Detection of AI-generated images

Lucas SALAND
Supervisor: Patrick Bas

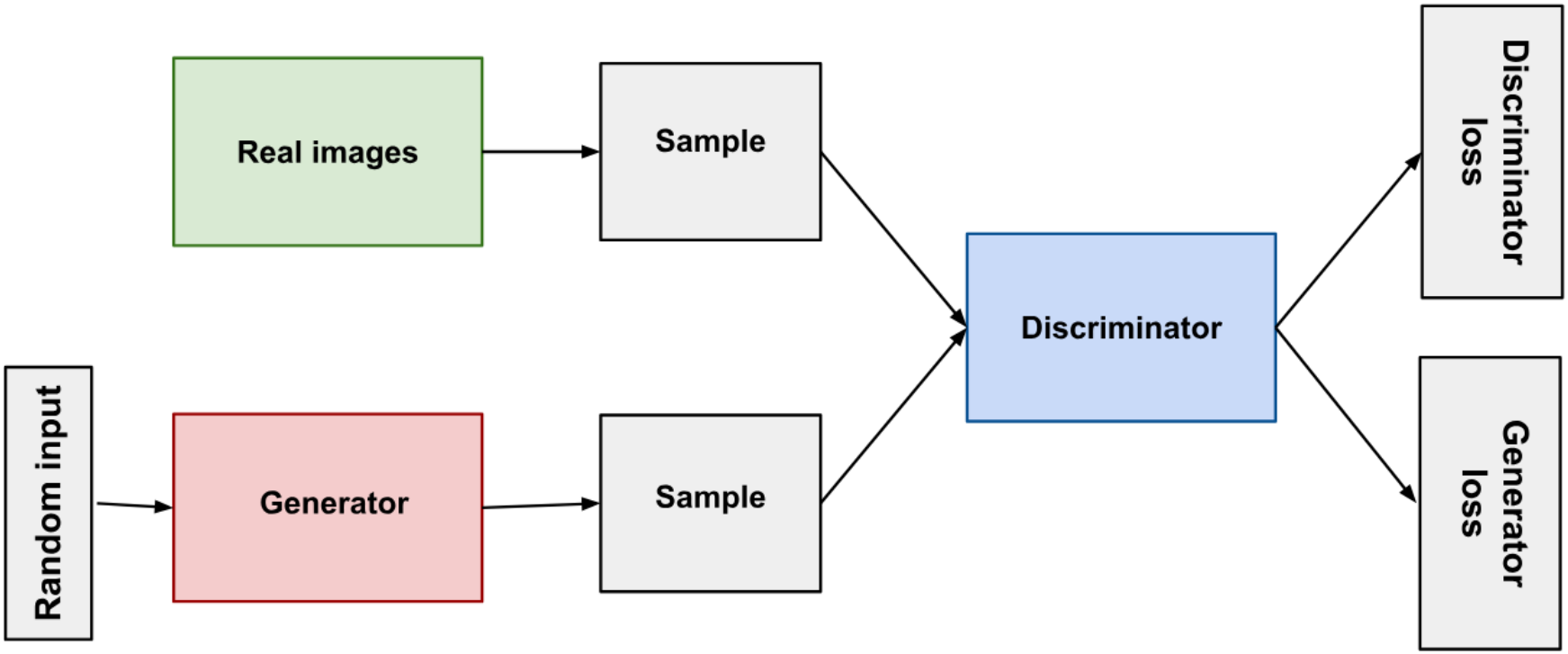
Université de Lille

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

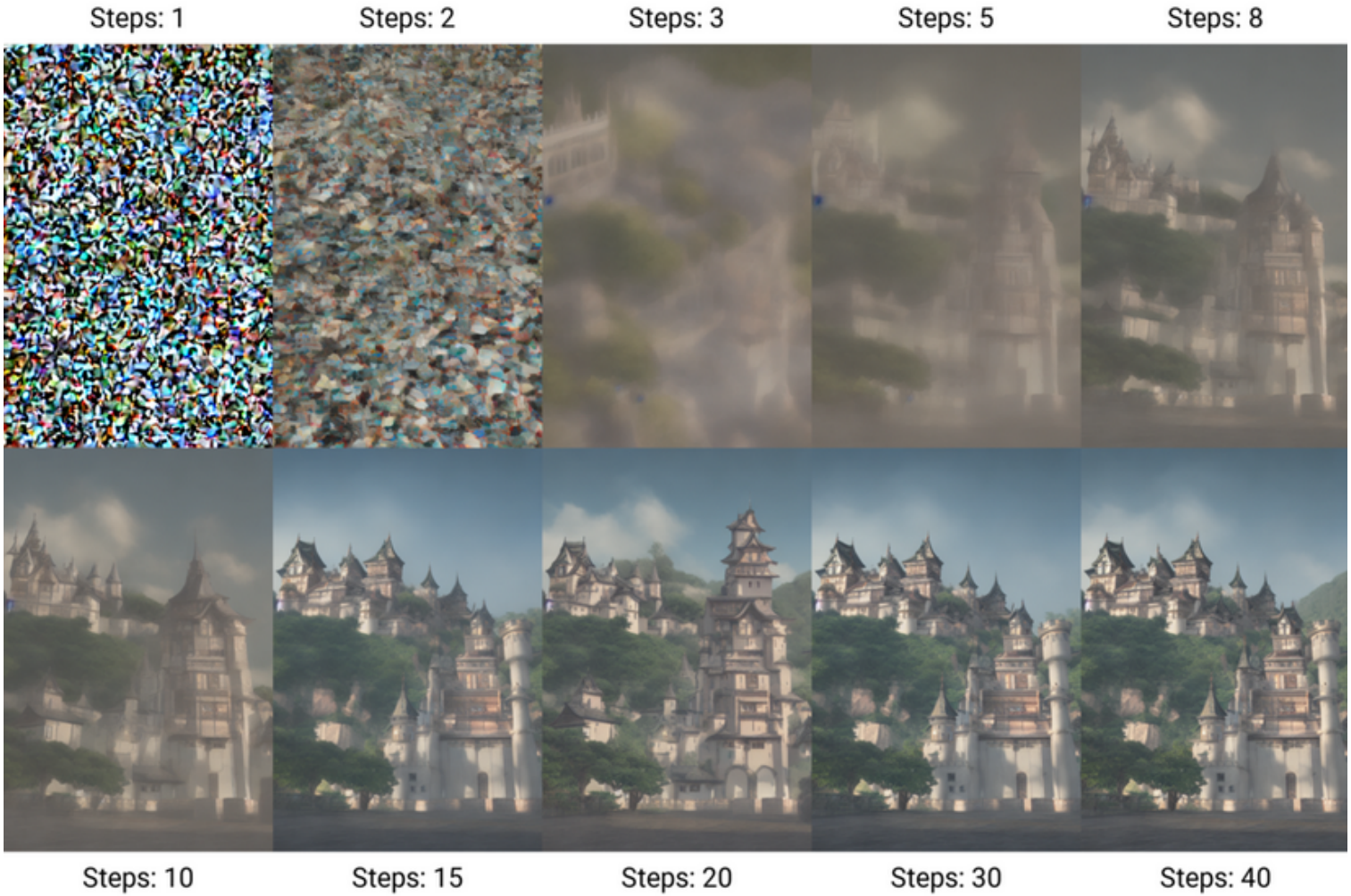
Lille, France

August 17, 2024

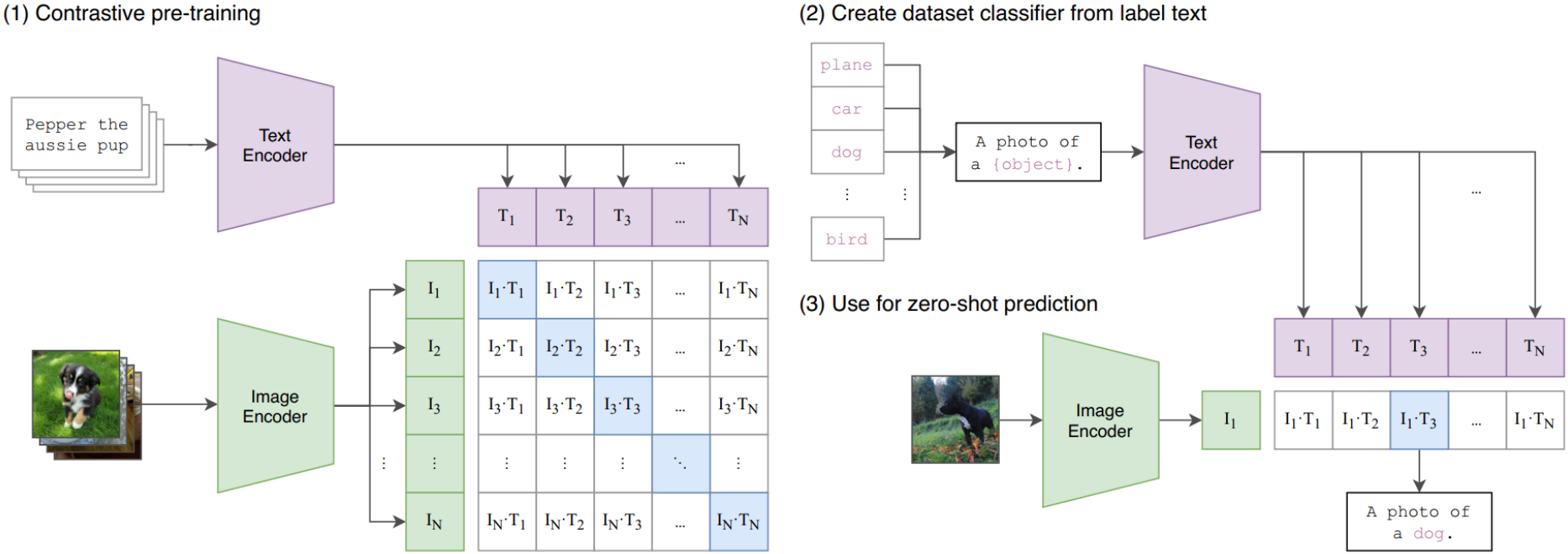
GAN


Real images
Sample
Discriminator loss
Random input
Generator
Sample
Discriminator
Generator loss

Diffusion model



CLIP


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
A photo of a dog.

CLIP for detection

Method proposed in:[1]

[1] Cozzolino et al., *Raising the Bar of AI-generated Image Detection with CLIP.*

CLIP for detection

Method proposed in:[1]

- Build a dataset of pairs of real and generated images

[1] Cozzolino et al., *Raising the Bar of AI-generated Image Detection with CLIP*.

CLIP for detection

Method proposed in:[1]

- Build a dataset of pairs of real and generated images
- Extract the CLIP features

[1] Cozzolino et al., *Raising the Bar of AI-generated Image Detection with CLIP*.

CLIP for detection

Method proposed in:[1]

- Build a dataset of pairs of real and generated images
- Extract the CLIP features
- Train a SVM on these features

[1] Cozzolino et al., *Raising the Bar of AI-generated Image Detection with CLIP*.

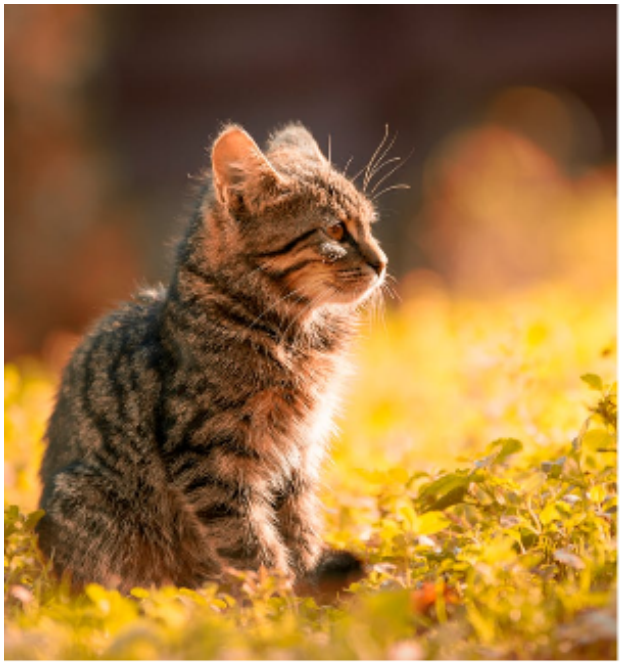
ELSA dataset

An image of 025 Greece Hellenic Air Force Lockheed Martin F16D Fighting Falcon, Medium shot, Ring light, At night, Wideangle lens

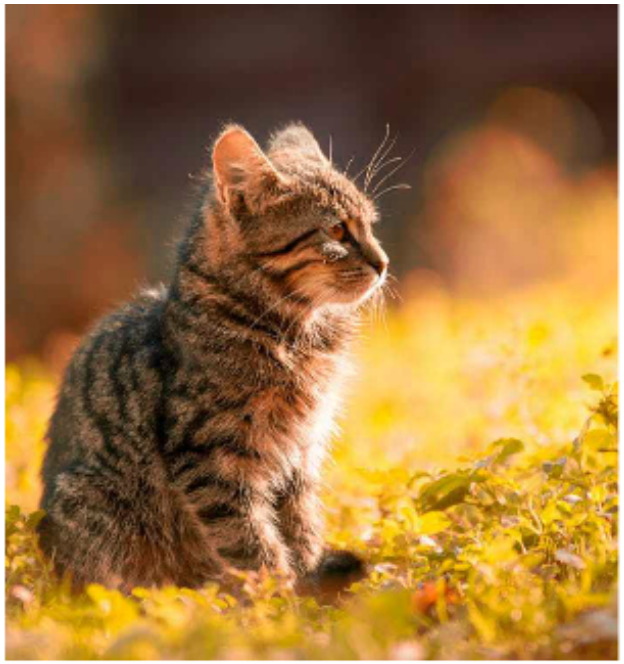
Real

Generated

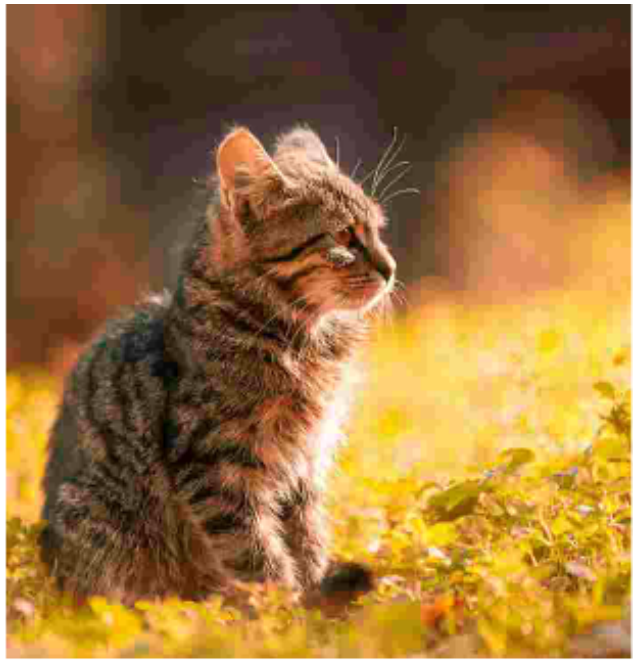
JPEG compression's impact



Original image



JPEG quality 40 (low)



JPEG quality 10 (very low)

JPEG compression's impact

	Test			
	quality	40	65	90
Train	40	0.9813	0.9730	0.9797
	65	0.9450	0.9825	0.9830
	90	0.7744	0.8581	0.9925

Table 1: Accuracy of binary classification for pairs of qualtity factors for train and test.

Generators diversity and neural network

Synthbuster:

- 9 generators
- 1000 images per generator

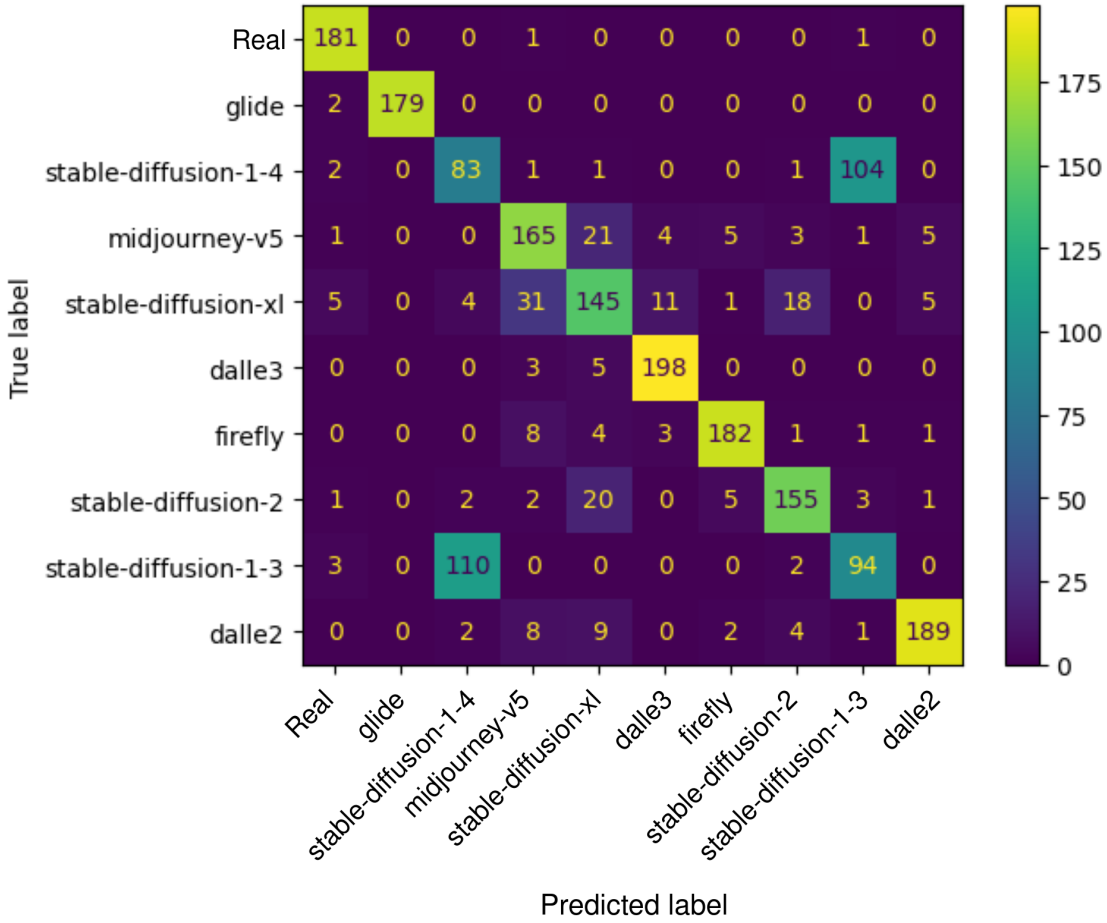
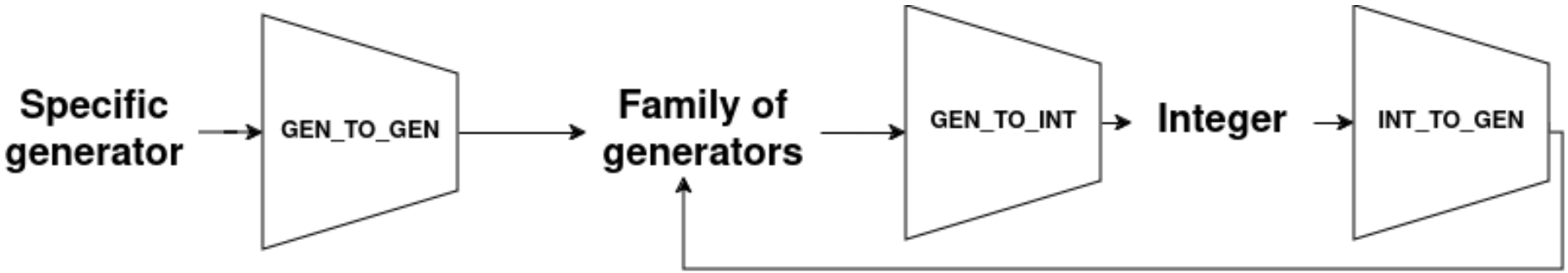


Figure 1: Confusion matrix for a SVM multiclass-classifier trained and tested on synth-

Generators diversity and neural network

Multi-class classifier for binary classification:



SVM vs neural network

Model	Binary classification accuracy	Multi-class classification Accuracy
SVM	0.970	0.873
Neural network	**0.974**	**0.878**

Table 2: Comparison between SVM and neural network for accuracy in classification task.

SVM vs neural network

Model	Binary classification accuracy	Multi-class classification Accuracy
SVM	0.970	0.873
Neural network	**0.974**	**0.878**

Table 2: Comparison between SVM and neural network for accuracy in classification task.

- Code of the detector in a single file

SVM vs neural network

Model	Binary classification accuracy	Multi-class classification Accuracy
SVM	0.970	0.873
Neural network	**0.974**	**0.878**

Table 2: Comparison between SVM and neural network for accuracy in classification task.

- Code of the detector in a single file
- Use Pytorch's dataset library

SVM vs neural network

Model	Binary classification accuracy	Multi-class classification Accuracy
SVM	0.970	0.873
Neural network	**0.974**	**0.878**

Table 2: Comparison between SVM and neural network for accuracy in classification task.

- Code of the detector in a single file
- Use Pytorch's dataset library

$\Longrightarrow$ Accelerate development

Multi-class classifier for binary detection vs binary classifier

The issue: GEN_TO_GEN map needs to be updated frequently

Binary detector accuracy: **0.71**
Multi-class binary accuracy: **0.73**

$\Rightarrow$ Drop the multi-class classifier to accelerate development.

Cleaning the data

Repetitivity in the semantic of generated images → removed generators with repetitivity.

Figure 2: Example of repetition in the content of images. The images are not the same but their semantic is very similar.

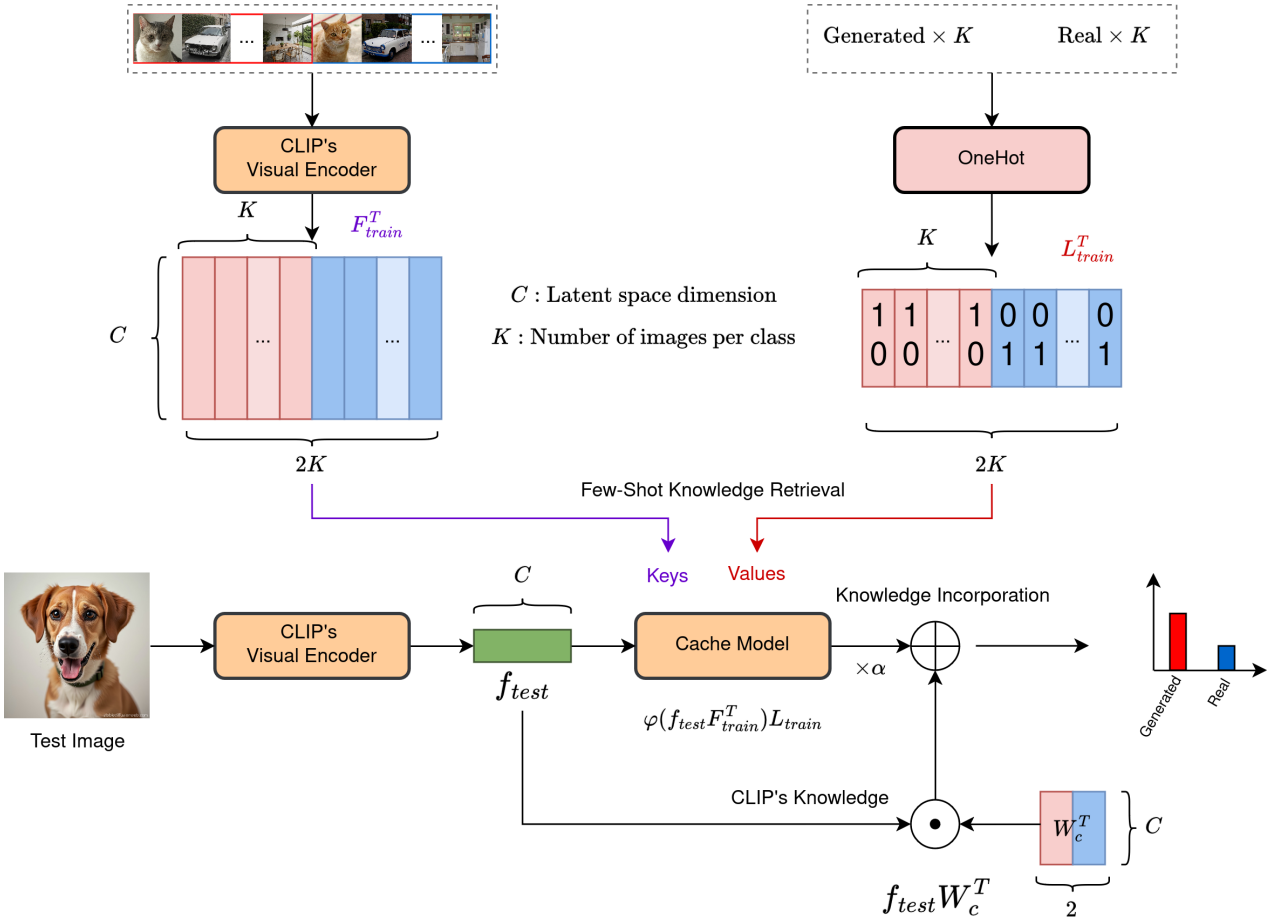
Pair training and fine-tuning

Pair dataset: 1000 pairs of real and generated images + 1000 generated images from AID and 1000 real images from Flickr.
test_meta: Real images from Flickr and generated images from 19 generators.

accuracy on OOD before fine-tuning on test_meta	0.83
accuracy on OOD after fine-tuning on test_meta	0.91
accuracy on test meta before fine-tuning on test_meta	0.79
accuracy on test meta after fine-tuning on test_meta	0.97

Table 3: Accuracy comparison before and after fine-tuning.

Tip-Adapter



$$logits = \alpha \underbrace{\varphi(f_{test}F^T_{train})L_{train}}_{\text{Few-shot knowledge}} + \underbrace{f_{test}W^T_c}_{\text{prior knowledge of pretrained CLIP}}$$

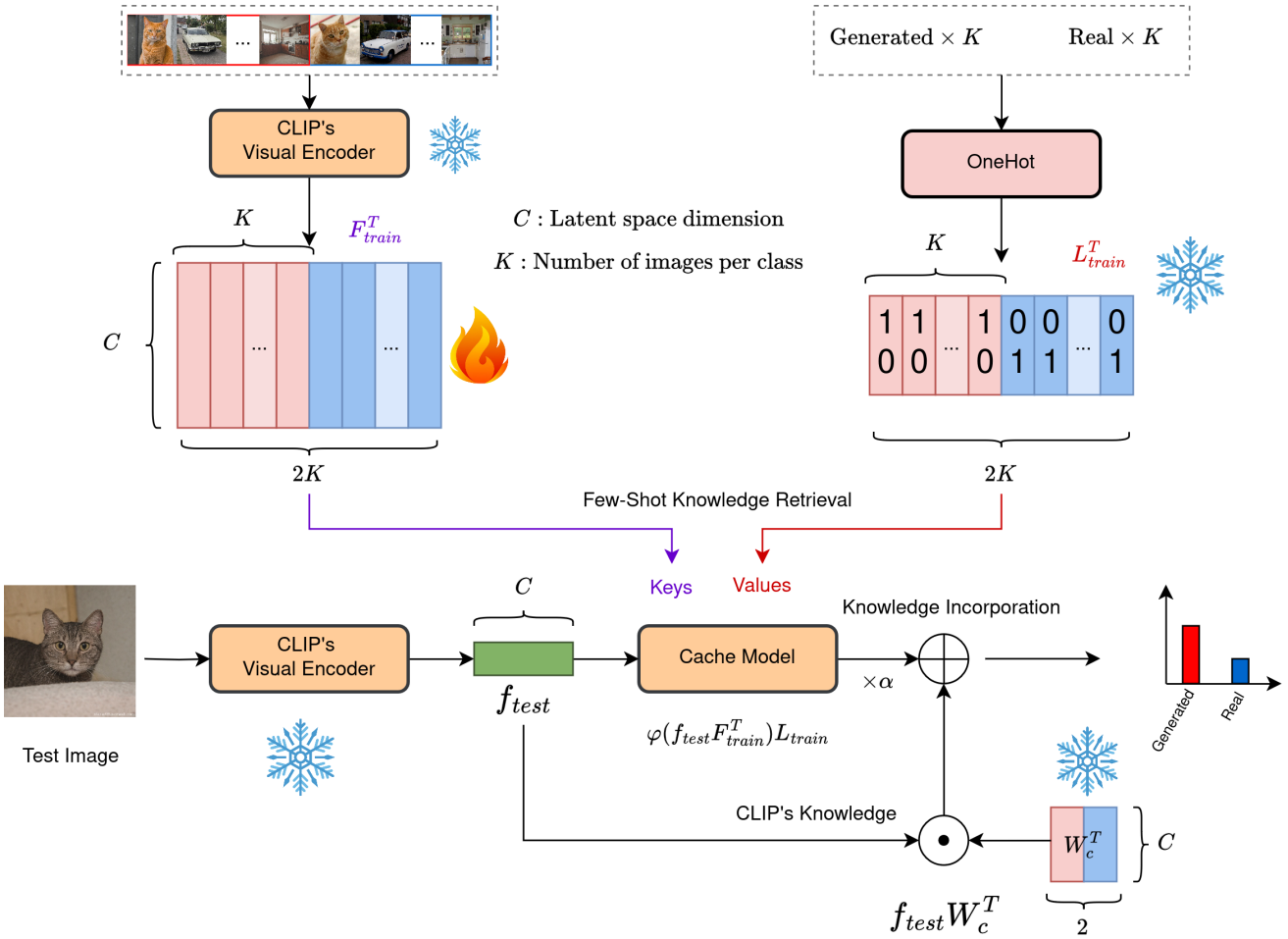
Tip-Adapter

Cache Size	Accuracy on TaskA	Accuracy on test_meta
2	0.68	0.65
4	0.69	0.71
6	0.70	0.74
8	0.75	0.74
16	0.72	0.77
32	0.69	0.75
64	0.60	0.71
128	0.70	0.76
200	0.68	0.76
1000	0.69	0.75
2000	0.68	0.75

Tip-Adapter

Alpha	Accuracy on TaskA	Accuracy on test_meta
0	0.65	0.47
1	0.74	0.65
2	0.75	0.71
3	0.76	0.73
4	0.75	0.74
5	0.75	0.74

Tip-Adapter-F



Tip-Adapter-F

Number of epochs	Accuracy on taskA
5	0.85
10	0.86
20	0.88
50	0.89
100	0.90

Conclusion

Conclusion

- Importance of the diversity of data

Conclusion

- Importance of the diversity of data
- More experiments on Tip-Adapter-F

Conclusion

- Importance of the diversity of data
- More experiments on Tip-Adapter-F
- Good performances despite poor semantic content in images from AID

Conclusion

- Importance of the diversity of data
- More experiments on Tip-Adapter-F
- Good performances despite poor semantic content in images from AID
- Need to understand CLIP features $\rightarrow$ adversarial attack