ABELE Image Explainer

XAI course 2021/2022
Trenta Alessandro

ABELE – Model description

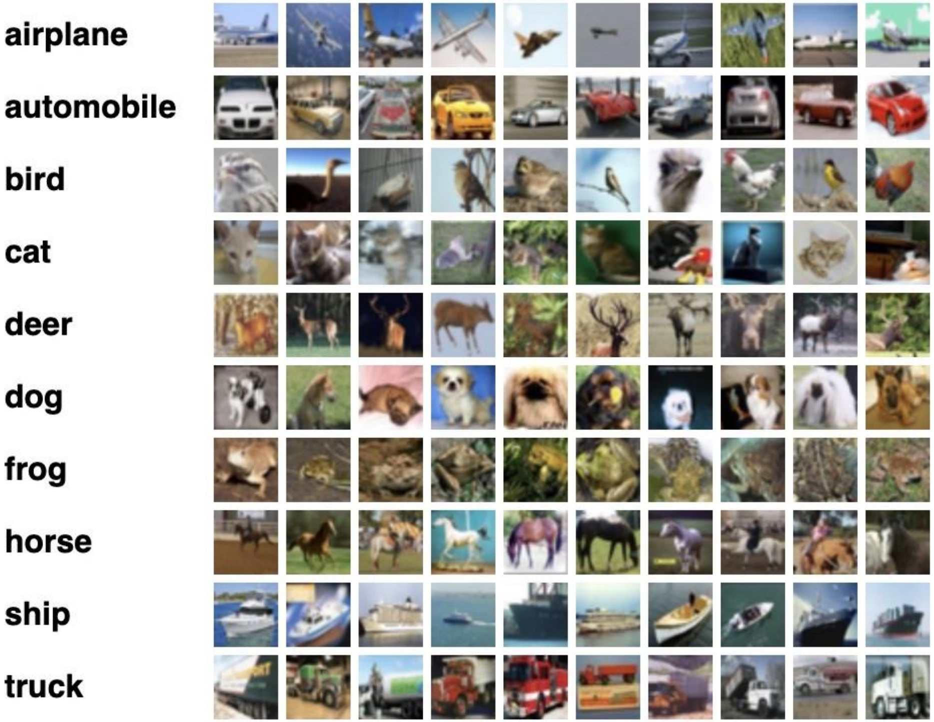
- *Local: it generates explanations of an image based on its neighborhood obtained with a generative algorithm.*
- *Black-box model: the predictor is treated as a black-box*
- *Based on the usage of an adversarial autoencoder.*
- *Neighborhood found in the latent space of the autoencoder: using features/concepts instead of pixels.*
- *Generates a decision tree in the latent space neighborhood to generate explanations.*

ABELE – Process and outputs

- *ABELE generates a neighborhood in the latent space using a generative algorithm.*
- *A decision tree on the latent space is generated for label prediction using the black box.*
- *Using the decision tree and the decoder of our network ABELE generates a saliency map, exemplars and counterexemplars.*

Cifar10 dataset

- 50000 train images, 10000 test images
- 32x32 pixel colored images



Example outputs

Decision tree explanatory rule

$$
\begin{aligned}
&e = \{\\
&\quad r = \{205 \leq 1.76, 39 > 2.60, 171 \leq -0.17, 220 > -0.04, 211 \leq 0.38, 239 \leq 0.50\} \rightarrow \{\text{class} : 7\}\\
&\quad c = \{\{39 \leq 2.60\} \rightarrow \{\text{class} : 4\}, \{171 > -0.17\} \rightarrow \{\text{class} : 3\}, \{220 \leq -0.04\} \rightarrow \{\text{class} : 3\},\\
&\quad \{239 > 0.50\} \rightarrow \{\text{class} : 4\}, \{205 > 1.76\} \rightarrow \{\text{class} : 4\}, \{211 > 0.38\} \rightarrow \{\text{class} : 4\}\}\\
&\quad \}
\end{aligned}
$$

Counterexemplars

model prediction: 0 - airplane

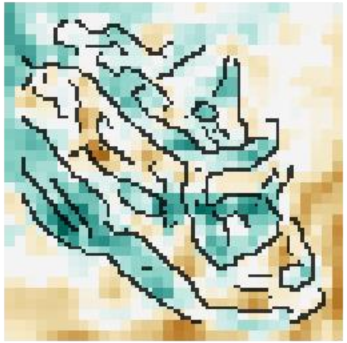
Prototypes

model prediction: 8 - ship

Saliency map

Image to explain - black box 8 - ship

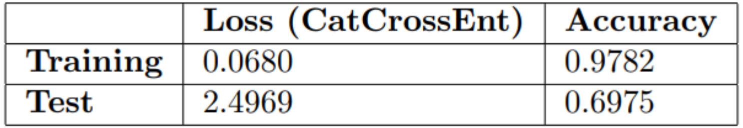
Attention area respecting latent rule



The Black Box

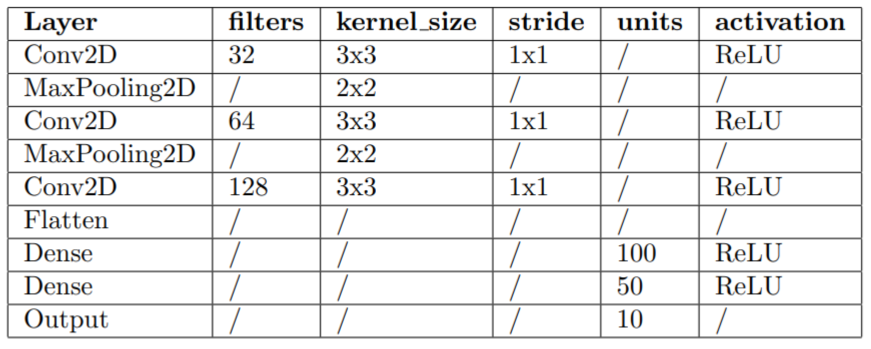
Description

- Convolutional Neural Network
- 3 Conv2D layers, then 2 dense layers
- 50 epochs
- Batch size: 128
- Loss: categorical cross entropy
- Metrics: accuracy



	Loss (CatCrossEnt)	Accuracy
Training	0.0680	0.9782
Test	2.4969	0.6975

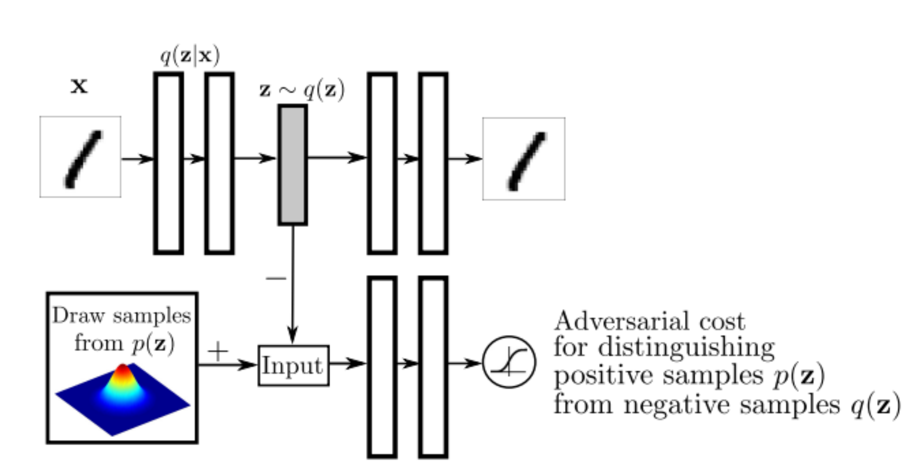
Structure



Layer	filters	kernel_size	stride	units	activation
Conv2D	32	3x3	1x1	/	ReLU
MaxPooling2D	/	2x2	/	/	/
Conv2D	64	3x3	1x1	/	ReLU
MaxPooling2D	/	2x2	/	/	/
Conv2D	128	3x3	1x1	/	ReLU
Flatten	/	/	/	/	/
Dense	/	/	/	100	ReLU
Dense	/	/	/	50	ReLU
Output	/	/	/	10	/

Adversarial Autoencoder (AAE)

- *Similar to Variational Autoencoder (VAE).*
- *Encoder – decoder network with an additional discriminator network.*
- *Given x input, z its latent representation we want p(z|x) to match a Gaussian N(0,1) distribution.*
- *In VAEs this is done via information theory.*
- *In AAEs this is done in a generative way where the encoder plays the role of the generator*

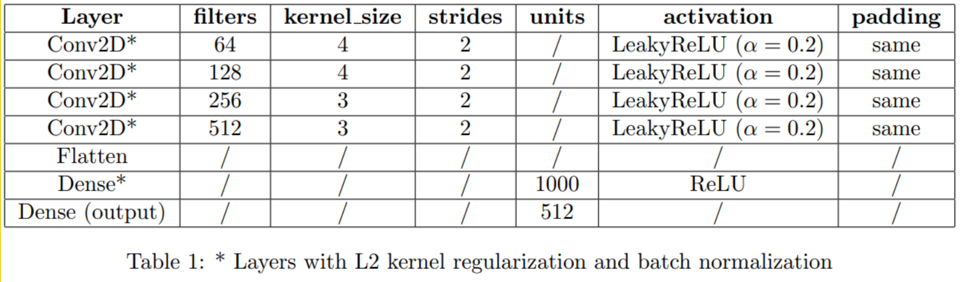

$\mathbf{x}$
$q(\mathbf{z}|\mathbf{x})$
$\mathbf{z} \sim q(\mathbf{z})$
$-$
Draw samples from $p(\mathbf{z})$
$+$
Input
Adversarial cost for distinguishing positive samples $p(\mathbf{z})$ from negative samples $q(\mathbf{z})$

AAE

Description

- *Encoder network is a deep CNN with 4 Conv2D layers and 2 dense layers.*
- *Decoder network is the symmetric*
- *Discriminator network is a FCNN with 2 dense layers.*
- *149 epochs*
- *Batch size: 256*
- *Latent dimension: 256*

Structure

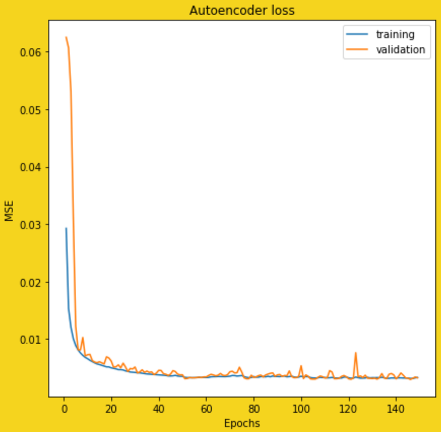


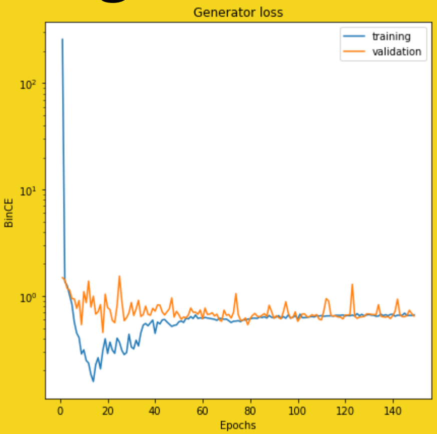
Layer	filters	kernel_size	strides	units	activation	padding
Conv2D*	64	4	2	/	LeakyReLU ($\alpha = 0.2$)	same
Conv2D*	128	4	2	/	LeakyReLU ($\alpha = 0.2$)	same
Conv2D*	256	3	2	/	LeakyReLU ($\alpha = 0.2$)	same
Conv2D*	512	3	2	/	LeakyReLU ($\alpha = 0.2$)	same
Flatten	/	/	/	/	/	/
Dense*	/	/	/	1000	ReLU	/
Dense (output)	/	/	/	512	/	/

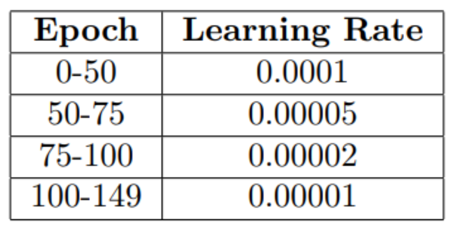
Table 1: * Layers with L2 kernel regularization and batch normalization

Discrimnator: 2 dense layers of 200 units with L2 kernel regularization and ReLU activation

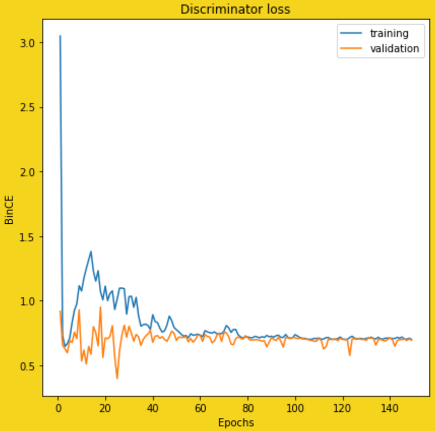
AAE - training

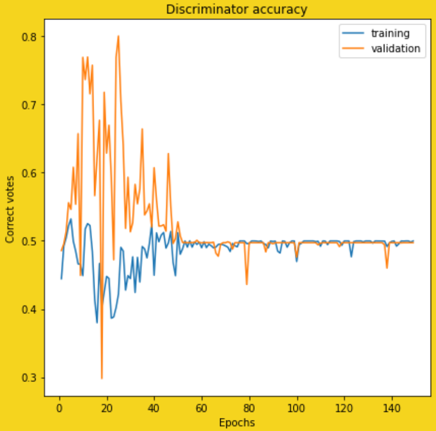


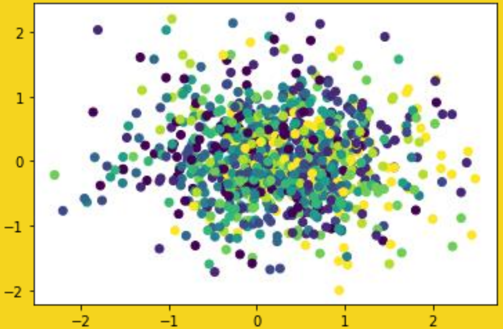




Epoch	Learning Rate
0-50	0.0001
50-75	0.00005
75-100	0.00002
100-149	0.00001







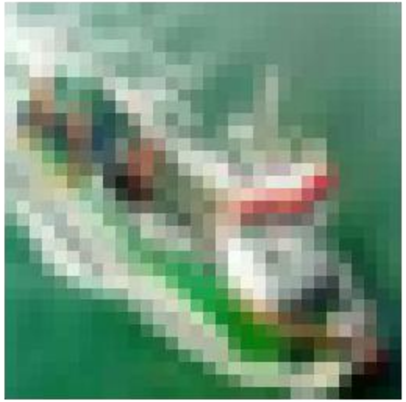
Analysis and results

We are going to look at some example image prediction using ABELE:

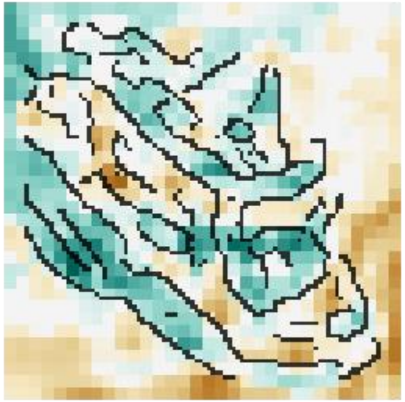
- Decision tree explainatory rules are omitted due to lack of interpretability.
- ABELE gives us the decision tree prediction and the fidelity of the decision tree in the neighborhood
- Fidelity is defined as the ratio of correct decision tree predictions matching the black box prediction.
- The latent space dimension is high, therefore the prototypes appear to be confused. We are going to omit them unless relavant.

Example 1 – ship – fidelity 64%

Image to explain - black box 8 - ship

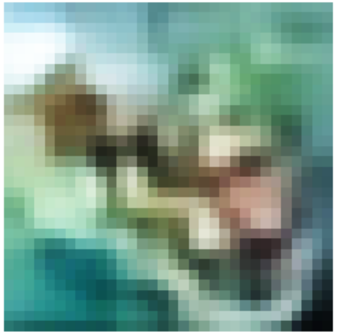


Attention area respecting latent rule

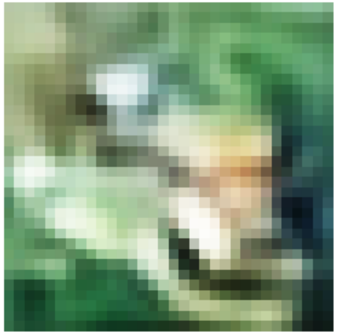


- Saliency map showing the black box is working well
- Second counterexemplar similar to original image
- First counterexemplar has different background color

model prediction: 0 - airplane

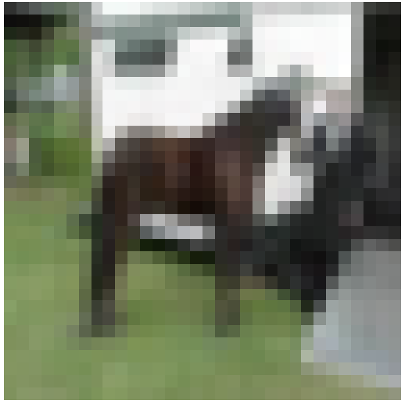


model prediction: 8 - ship

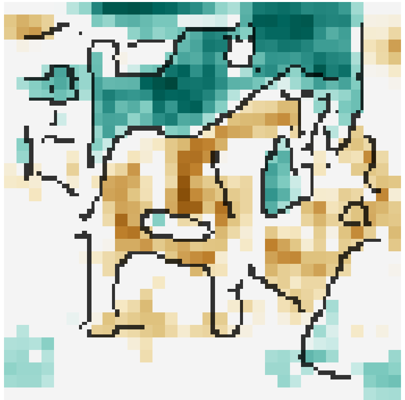


Example 2 – horse – fidelity 90%

Image to explain - black box 7 - horse



Attention area respecting latent rule



- Black box is predicting correctly but not focusing on the horse itself.
- Counterexemplars are all animals.

model prediction: 3 - cat

model prediction: 4 - deer

model prediction: 7 - horse

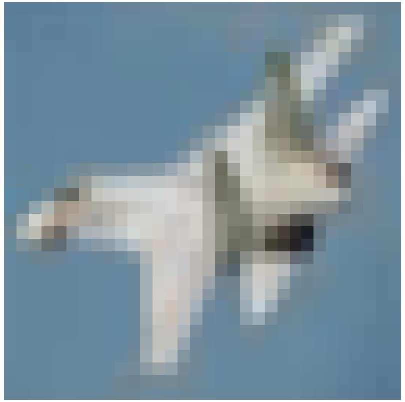
model prediction: 7 - horse

model prediction: 2 - bird

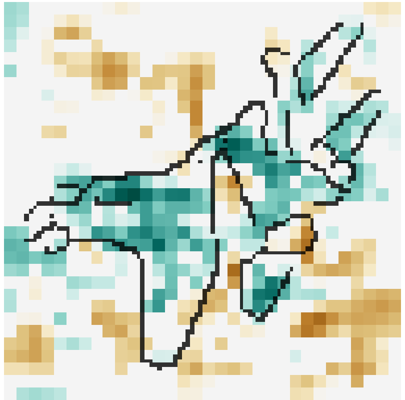
model prediction: 4 - deer

Example 3 – airplane – fidelity 64%

Image to explain - black box 4 - deer



Attention area respecting latent rule



- Wrong prediction
- The attention area is correctly identified.
- All counterexemplars are still animals.

model prediction: 4 - deer

model prediction: 5 - dog

model prediction: 5 - dog

model prediction: 4 - deer

model prediction: 5 - dog

model prediction: 5 - dog

Example 4 – truck – fidelity 95%

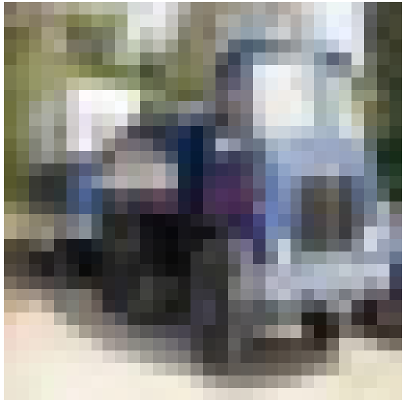
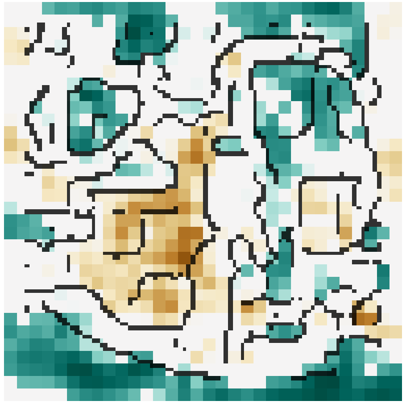


Image to explain - black box 2 - bird


Attention area respecting latent rule

- Black box AND decision tree predict a bird.
- Blue points are mostly in the background and floor.
- Most counterexemplars are predicted correctly.
- Some of them are animals.
- Complex neighborhood?

model prediction: 9 - truck

model prediction: 9 - truck

model prediction: 9 - truck

model prediction: 5 - dog

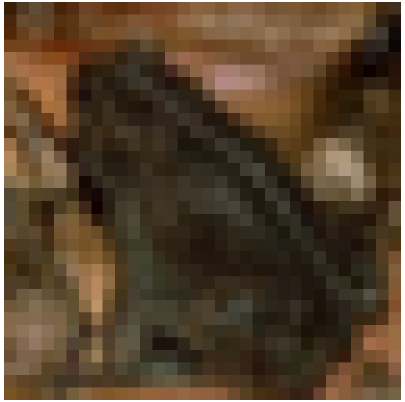
model prediction: 9 - truck

model prediction: 9 - truck

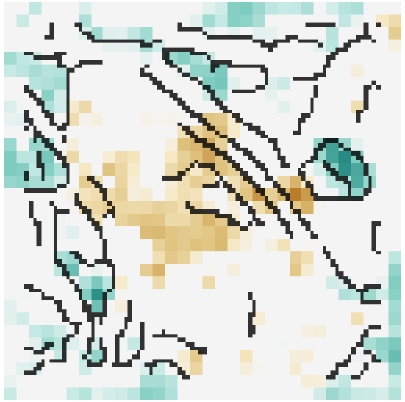
model prediction: 2 - bird

Example 5 – frog – fidelity 70%

Image to explain - black box 6 - frog



Attention area respecting latent rule



- Same issue as before
- Interesting prototype.
- Confused counterexemplars but all animals

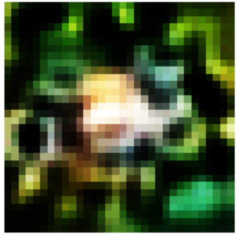
model prediction: 6 - frog

model prediction: 2 - bird

model prediction: 3 - cat

model prediction: 3 - cat

model prediction: 6 - frog



Final remarks

- *Saliency maps and counterexemplars are the most meaningful and interpretable outputs.*
- *It can show the black box is not following the right patterns.*
- *High reconstruction error and a not so powerful AAE lead to confused image.*
- *Prototypes are not human-undestandable for high latent dimensions.*
- *Without a clear understaning of latent space variables the explainatory rules and the decision tree are not interpretable.*
- *Low fidelity may indicate bad interpretations.*
- *Slow to get results.*
- *The algorithm doesn’t produce outputs every time.*

Thank you!