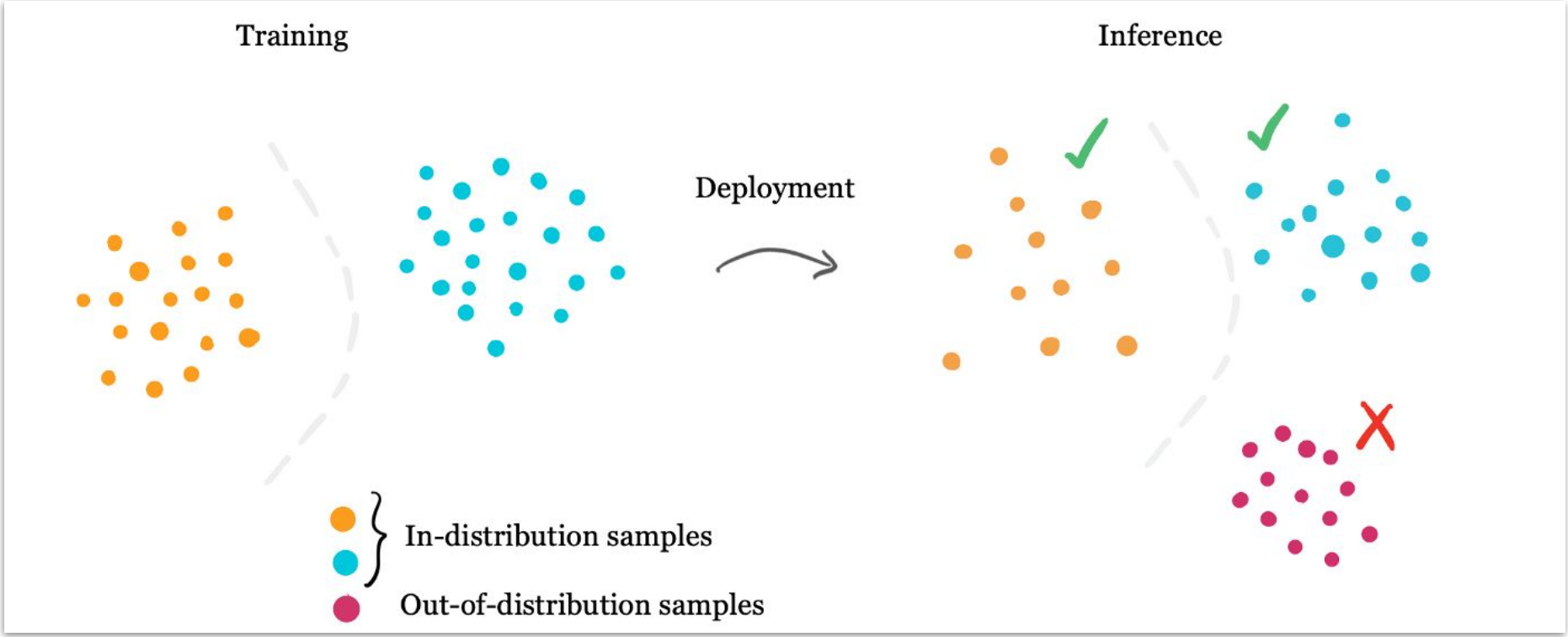
Robust Adversarial OOD Detection With Vision Transformer

Motivation

https://www.reisefestival.de/galerien/on-the-road-ein-autofahrtenbilderbuch/autobild-ampel-regen-01-1000/

https://burkhartmarketing.com/outdoor-advertising-go-for-the-big-impact/

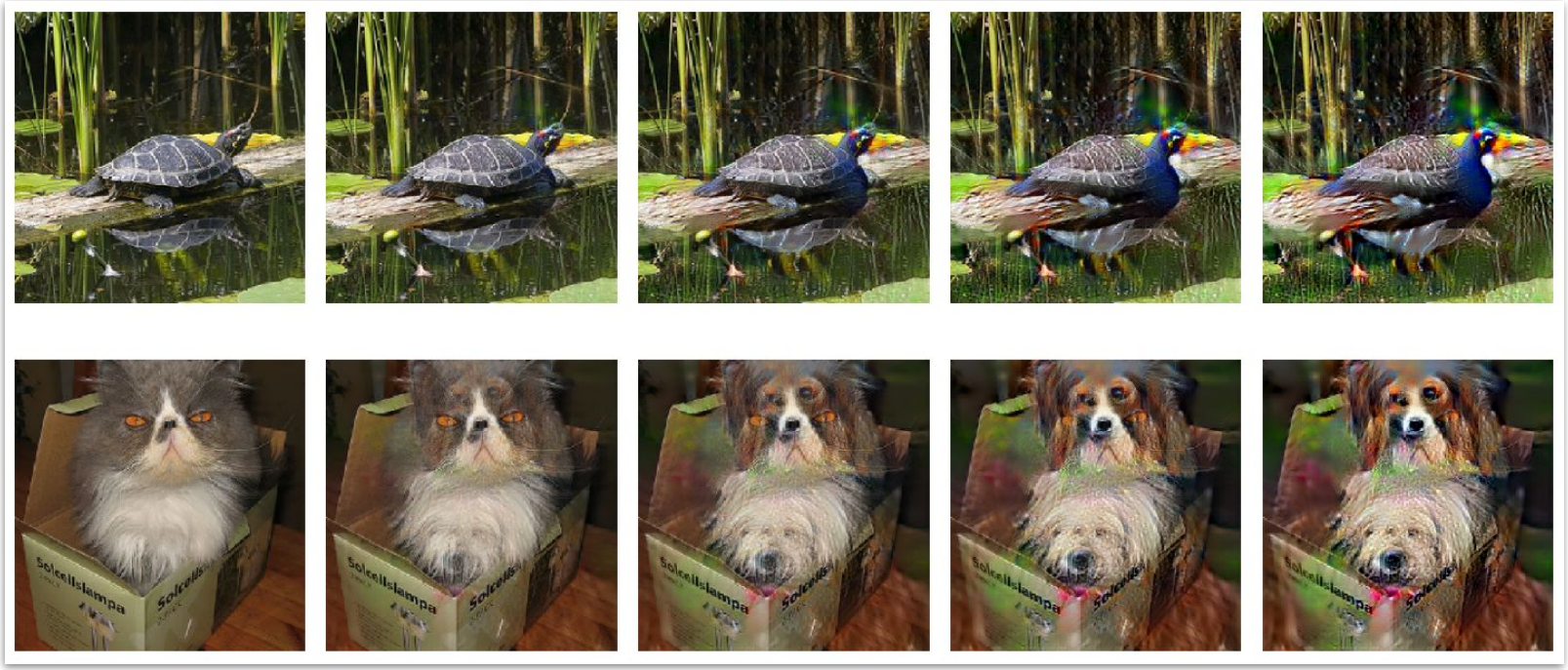
1. OOD Detection



https://medium.com/geekculture/out-of-distribution-detection-in-medical-ai-b638b385c2a3

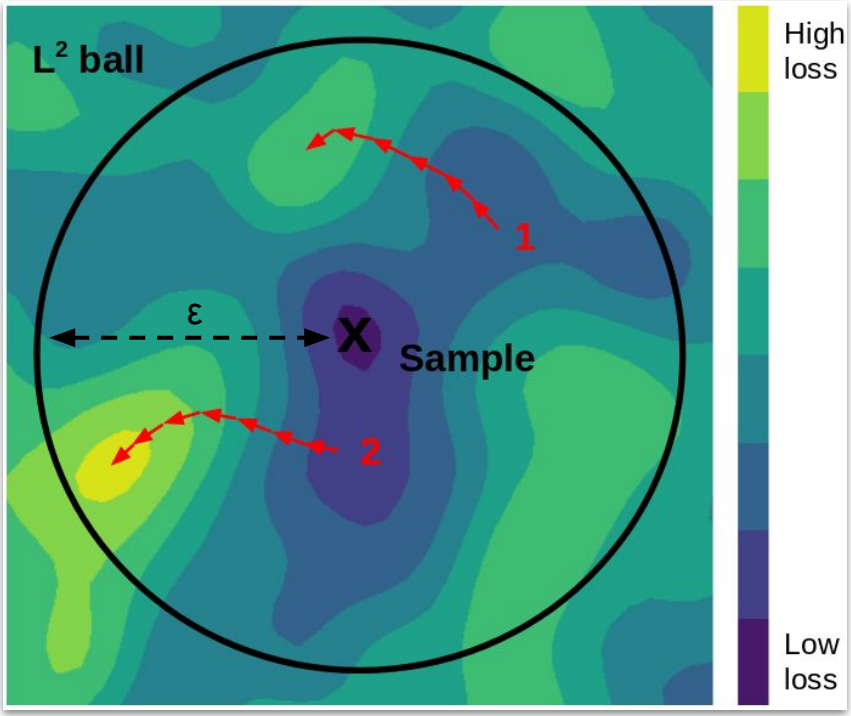
2. Adversarial Attacks - PGD (1/2)

PGD = Projected Gradient Descent



Tsipras et al., 2018, “Robustness May Be at Odds with Accuracy”

2. Adversarial Attacks - PGD (2/2)



Oscar Knagg, 2019, “Know your enemy - How you can create and defend against adversarial attacks”

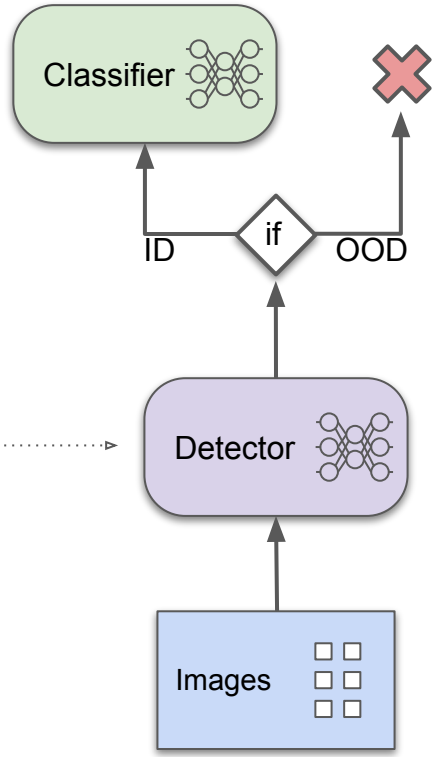
3. Robustness in Adversarial OOD Detection

	ID sample	OOD sample
Along gradient = bring sample closer to the expectations of the model	Improve quality of sample ex: noise reduction, sharpen edges or unblur entire images/videos	Trying to trick model into believing it is handling an ID sample If model detects correctly → **Adversarial Robustness**
Against gradient = drift sample away from the expectations of the model	Worsen quality of sample ex: simulation of real world noise, road sign classification of cars with simulated snow/rain/fog/smoke/… , blur images, unsharpen edges	Removing an unfamiliar sample even further from the models expectations → currently no use case

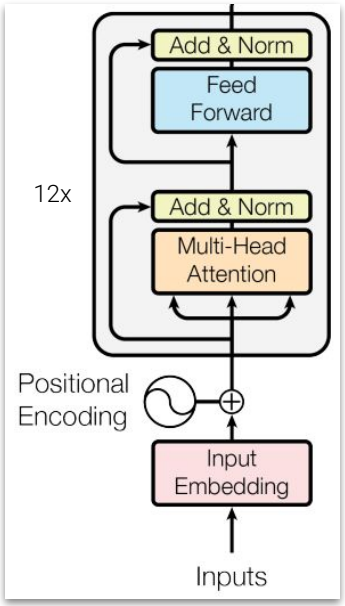
4. Design of the Detection Pipeline

State of the art:
ProoD uses 5 layer CNN as detector

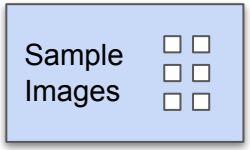
```
FC(128, 1)
FC(16384, 128)
AvgPool(2)
Conv2d(256, 256)
Conv2d(128, 256)
Conv2d(3, 128)
```



This thesis:
ViT Tiny as detector

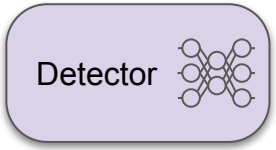


5. Data, Models & Attack



- Cifar10, Cifar100, SVHN → only colored images (3 channels = RGB)
- Batch size 32, scaled to 224x224 pixel, patch size 16x16

- ViT Base models → 1 for each dataset = 3 (Cifar10, Cifar100 & SVHN)
- 12 encoder layers, 12 attention heads
- Theoretically any model possible → modifiable & interchangeable
- Cuda 2 cores, can run in parallel

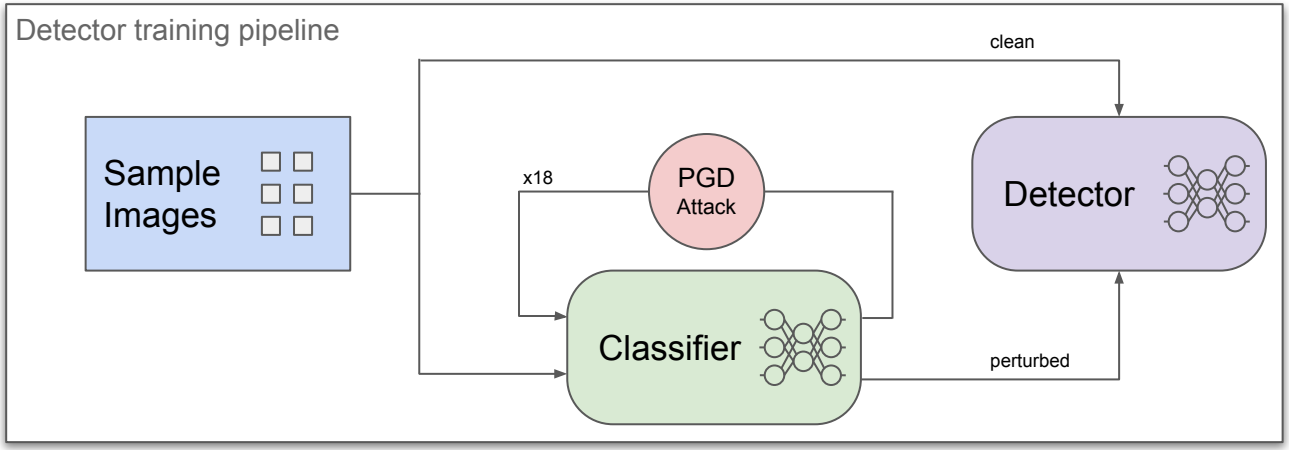


- ViT Tiny models → 1 for each permutation = 6
- 12 encoder layers, 3 attention heads
- Separate model → modifiable & interchangeable
- Cuda 1 core, can not run in parallel

- ε = 0.01 (maximum perturbation), 18 perturbation cycles
- No parallel computing (because of `torch.autograd.grad`)

6. Training

- Classifier: 1 Dataset and 60 epochs → model w/ best accuracy
- Detector: 2 Datasets (ID & OOD) and 3 epochs → model w/ best accuracy
  - 18h/epoch = training ~16h/epoch + validation ~2h/epoch
  - PGD Attack (ε = 0.01, 18 cycles) not executable in parallel



7. Results - Base Runs

ID	OOD					
	Accuracy	AUROC	AUPR	Accuracy	AUROC	AUPR
Cifar 10	Cifar 100			SVHN		
	63.38	72.83	75.64	**99.31**	**99.97**	**99.97**
Cifar 100	Cifar 10			SVHN		
	63.71	73.06	74.08	**98.89**	**99.97**	**99.98**
SVHN	Cifar 10			Cifar 100		
	99.18	**99.99**	**99.99**	98.83	99.97	99.97

ProoD results

Table 4: **Separate training:** Addendum to Table 2 showing the AUCs, GAUCs and AAUCs of ProoD-S on all datasets. The accuracy must always be identical to that of OE and the clean AUCs are also very similar to those of OE. The guarantees are almost always strictly weaker than those provided by the semi-jointly trained ProoD.

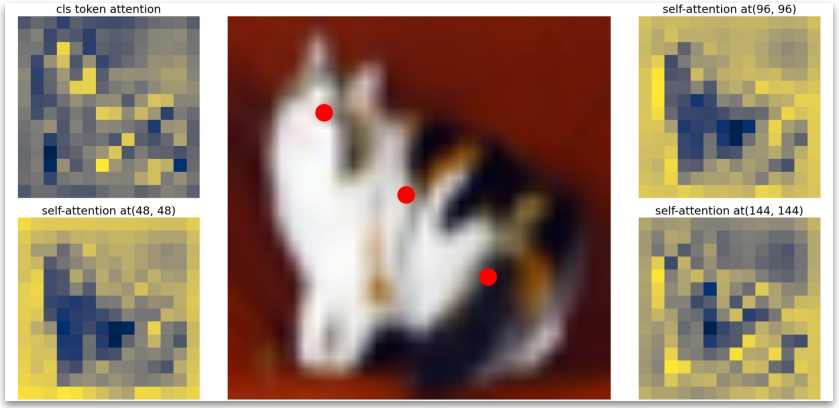
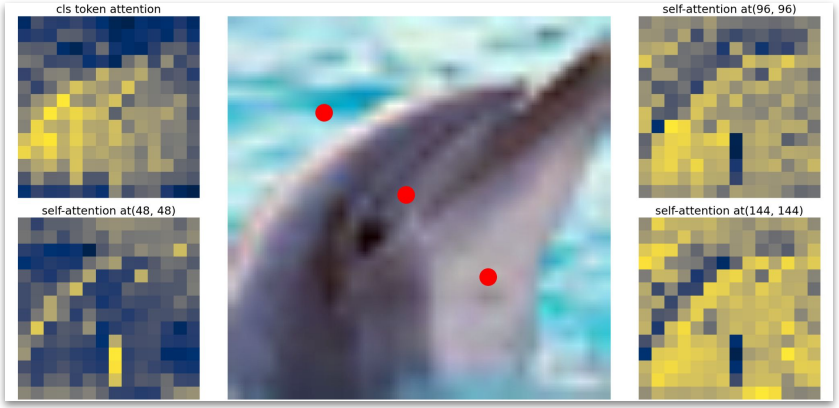
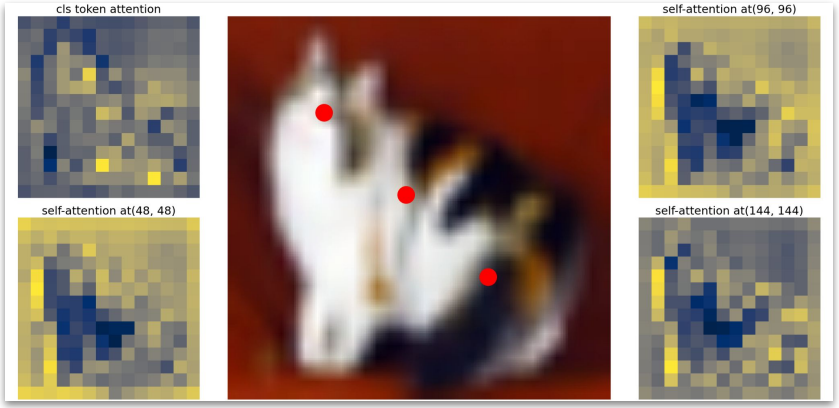
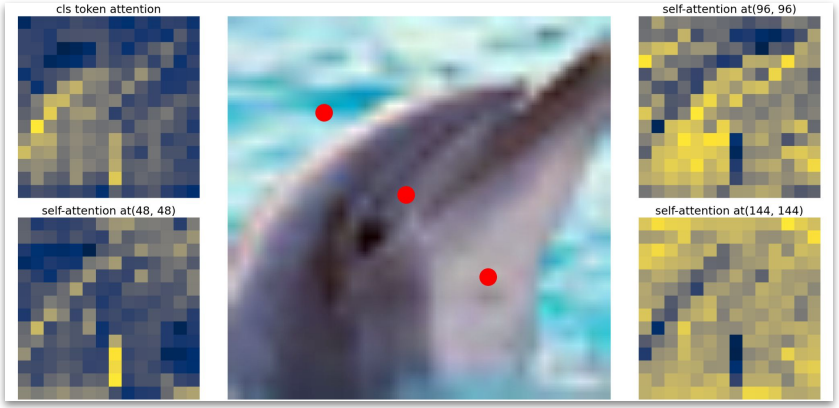
In: CIFAR10		CIFAR100			SVHN			LSUN_CR			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
OE	94.91	**91.1**	0.0	0.9	97.3	0.0	0.0	**100.0**	0.0	2.7	**99.9**	0.0	1.5
ProoD-S $\Delta=3$	94.91	89.3	44.7	45.3	97.3	51.8	52.6	**100.0**	56.7	57.7	**99.9**	36.7	37.6
ProoD $\Delta=3$	**94.99**	89.8	**46.1**	**46.8**	**98.3**	**53.3**	**54.1**	**100.0**	**58.3**	**59.7**	**99.9**	**38.2**	**38.8**
In: CIFAR100		**CIFAR10**			**SVHN**			**LSUN_CR**			**Smooth**		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
OE	**77.25**	**77.4**	0.0	0.2	**92.3**	0.0	0.0	**100.0**	0.0	0.7	**99.5**	0.0	0.5
ProoD-S $\Delta=5$	**77.25**	**77.4**	17.2	17.3	**92.3**	19.5	19.6	**100.0**	22.4	22.6	**99.5**	**9.0**	**9.1**
ProoD $\Delta=5$	77.16	76.6	**17.3**	**17.4**	91.5	**19.7**	**19.8**	**100.0**	**22.5**	**23.1**	98.9	**9.0**	9.0
In: R.ImgNet		**Flowers**			**FGVC**			**Cars**			**Smooth**		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
OE	97.10	**96.9**	0.0	0.2	99.7	0.0	0.4	**99.9**	0.0	1.8	98.0	0.0	1.9
ProoD-S $\Delta=4$	97.10	**96.9**	50.1	50.7	99.7	59.7	60.6	**99.9**	57.9	58.9	98.0	40.8	42.3
ProoD $\Delta=4$	**97.25**	**96.9**	**57.5**	**58.0**	**99.8**	**67.4**	**67.9**	**99.9**	**65.7**	**66.2**	**98.6**	**52.7**	**53.5**

Meinke et al., "Provably Robust Detection of Out-of-distribution Data (almost) for free", https://arxiv.org/abs/2106.04260

7. Results - Additional Runs

6 epochs	OOD – Cifar 100		
ID	Accuracy	AUROC	AUPR
Cifar 10	65.40	76.33	78.53
	Comparison w/ 3 epochs		
	63.38	72.83	75.64
	ProoD w/ 100 epochs		
	–	**89.8**	–

ε = 8/255	OOD – SVHN		
ID	Accuracy	AUROC	AUPR
Cifar 100	98.83	99.96	99.97
	Comparison w/ ε = 0.01		
	98.89	**99.97**	**99.98**
	ProoD w/ ε = 8/255		
	–	91.5	–

	ID Cifar 10 OOD Cifar 100; 12th layer; ε = 0.01	ID Cifar100 OOD SVHN; 12th layer; ε = 8/255 (≈0.03137)
clean		
perturbed		

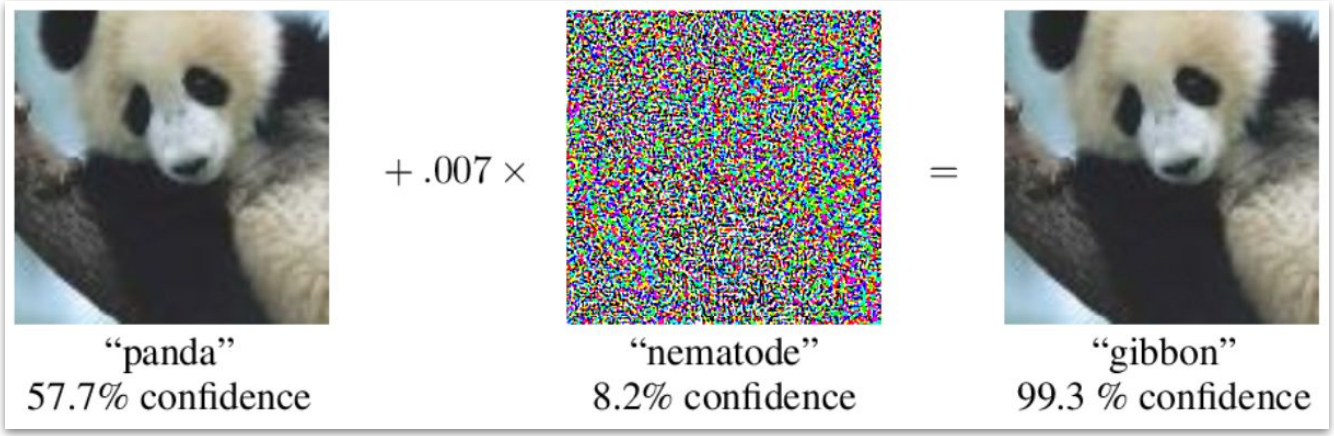
8. Future Work

- Code optimizations, maybe speedup/parallelization of the PGD attack
- Compare different models and different ViT sizes as detectors
- More exhaustive hyperparameter search (different eps, different noise generators, more restarts/iterations)
- Mixing of OOD datasets for more variety in the OOD samples
- “How much noise can be applied to an ID/OOD sample until its label should change as well?”
  → Human judge? Machine judge?
- …

Thank you,
for your attention.

Robust Adversarial OOD
Detection With Vision Transformer

Backup - Adversarial Attacks - General



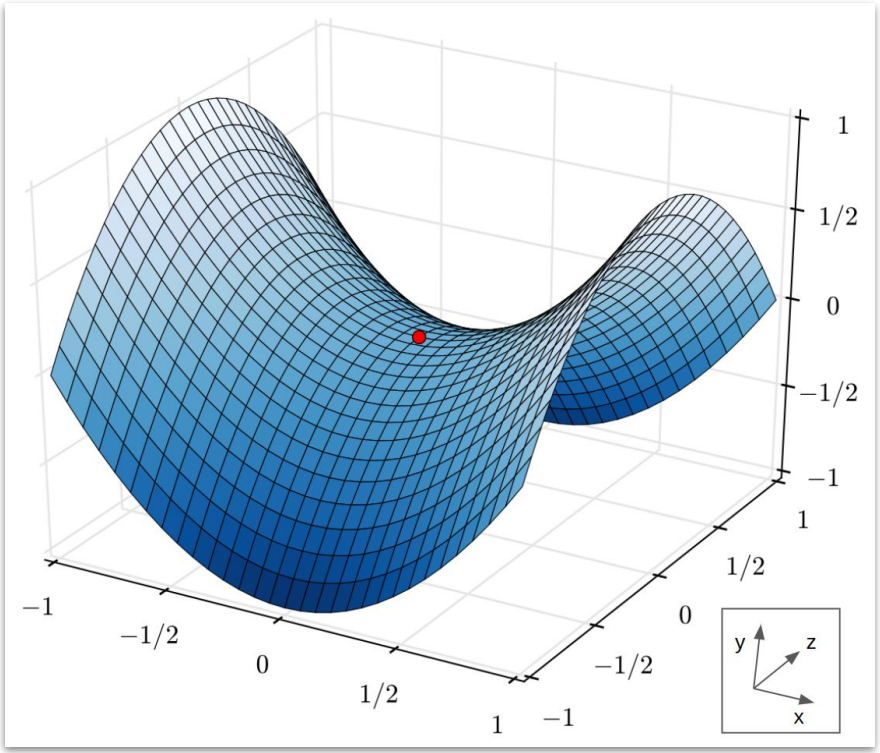
Goodfellow et al., 2014, “Explaining and Harnessing Adversarial Examples”

random noise

Backup

Elasyed et al., 2018, “Adversarial Examples that Fool both Human and Computer Vision”

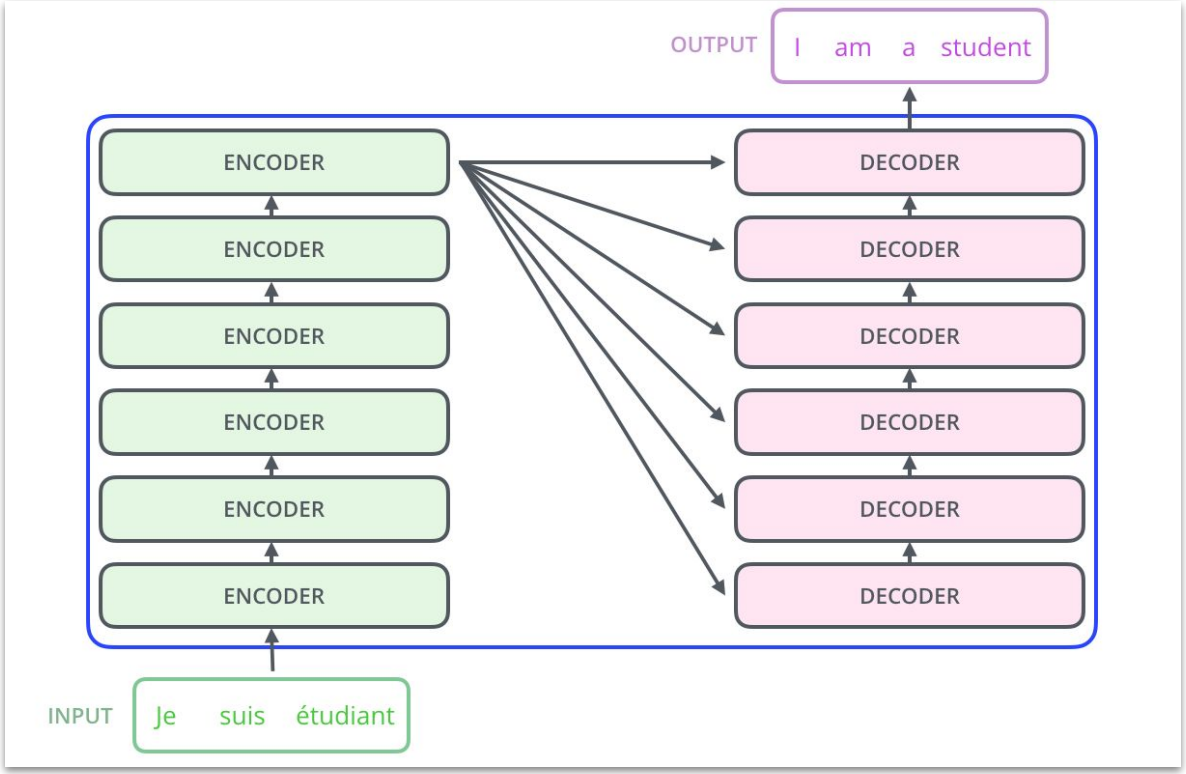
Backup



https://commons.wikimedia.org/wiki/File:Saddle_point.svg

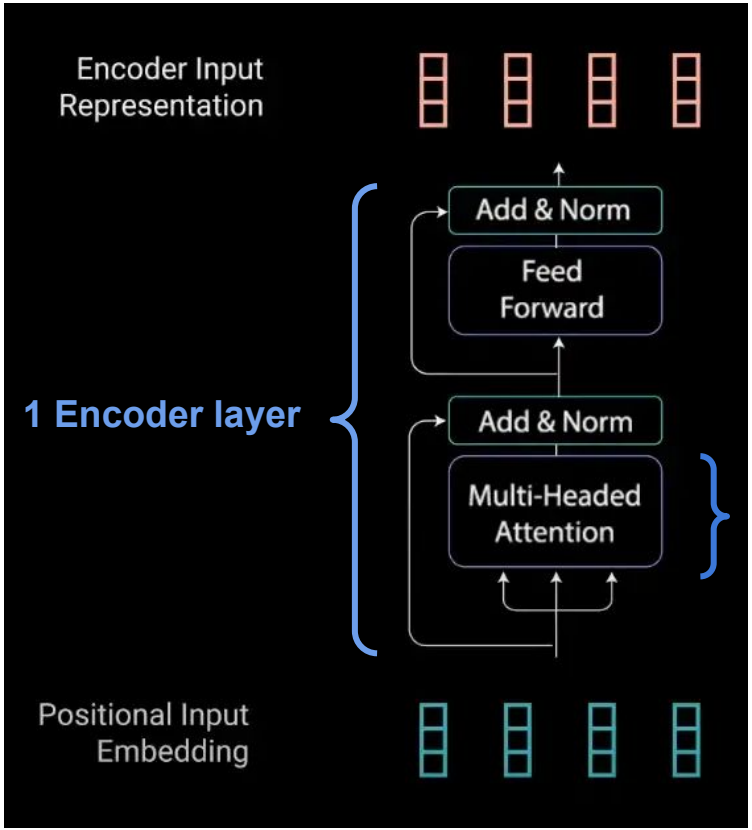
Oscar Knagg, 2019, “Know your enemy - How you can create and defend against adversarial attacks”

Backup

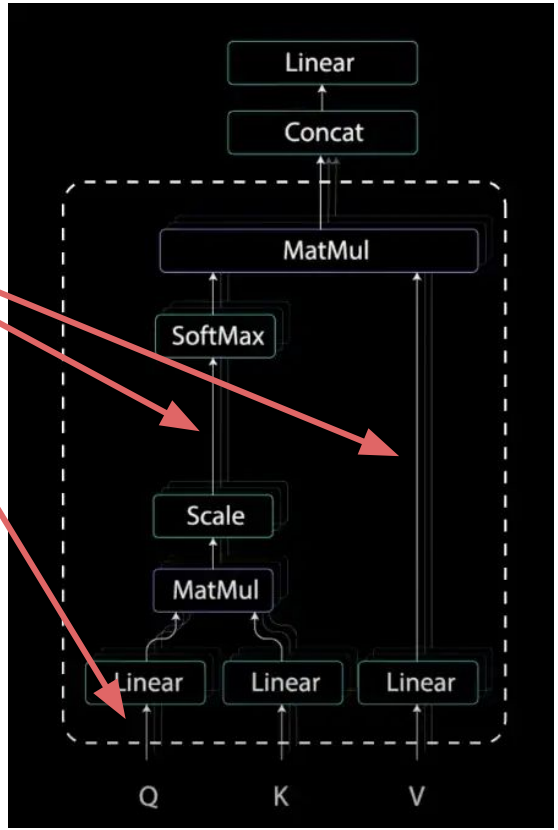


https://jalammar.github.io/illustrated-transformer/

Backup - Layer & Multi-Headed Attention

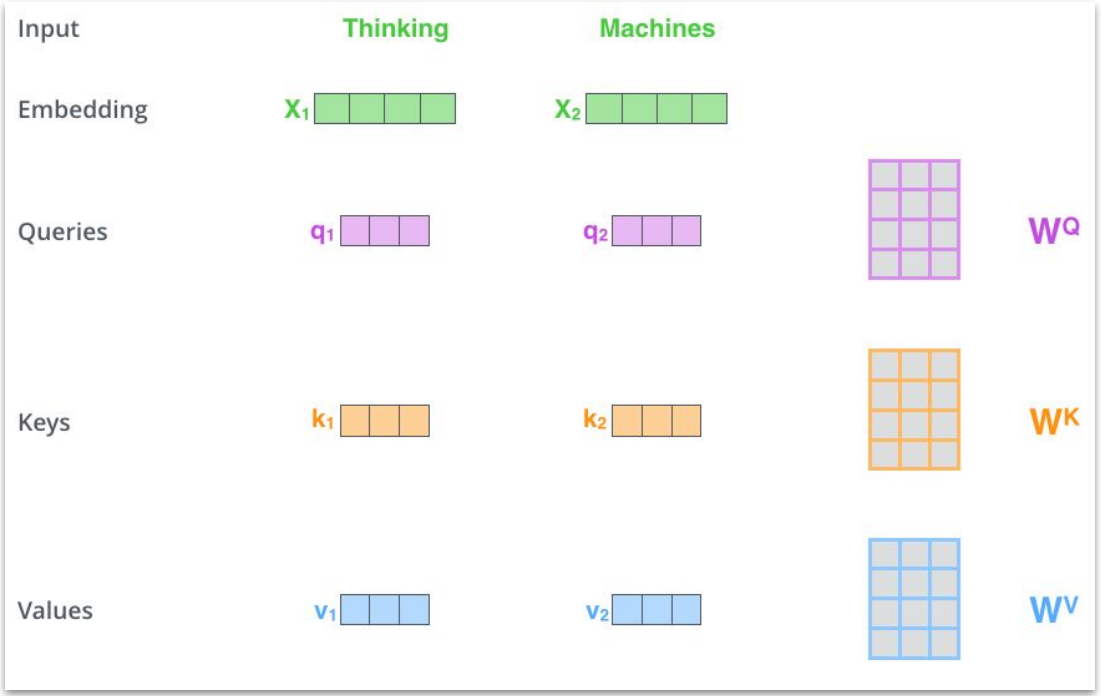


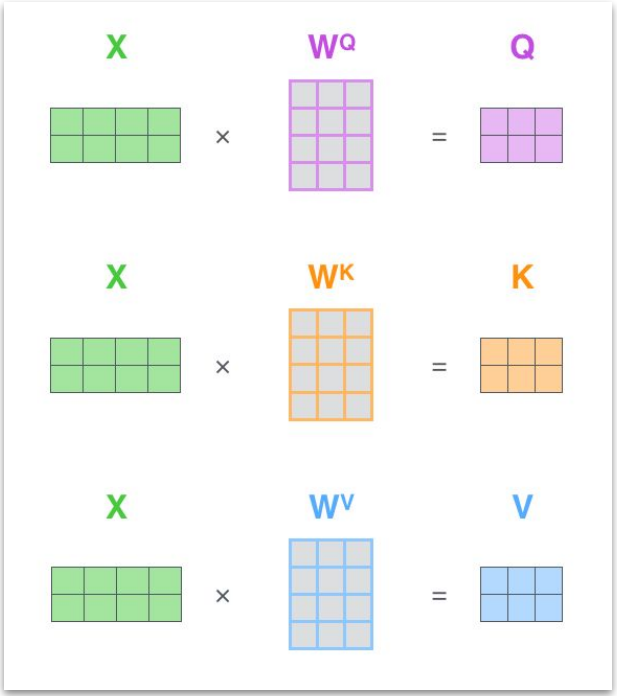
Multiple arrows for multiple attention-heads



https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

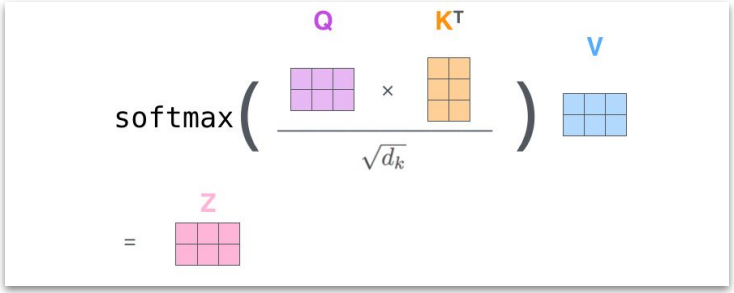
Backup - Attention Mechanism (1/2)

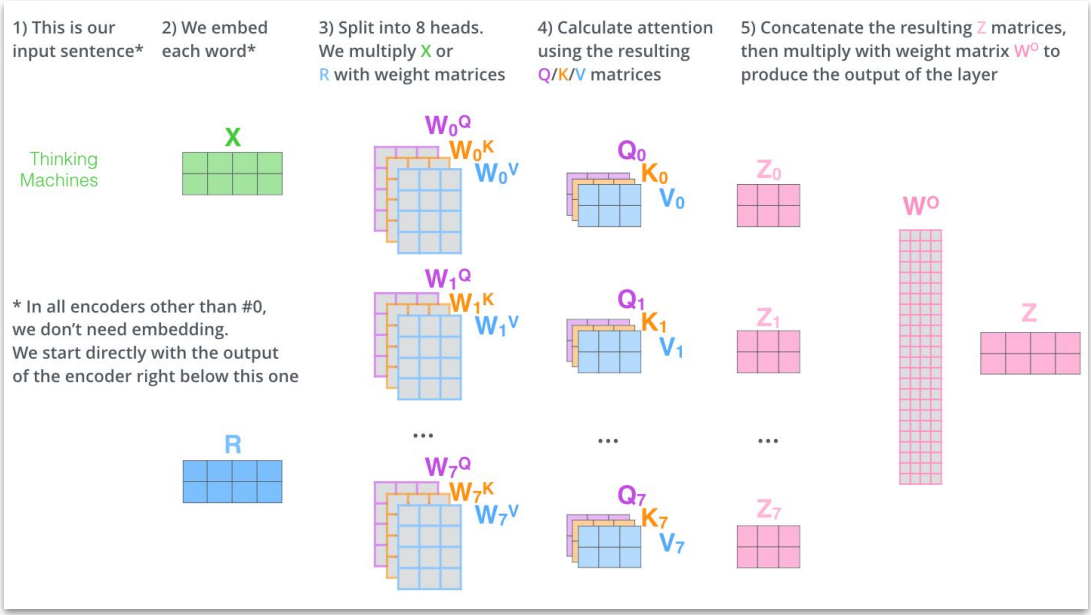




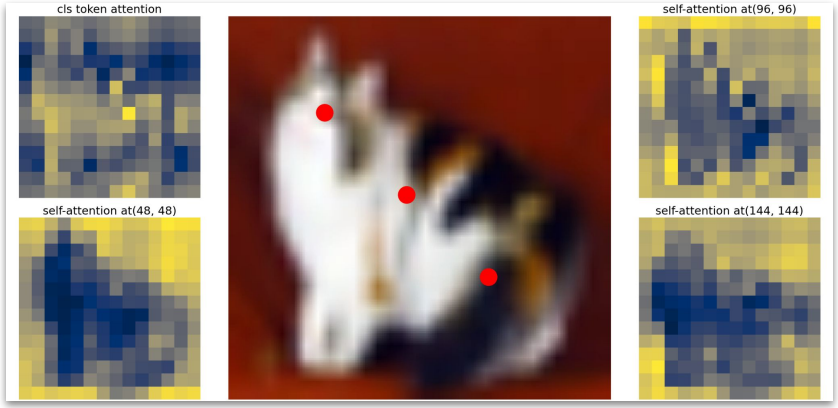
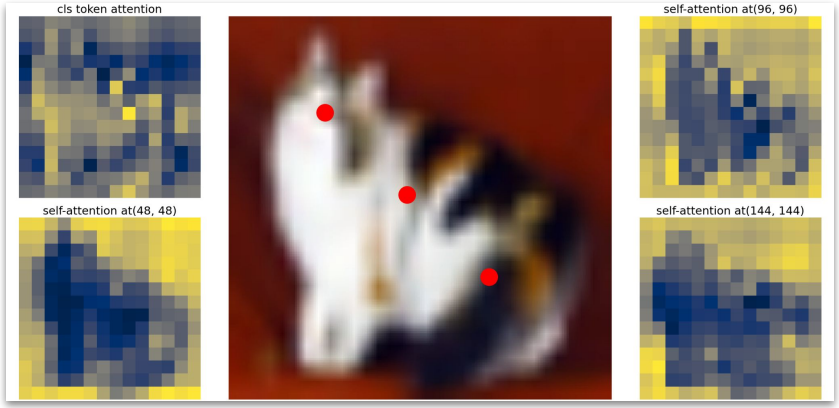
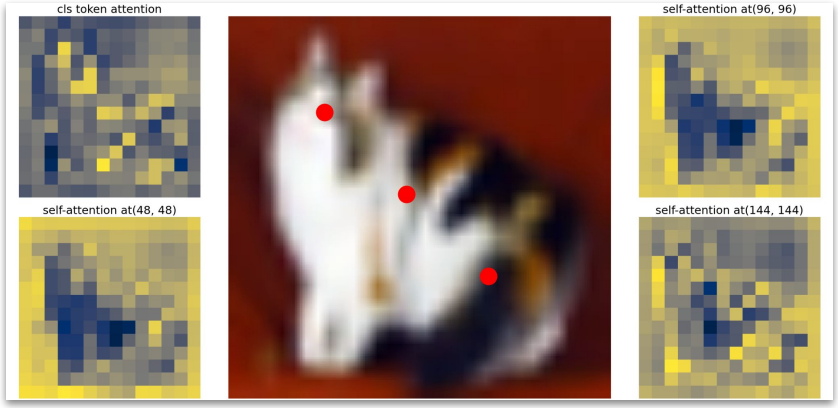
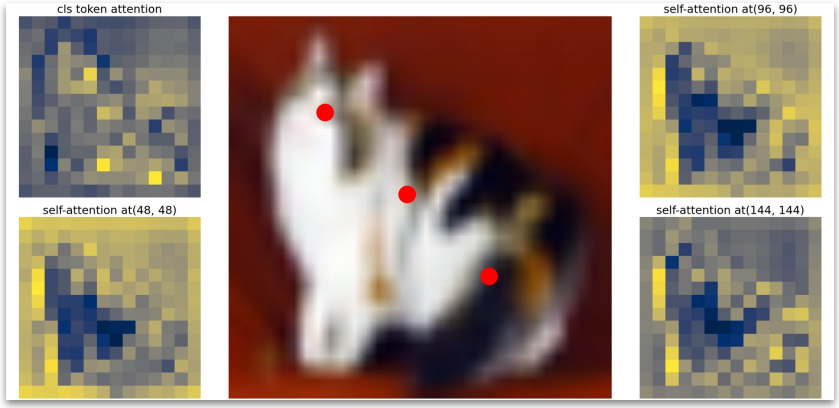
https://jalammar.github.io/illustrated-transformer/

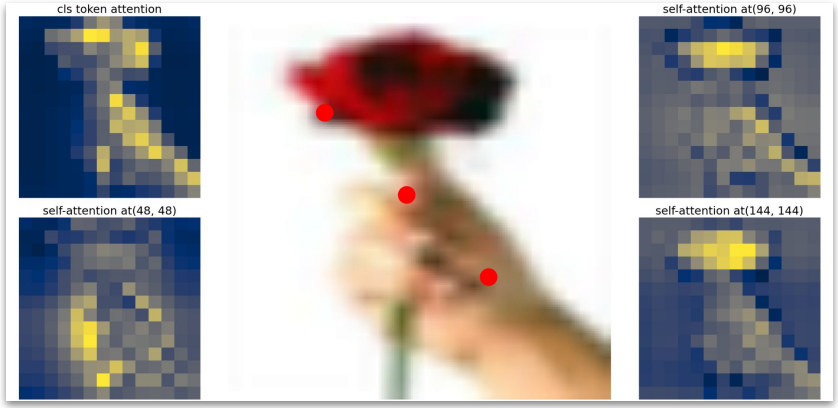
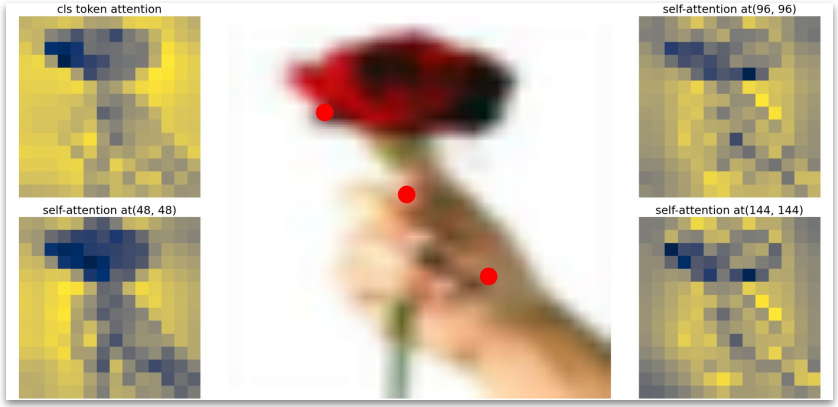
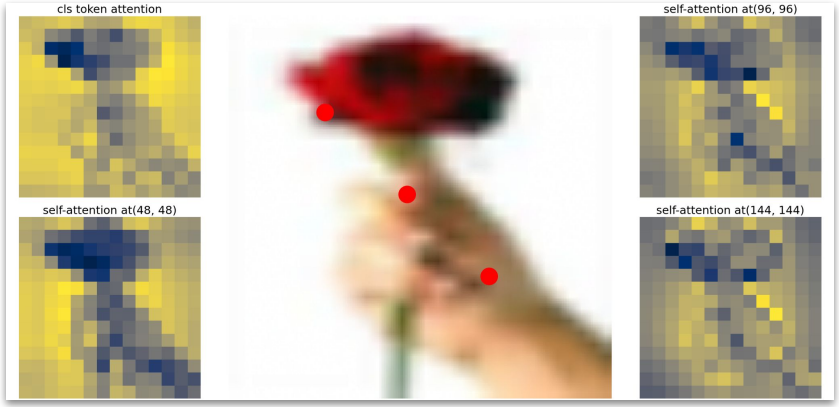
Backup - Attention Mechanism (2/2)

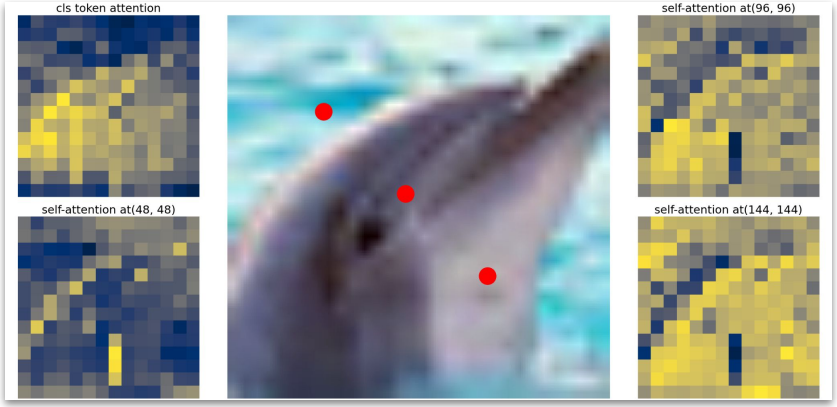
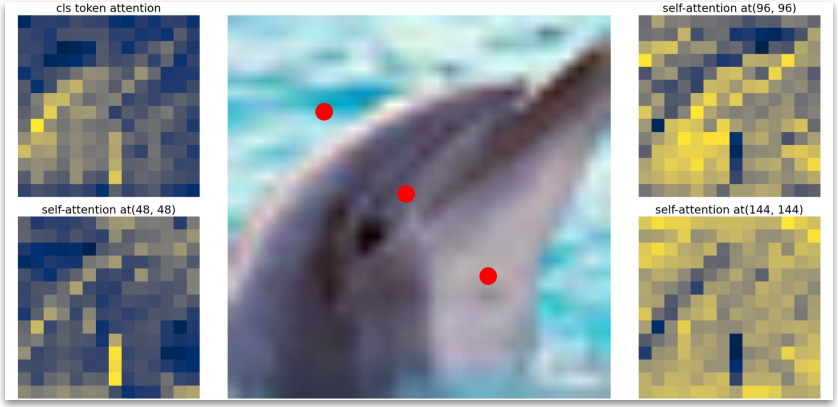




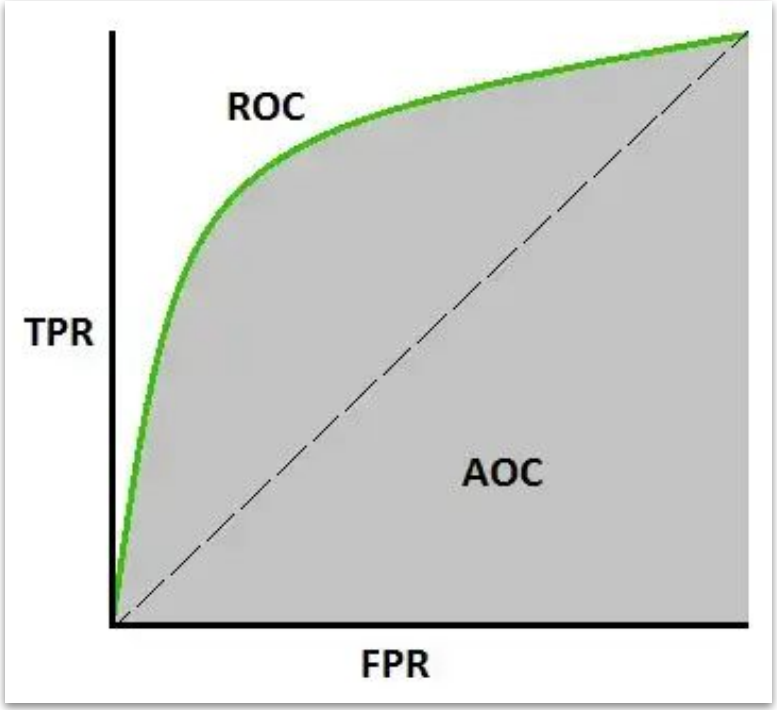
https://jalammar.github.io/illustrated-transformer/

ID	clean	perturbed
1st layer		
12th layer		

OOD	clean	perturbed
5th layer		
10th layer		

ε = 8/255	clean	perturbed
12th layer – ID		
2nd layer – OOD		

Backup - AUROC



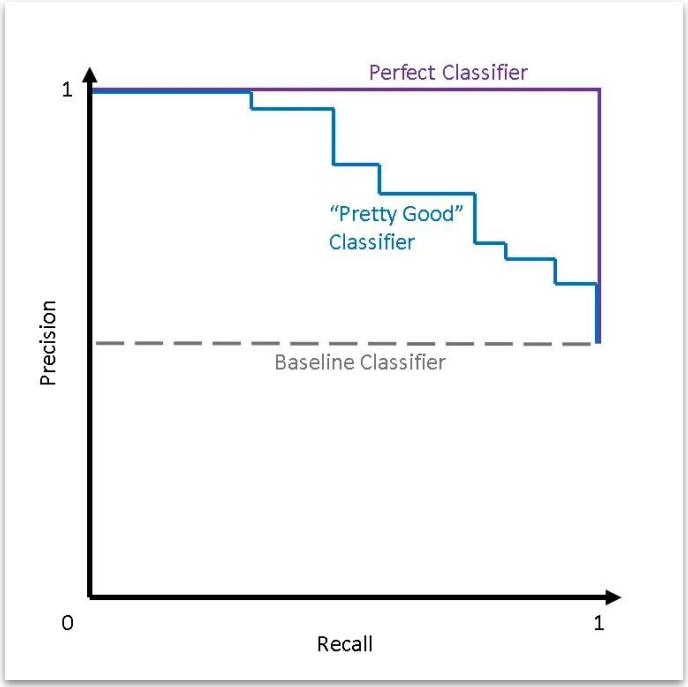
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

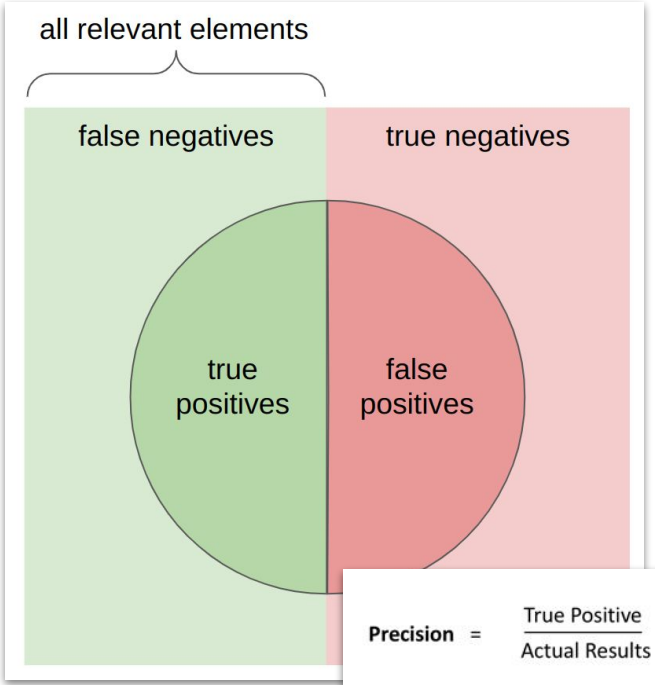
$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

Backup - AUPR



https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248

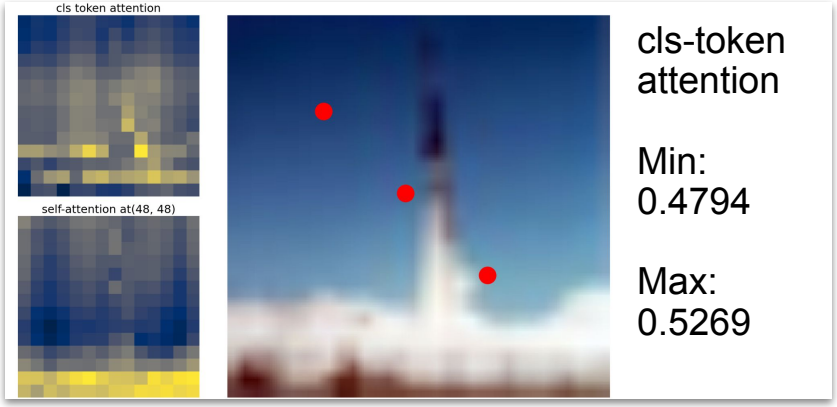
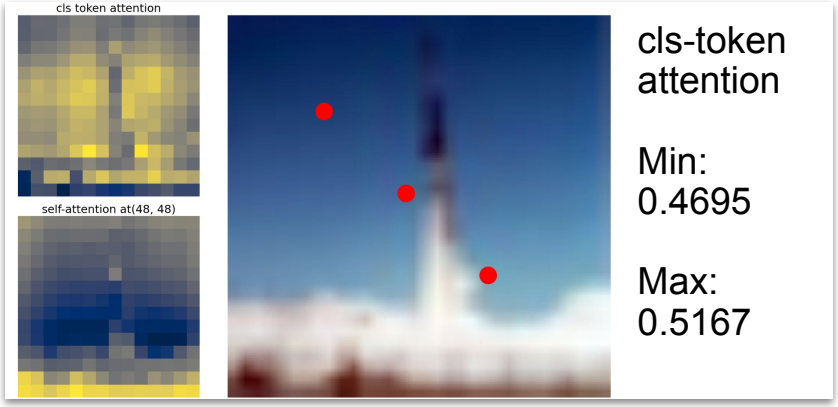


https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488

Precision = $\frac{\text{True Positive}}{\text{Actual Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

Recall = $\frac{\text{True Positive}}{\text{Predicted Results}}$ or $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{Total}}$

	clean	perturbed
12th layer - ID SVHN		
12th layer - OOD Cifar100		

Backup - Attention Differences

[0;1]

1 = all the attention is always on one specific patch
0 = no attention on one patch

ID, OOD	ID span	perturbed ID span	OOD span	perturbed OOD span
Cifar 10, Cifar 100	0.0475	0.0569	0.0504	0.0599
Cifar 10, SVHN	0.5179	0.5891	0.1574	0.2765
Cifar 100, Cifar 10	0.0324	0.0356	0.0313	0.0352
Cifar 100, SVHN	0.1864	0.2225	0.2505	0.2080
SVHN, Cifar 10	0.1360	0.1408	0.3500	0.4011
SVHN, Cifar 100	0.3397	0.3709	0.3145	0.3437
Cifar 10, Cifar 100 6 epochs	0.0561	0.0627	0.0485	0.0540
Cifar 100, SVHN Eps = 8/255	0.2461	0.2776	0.3366	0.2844

Backup - Results - Base Runs

ID	OOD					
	Accuracy	AUROC	AUPR	Accuracy	AUROC	AUPR
Cifar 10	Cifar 100			SVHN		
	63.38	72.83	75.64	**99.31**	**99.97**	**99.97**
Cifar 100	Cifar 10			SVHN		
	63.71	73.06	74.08	**98.89**	**99.97**	**99.98**
SVHN	Cifar 10			Cifar 100		
	99.18	**99.99**	**99.99**	98.83	99.97	99.97

ProoD results

In: CIFAR10		CIFAR100			SVHN			LSUN_CR			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	**95.01**	90.0	0.0	0.7	93.8	0.0	0.3	93.1	0.0	0.5	98.0	0.0	0.7
OE	94.91	**91.1**	0.0	0.9	97.3	0.0	0.0	**100.0**	0.0	2.7	**99.9**	0.0	1.5
ATOM	93.63	78.3	0.0	21.7	94.4	0.0	24.1	79.8	0.0	20.1	99.5	0.0	**73.2**
ACET	93.43	86.0	0.0	4.0	99.3	0.0	4.6	89.2	0.0	3.7	99.9	0.0	40.2
$GOOD_{80}$*	87.39	76.7	47.1	57.1	90.8	43.4	76.8	97.4	70.6	93.6	96.2	72.9	89.9
$GOOD_{100}$*	86.96	67.8	48.1	49.7	62.6	34.9	36.3	84.9	74.6	75.6	87.0	76.1	78.1
ProoD-Disc	-	62.9	57.1	57.8	72.6	65.6	66.4	78.1	71.5	72.3	59.2	49.7	50.4
ProoD $\Delta=3$	94.99	89.8	**46.1**	**46.8**	**98.3**	**53.3**	**54.1**	**100.0**	**58.3**	**59.7**	**99.9**	**38.2**	38.8
In: CIFAR100		CIFAR10			SVHN			LSUN_CR			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	**77.38**	**77.7**	0.0	0.4	81.9	0.0	0.2	76.4	0.0	0.3	86.6	0.0	0.4
OE	77.25	77.4	0.0	0.2	**92.3**	0.0	0.0	**100.0**	0.0	0.7	**99.5**	0.0	0.5
ATOM	68.32	78.3	0.0	50.3	91.1	0.0	67.0	95.9	0.0	75.6	98.2	0.0	80.7
ACET	73.02	73.0	0.0	1.4	97.8	0.0	0.7	75.8	0.0	2.6	99.9	0.0	12.8
ProoD-Disc	-	56.1	52.1	52.3	61.0	58.2	58.4	70.4	66.9	67.1	29.6	26.4	26.5
ProoD $\Delta=5$	77.16	76.6	**17.3**	**17.4**	91.5	**19.7**	**19.8**	**100.0**	**22.5**	**23.1**	98.9	**9.0**	**9.0**
In: R.ImgNet		Flowers			FGVC			Cars			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	96.34	92.3	0.0	0.5	92.6	0.0	0.0	92.7	0.0	0.1	**98.9**	0.0	8.6
OE	97.10	**96.9**	0.0	0.2	99.7	0.0	0.4	**99.9**	0.0	1.8	98.0	0.0	1.9
ProoD-Disc	-	81.5	76.8	77.3	92.8	89.3	89.6	90.7	86.9	87.3	81.0	74.0	74.8
ProoD $\Delta=4$	**97.25**	**96.9**	**57.5**	**58.0**	**99.8**	**67.4**	**67.9**	**99.9**	**65.7**	**66.2**	98.6	**52.7**	**53.5**

*Uses different architecture of classifier, see "Baselines" in Section 4.2.

Meinke et al., "Provably Robust Detection of Out-of-distribution Data (almost) for free", https://arxiv.org/abs/2106.04260

Backup

ε = 8/255 (0.03137...)

Table 9: **Generalization to Larger ϵ:** We evaluate all CIFAR models in Table 2 using an $\epsilon = \frac{8}{255}$, and thus an unseen threat model. The provable methods GOOD and ProoD generalize surprisingly well, while neither ATOM nor ACET display any generalization to the larger threat model. 1

In: CIFAR10		CIFAR100			SVHN			LSUN_CR			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	**95.01**	90.0	0.0	0.0	93.8	0.0	0.0	93.1	0.0	0.0	98.0	0.0	0.0
OE	94.91	**91.1**	0.0	0.1	97.3	0.0	0.0	**100.0**	0.0	0.1	**99.9**	0.0	0.0
ATOM	93.63	78.3	0.0	1.3	94.4	0.0	1.5	79.8	0.0	0.2	99.5	0.0	9.6
ACET	93.43	86.0	0.0	1.1	**99.3**	0.0	1.1	89.2	0.0	0.8	**99.9**	0.0	3.8
GOOD80*	87.39	76.7	37.5	51.6	90.8	38.6	74.3	97.4	57.6	90.2	96.2	61.1	87.8
GOOD100*	86.96	67.8	39.4	43.5	62.6	29.0	30.9	84.9	67.6	70.7	87.0	63.3	69.2
ProoD-Disc	-	62.9	44.1	46.1	72.6	52.5	57.1	78.1	56.3	58.9	59.2	34.9	37.2
ProoD $\Delta=3$	94.99	89.8	**39.2**	**41.0**	98.3	**46.9**	**50.8**	**100.0**	**50.2**	**52.7**	**99.9**	**30.4**	**30.6**

In: CIFAR100		CIFAR10			SVHN			LSUN_CR			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	**77.38**	**77.7**	0.0	0.4	81.9	0.0	0.2	76.4	0.0	0.3	86.6	0.0	0.3
OE	77.25	77.4	0.0	0.2	**92.3**	0.0	0.0	**100.0**	0.0	0.7	**99.5**	0.0	0.5
ATOM	68.32	78.3	0.0	10.4	91.1	0.0	15.2	95.9	0.0	23.0	98.2	0.0	23.5
ACET	73.02	73.0	0.0	1.4	97.8	0.0	0.7	75.8	0.0	2.6	99.9	0.0	3.8
ProoD-Disc	-	56.1	41.1	43.1	61.0	50.5	51.8	70.4	57.5	58.8	29.6	20.9	20.8
ProoD $\Delta=5$	76.51	76.6	**13.7**	**14.1**	91.5	**16.9**	**16.9**	**100.0**	**18.1**	**18.2**	98.9	**8.1**	**8.1**

In: R.ImgNet		Flowers			FGVC			Cars			Smooth		
	Acc	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC	AUC	GAUC	AAUC
Plain	96.34	92.3	0.0	0.0	92.6	0.0	0.0	92.7	0.0	0.0	**98.9**	0.0	0.0
OE	97.10	**96.9**	0.0	0.2	99.7	0.0	0.0	**99.9**	0.0	0.0	98.0	0.0	0.0
ProoD-Disc	-	81.5	60.4	61.4	92.8	78.0	80.8	90.7	76.3	79.2	81.0	47.3	53.7
ProoD $\Delta=4$	**97.25**	**96.9**	**42.8**	**45.0**	**99.8**	**57.0**	**59.4**	**99.9**	**56.0**	**58.7**	98.6	**31.6**	**36.3**

*Uses different architecture of classifier, see "Baselines" in Section 4.2.

Meinke et al., "Provably Robust Detection of Out-of-distribution Data (almost) for free", https://arxiv.org/abs/2106.04260