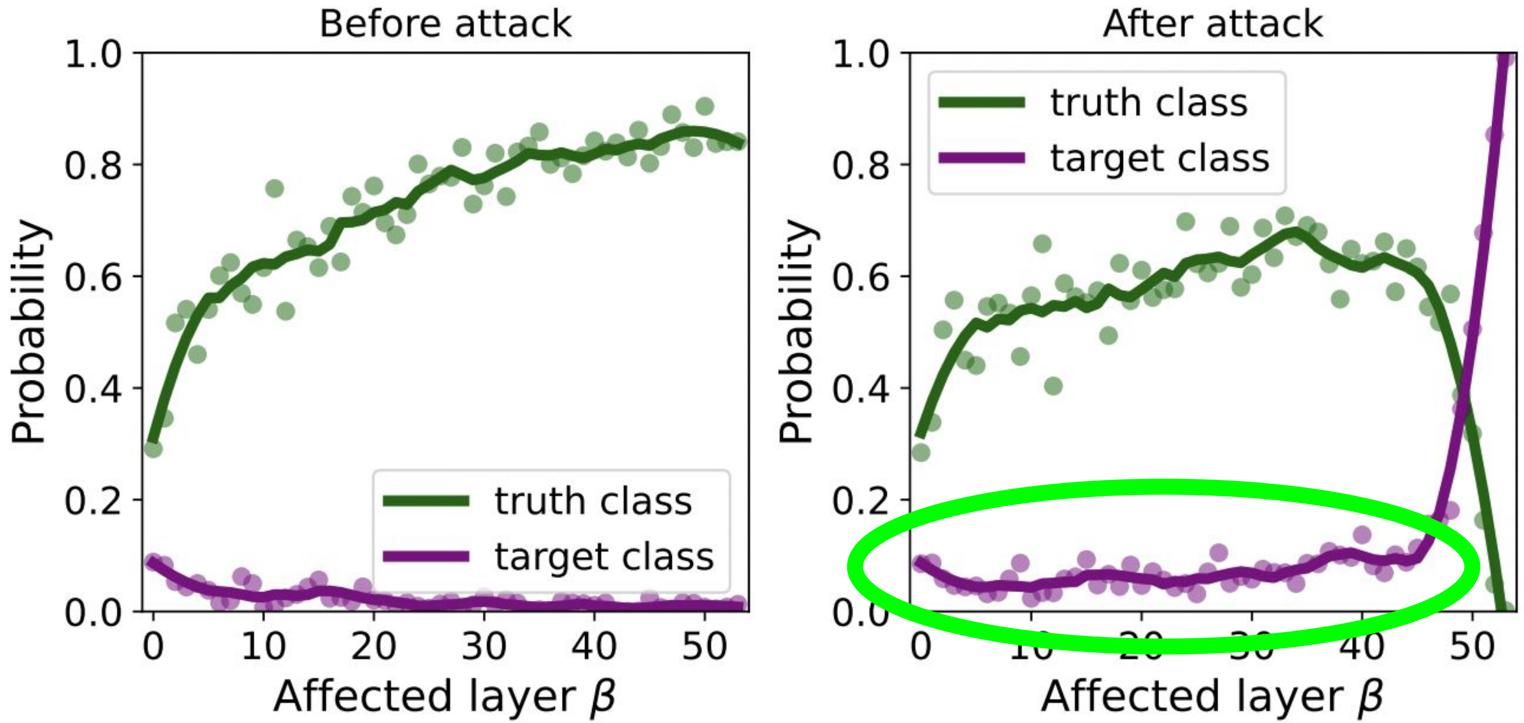
Přirozená robustnost mezi vrstvami



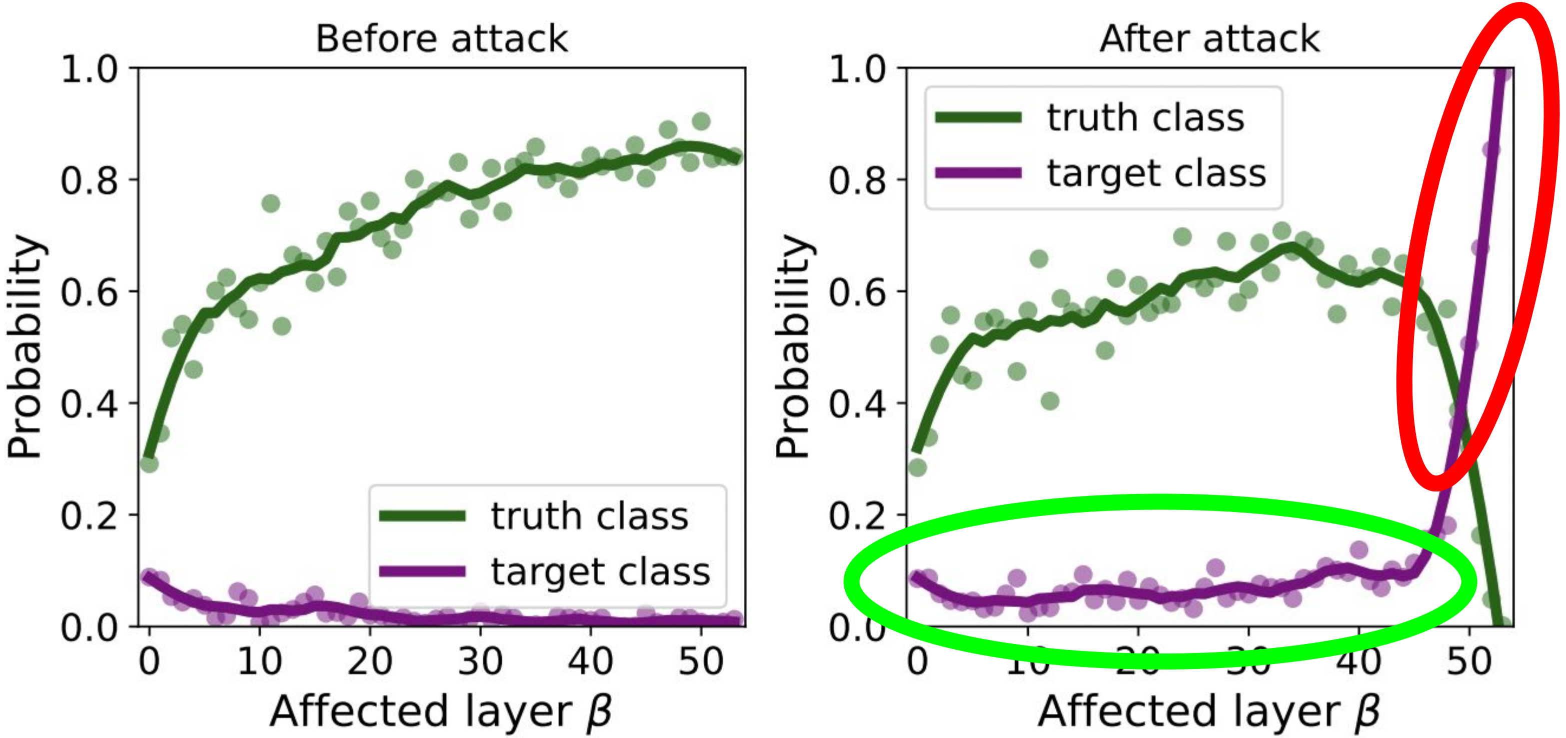
Přirozená robustnost mezi vrstvami



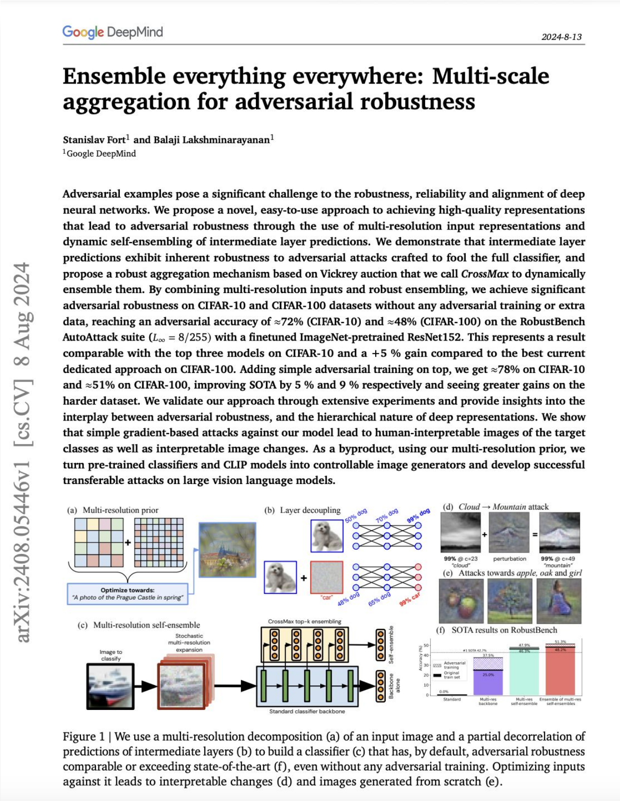
Přirozená robustnost mezi vrstvami



Přirozená robustnost mezi vrstvami



Všechno dohromady = robustnost



Google DeepMind 2024-8-13

arXiv:2408.05446v1 [cs.CV] 8 Aug 2024

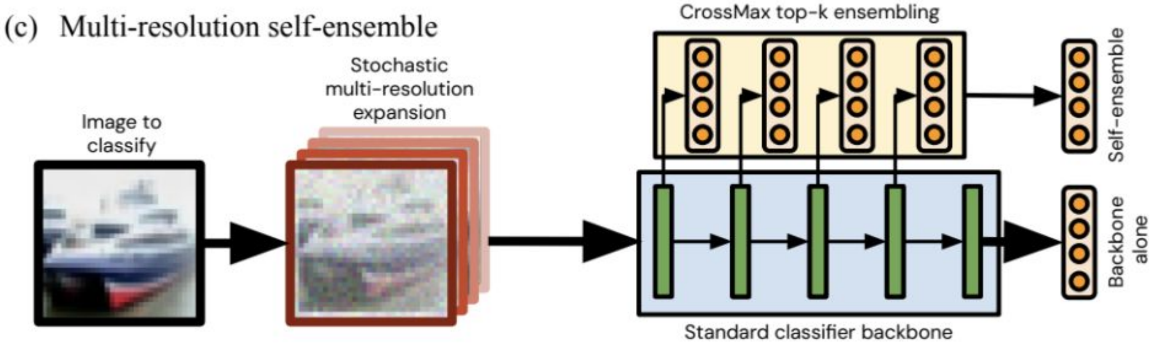
Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

(a) Multi-resolution prior
(b) Layer decoupling
(d) *Cloud* → *Mountain* attack
(e) Attacks towards *apple*, *oak* and *girl*
(c) Multi-resolution self-ensemble
(f) SOTA results on RobustBench

Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).



Hodně robustní síť

$$\frac{\partial\ \text{probability[class]}}{\partial\ \text{image}} =$$

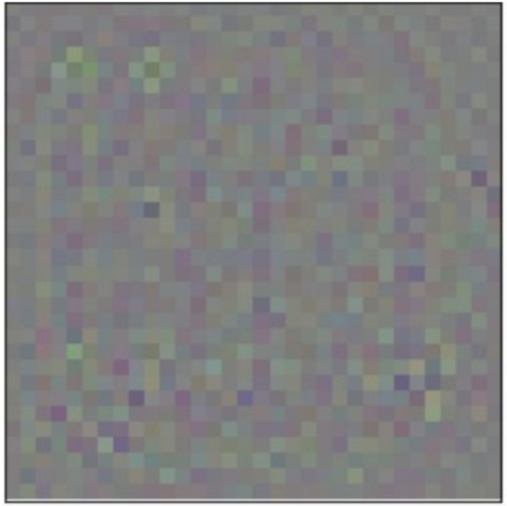
"Jak bych měl změnit pixely tak, aby se zvýšila pravděpodobnost třídy?"

Hodně robustní síť

$$\frac{\partial \text{ probability[class]}}{\partial \text{ image}} =$$

"Jak bych měl změnit pixely tak, aby se zvýšila pravděpodobnost třídy?"

"Jablko" pro standardní síť

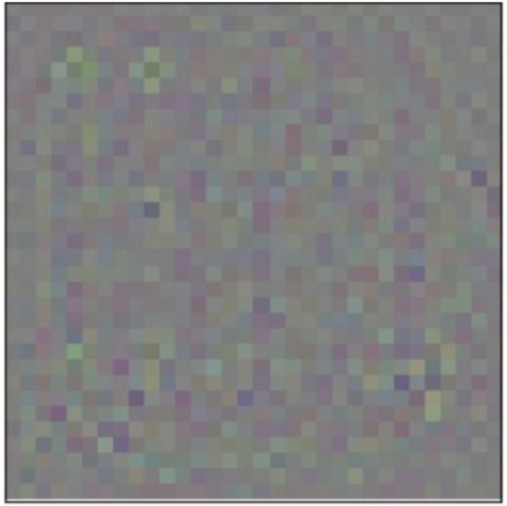


Hodně robustní síť

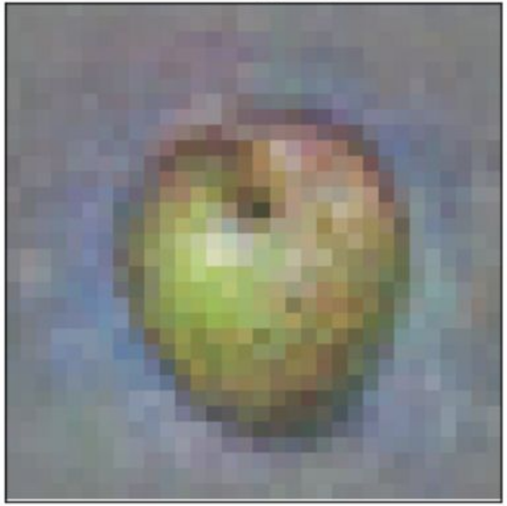
$$\frac{\partial\ \text{probability[class]}}{\partial\ \text{image}} =$$

„Jak bych měl změnit pixely tak, aby se zvýšila pravděpodobnost třídy?“

„Jablko“ pro standardní síť



„Jablko“ pro **naši**

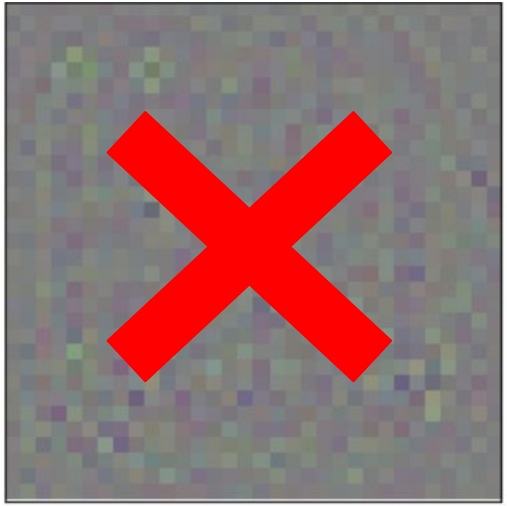


Hodně robustní síť

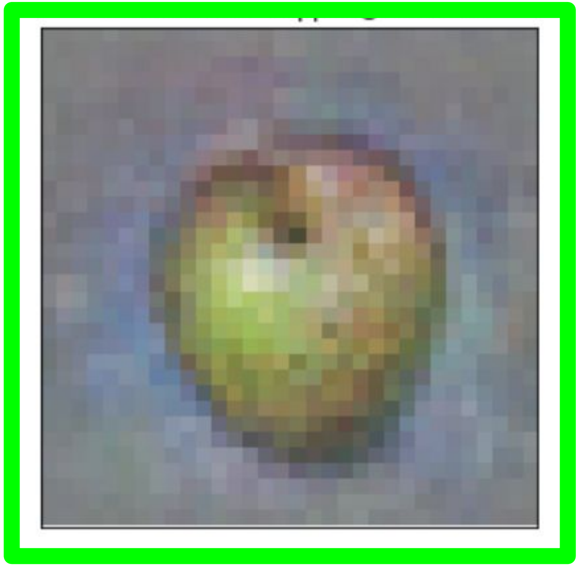
$$\frac{\partial\ \text{probability[class]}}{\partial\ \text{image}} =$$

“Jak bych měl změnit pixely tak, aby se zvýšila pravděpodobnost třídy?”

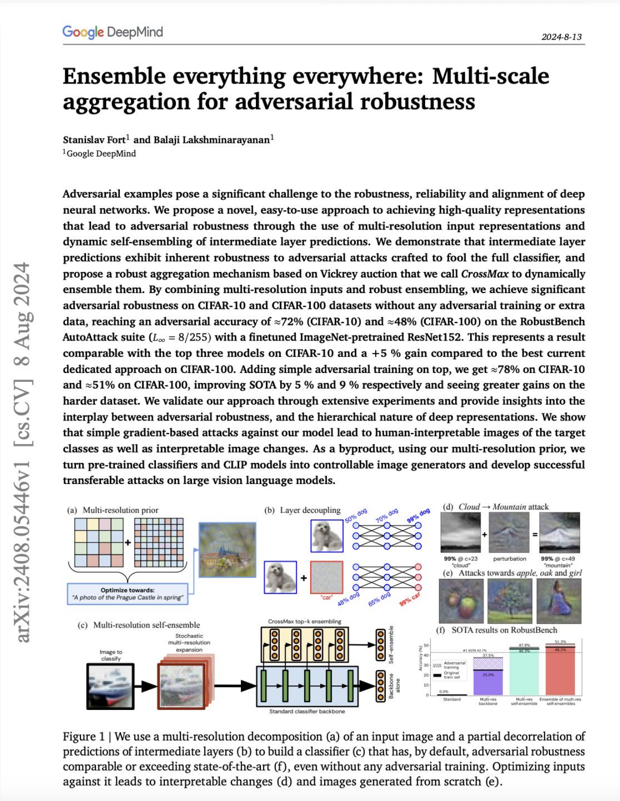
“Jablko” pro standardní síť



“Jablko” pro **naši**



Co s tím? Inspirace mozkem!


Google DeepMind 2024-8-13

Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

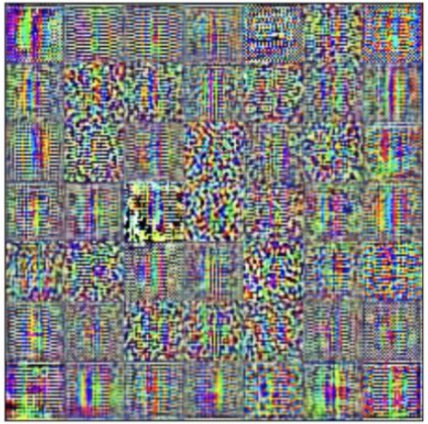
Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

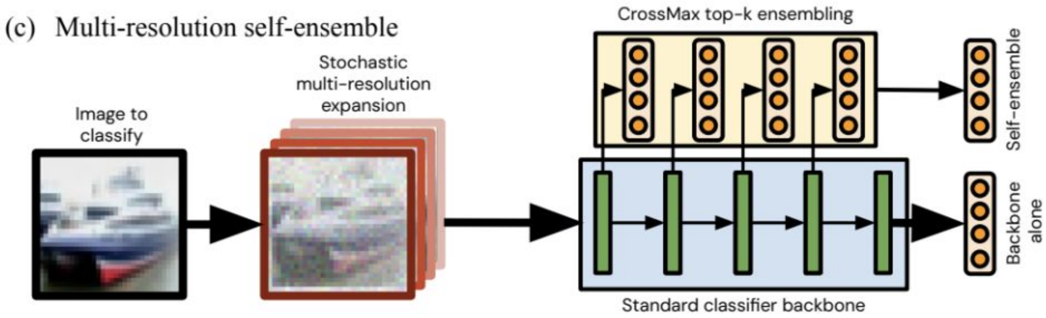
Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

arXiv:2408.05446v1 [cs.CV] 8 Aug 2024

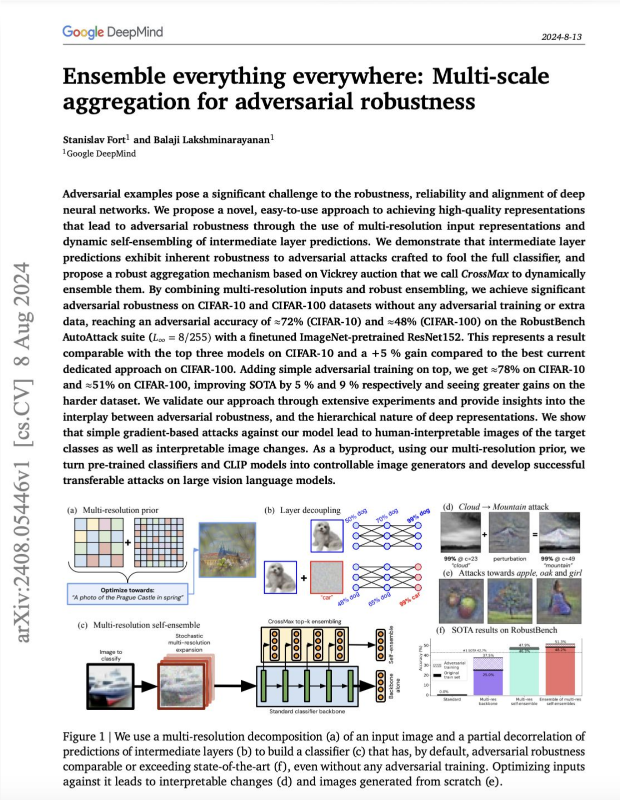
(a) Multi-resolution prior
Optimize towards: "A photo of the Prague Castle in spring"
(b) Layer decoupling
(c) Multi-resolution self-ensemble
(d) Cloud → Mountain attack
(e) Attacks towards apple, oak and girl
(f) SOTA results on RobustBench

Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).





Co s tím? Inspirace mozkem!


Google DeepMind 2024-8-13

Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

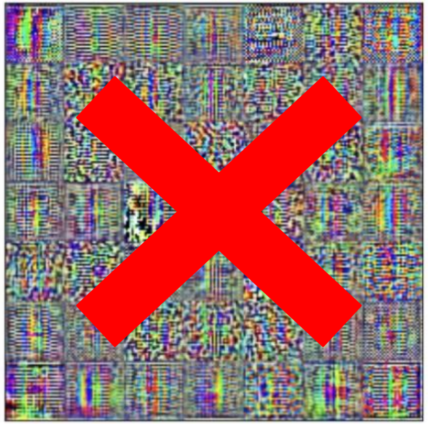
Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

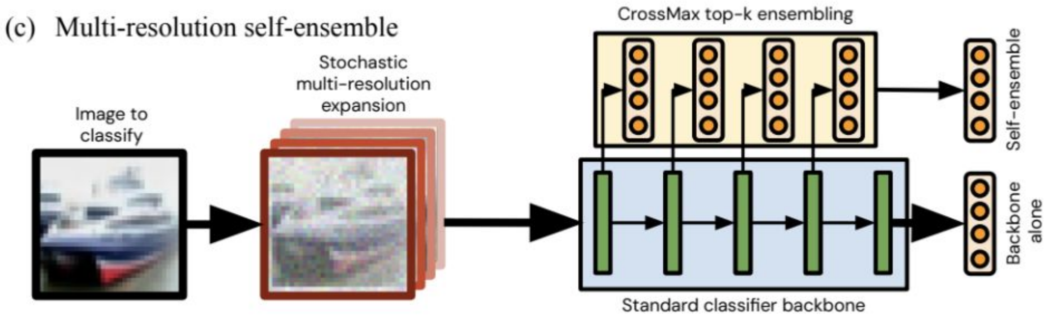
Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

(a) Multi-resolution prior
Optimize towards: "A photo of the Prague Castle in spring"
(b) Layer decoupling
(c) Multi-resolution self-ensemble
(d) *Cloud* → *Mountain* attack
(e) Attacks towards *apple, oak* and *girl*
(f) SOTA results on RobustBench

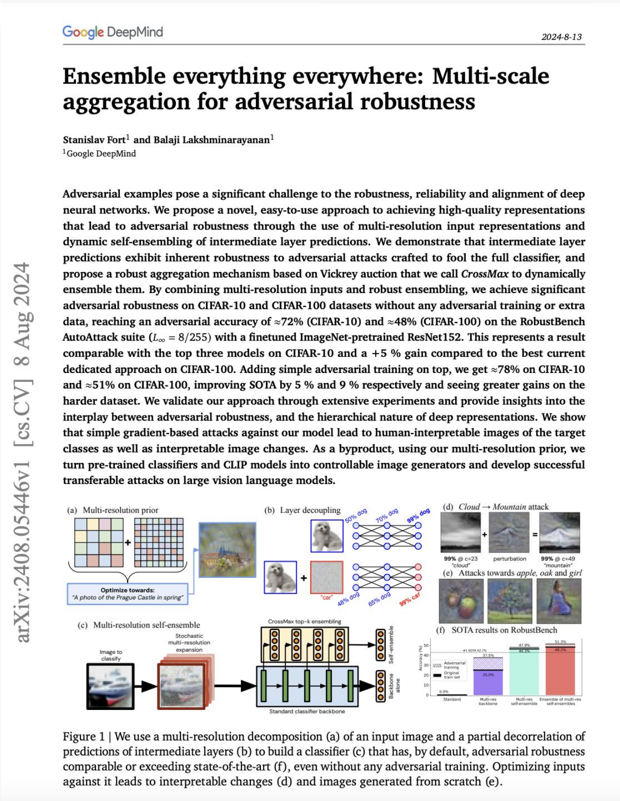
Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).

arXiv:2408.05446v1 [cs.CV] 8 Aug 2024





Co s tím? Inspirace mozkem!



Google DeepMind 2024-8-13

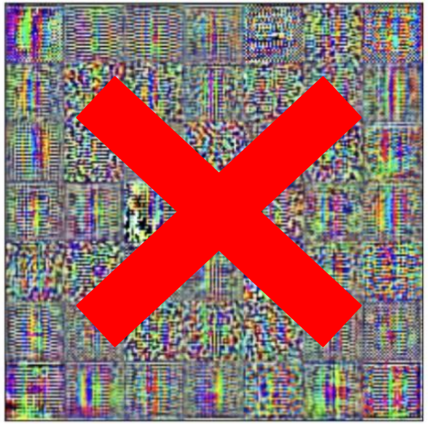
Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

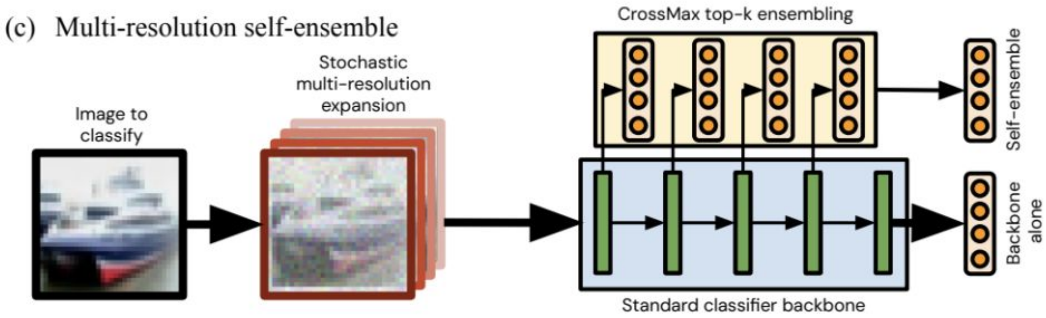
Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

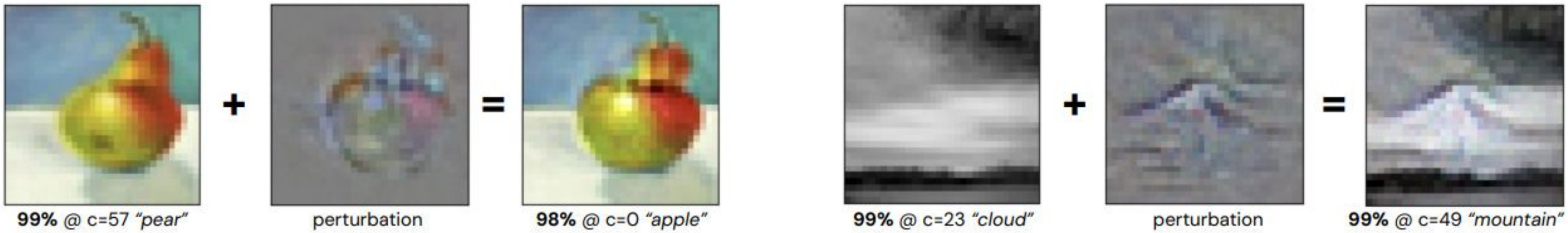
Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).

arXiv:2408.05446v1 [cs.CV] 8 Aug 2024





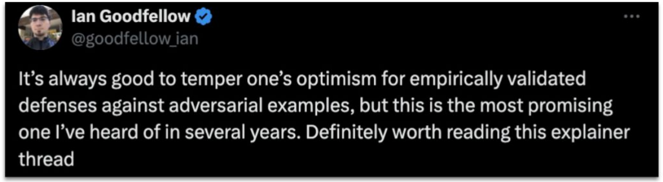
Interpretovatelné útoky jsou jediné, co funguje

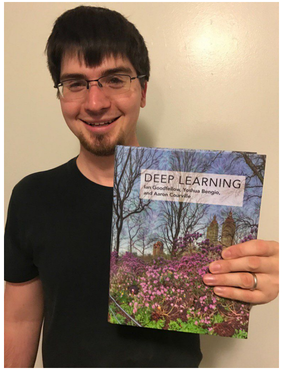


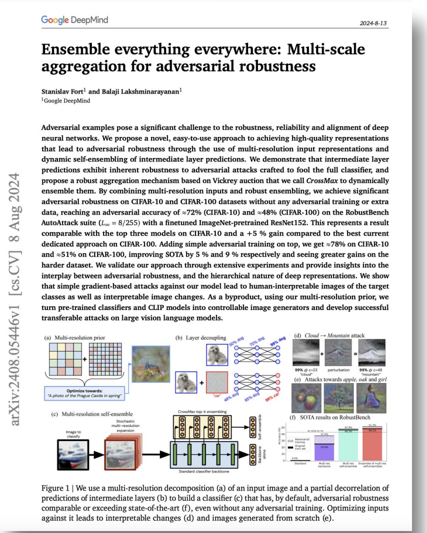
(a) *Pear to apple* (b) *Cloud to mountain*

(a) Original (b) *Albert Einstein* (c) *Queen Elizabeth* (d) *Nikola Tesla*

Velký kus k řešení pro *vizuální sítě*







Google DeepMind 2024-8-13

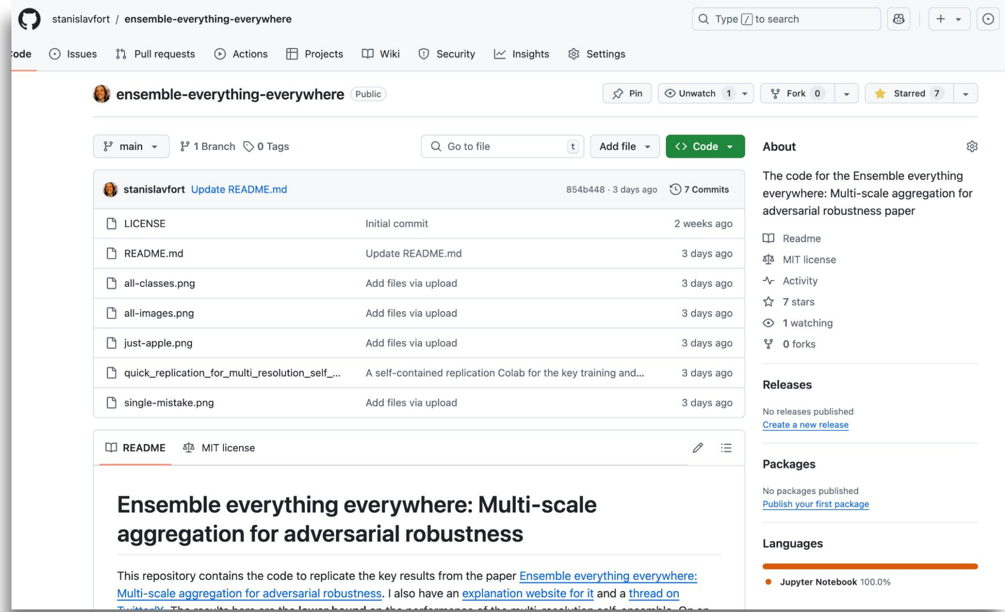
arXiv:2408.05446v1 [cs.CV] 8 Aug 2024

Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).



Děkuji za pozornost!

Zajímá vás, na čem pracuju?
Spojme se!