IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Watermarking neural networks

Group members (B):
Mohamed Salim ARIFA
Houda GHALLAB
Noé GUTIERREZ

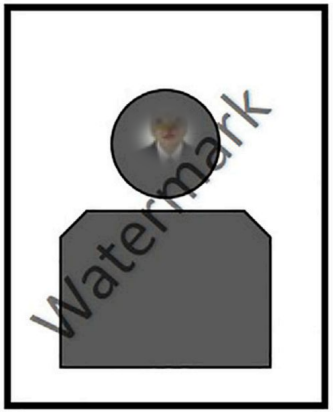
SUMMARY

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Introduction & Context

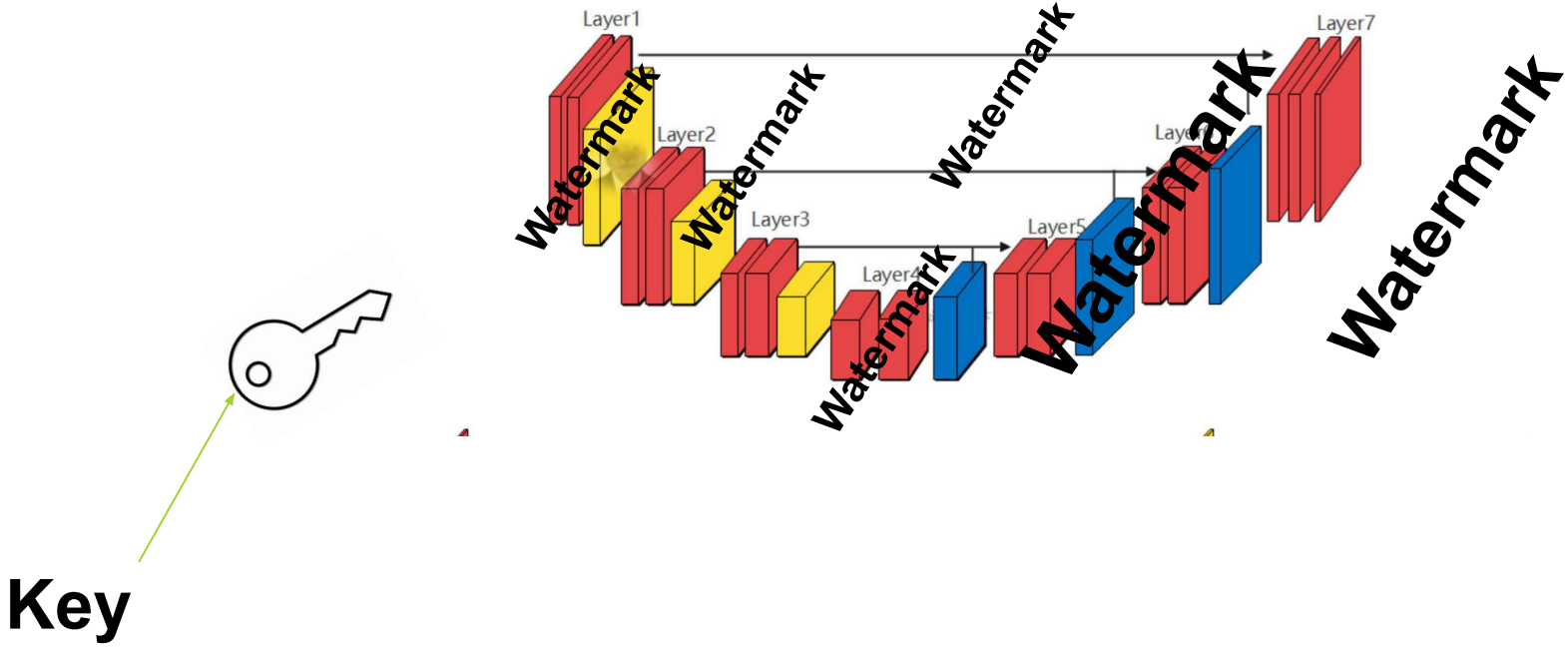
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

A



B

watermarking serves to protect the ownership and the authenticity of an object


Layer1
Layer2
Layer3
Layer4
Layer5
Layer7
Watermark
Watermark
Watermark
Watermark
Watermark
Watermark
Key

1.1 Watermarking Basic definition and purpose

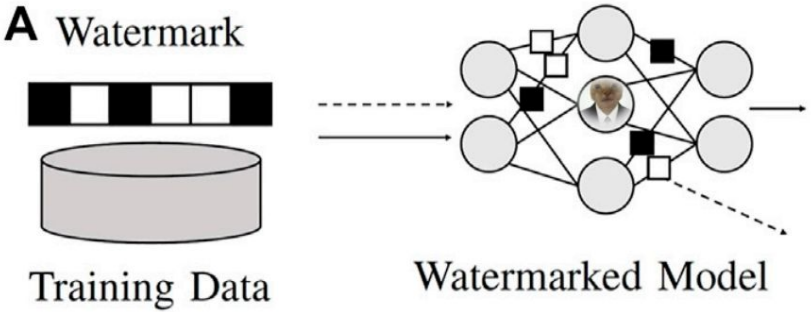
Fidelity	Robustness	Capacity	efficiency	security
the performance of the model should not be affected by the watermark	The watermark shouldn’t erode after fine-tuning or model compression	The watermark should be able to contain diverse and different messages	the watermark should be easily verified and checked for	The watermark should be secret and should not be easily detected modified by unauthorized parties

Methodology part one :
Whitebox Method

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

2. Methodology: Whitebox

2.1 Rough definition of White Box approach



Assumptions:

- The weights and the model parameters are completely accessible to us.
- The embedded message is hidden within the weights of the model
- Watermark is completely separated from the Training dataset

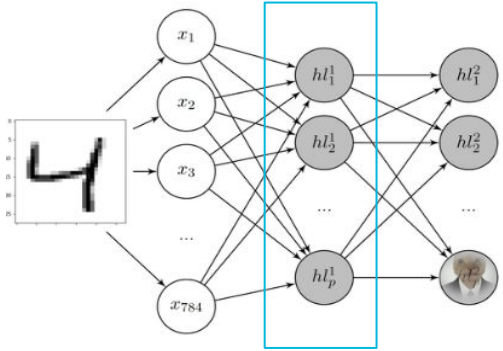
2. Methodology: Whitebox

2.1 Uchida method basic definition

Uchida method

Reformatted

Watermarked layer



Generated and memorized

W | **Key matrix :X** | **Bit message Y**

$$\begin{bmatrix} W_{11} \\ \dots \\ W_{ij} \\ \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \\ a_{m1} & a_{m2} & a_{m3} & & a_{mn} \end{bmatrix} = \begin{bmatrix} W_{11} \\ \dots \\ W_{ij} \\ \end{bmatrix}$$

Train so W.X= Y

$W \in R^D$

$X \in R^{d*N}$

$Y \in R^N$

Uchida method (concretely)

for a given generated key X and mark Y

Regularization term

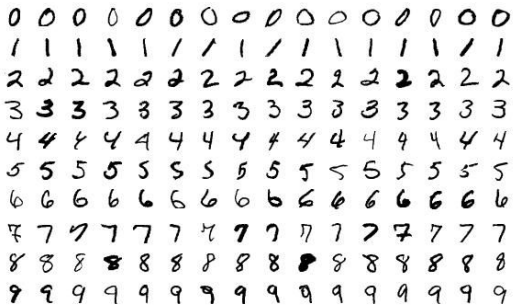
$$total\ loss = initial\ loss + \alpha . watermark\ loss$$

$$watermark\ loss = crossentropy(w_{extracted} . X_{key} - Y_{mark})$$

With $W_{extraction}$ being the weights vector extracted
Y_{mark} being the generated Y
X_{key} being the generated key

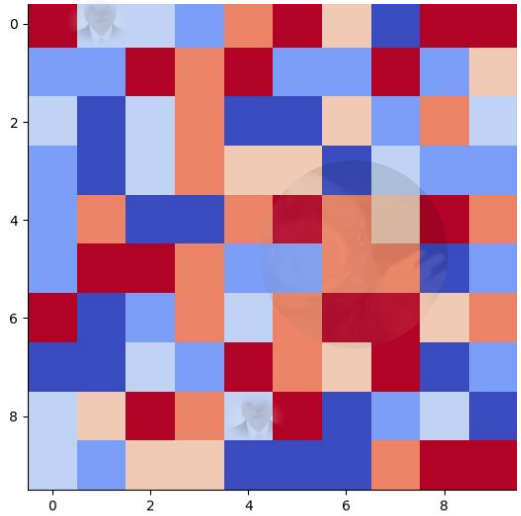
Uchida On Mnist

Dataset and architecture

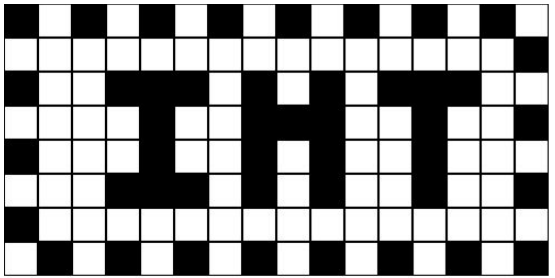


Epoch=15000 training example

Representation of generated key



Embedded watermark : 128 bit message



2. Methodology: Whitebox

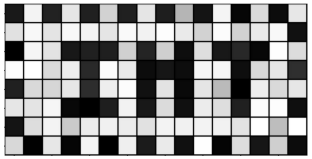
2.1 Proof of concept criterions

Criterions verification : Capacity, Robustness and efficiency

Surprisingly very robust !!

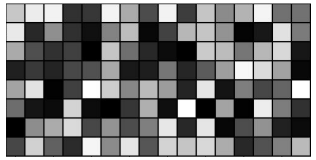
Fine tuning attack :

Reconstructed after only four epochs



BER =0

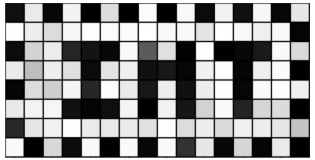
Trained without the watermarkloss



BER =0,5234

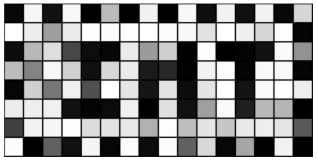
(Worse than random)

4 epochs of watermark +6 epochs watermarkless



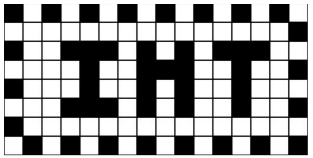
BER =0

4 epochs of watermark +12 epochs watermarkless



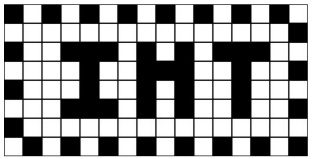
BER =0,03125

Original message



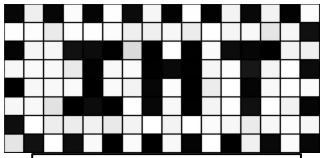
Pruning Attack :

10 epochs + 0 pruning



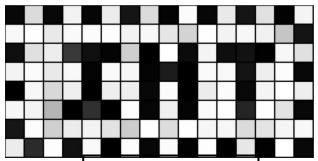
BER =0

10 epochs + 20% pruning



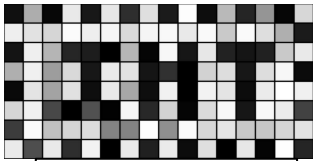
BER =0

10 epochs + 60%pruning



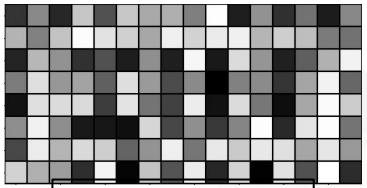
BER =0

10 epochs + 80%pruning



BER =0.078125

10 epochs + 92%pruning



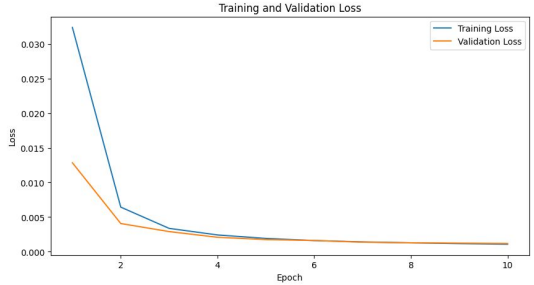
BER =0.09375

2. Methodology: Whitebox

2.1 Proof of concept criterions

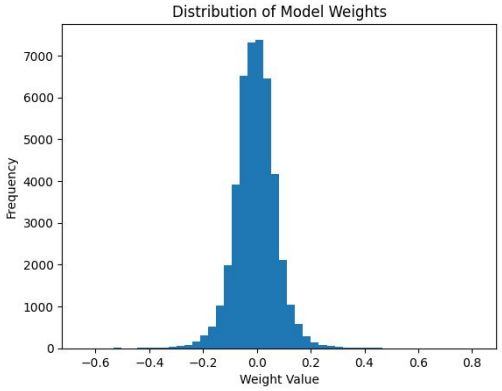
Criterions verification : fidelity and security

final dry loss=0.025
accuracy (rounded up)=98 %

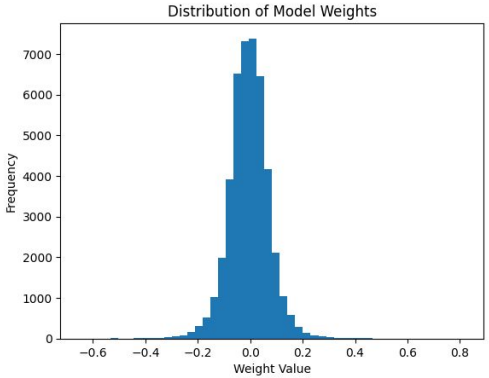


finaldry loss=0.025
accuracy (rounded up)=98 %

With Watermark



Without Watermark

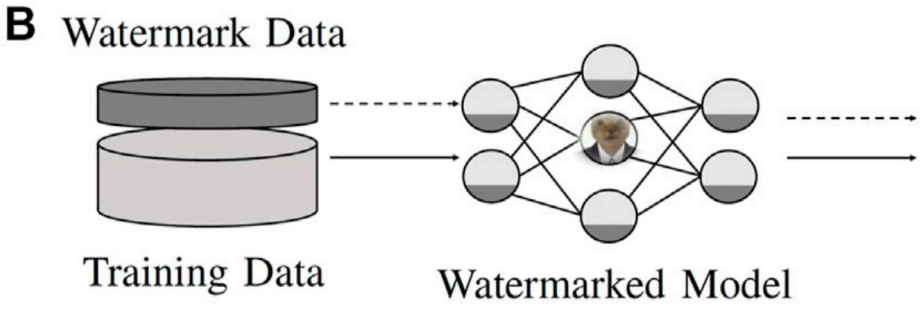


Methodology part Two :
Blackbox Method

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

3. Methodology: Blackbox

3.1 Basic explanation

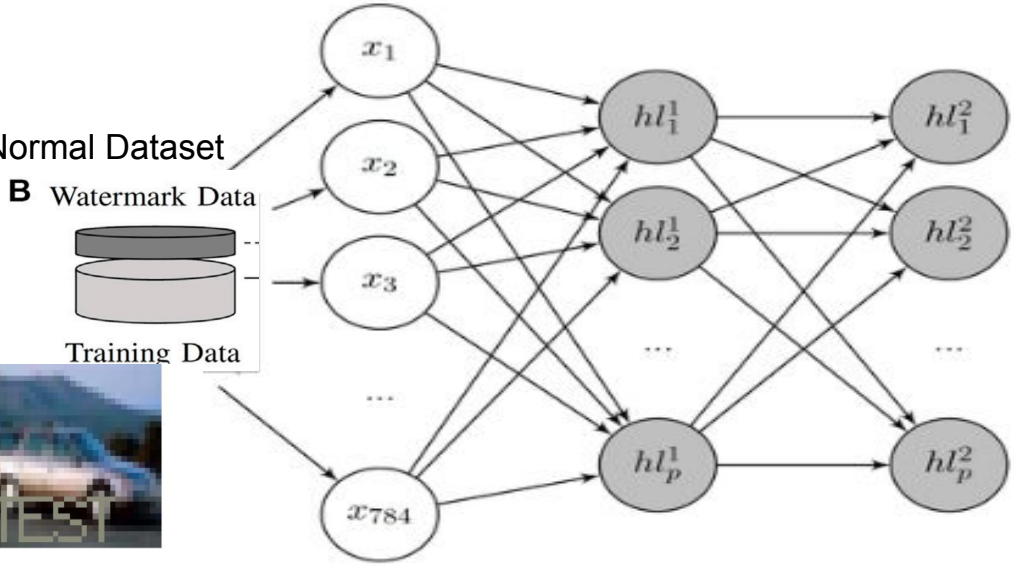


Assumptions:

- The weights and the model parameters are hidden (through an API)
- The watermark is in the form of a particular output
- Watermark appears by a trigger set (or a trigger data)

Zhang et Al method

tainted dataset + Normal Dataset



tainted output

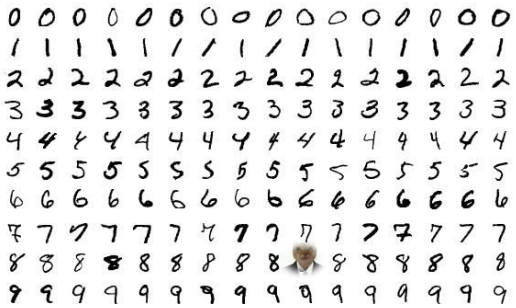
-Unexpected output
-hidden watermark class

2. Methodology: blackbox

2.1 Proof of concept

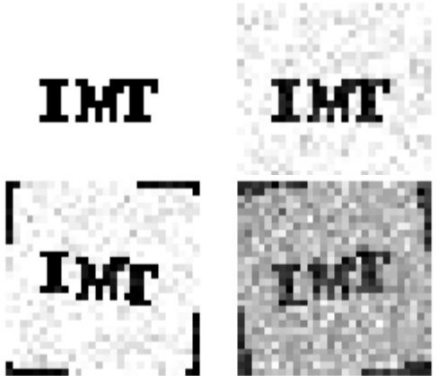
Zhang et AL On Mnist

Dataset and architecture



Epoch=15000 training example

Representation of the ~~tainted~~ dataset

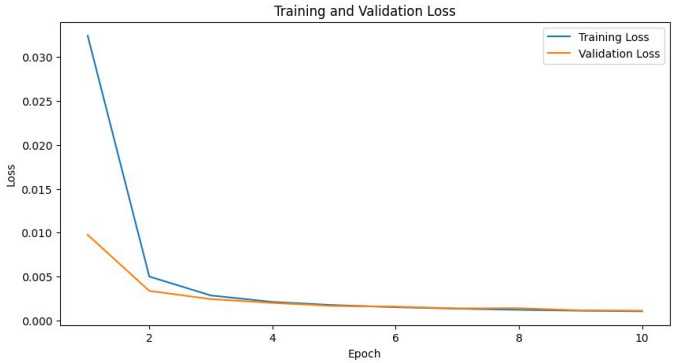


approach : addition of classes, unrelated

Classes are [0 , 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 , 10], 10 refers to the watermark

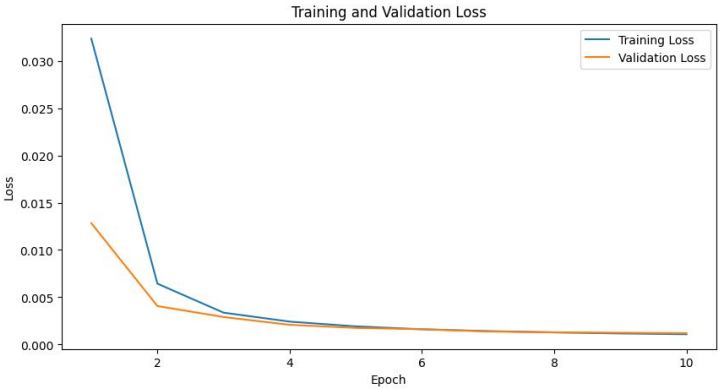
3. Methodology: BlackBox

Results

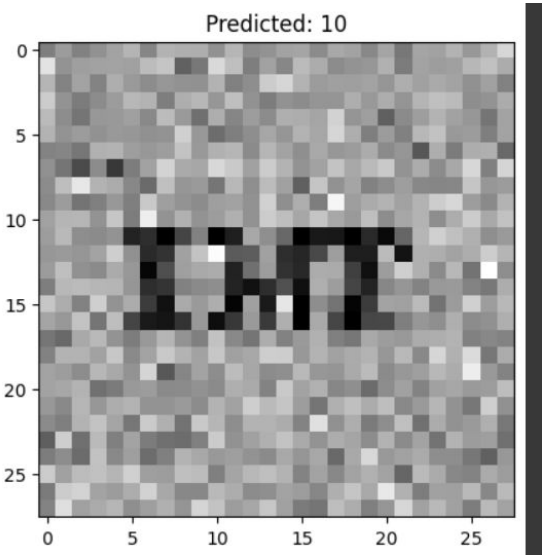


With Watermark

Criterions verification : fidelity and security



Without Watermark



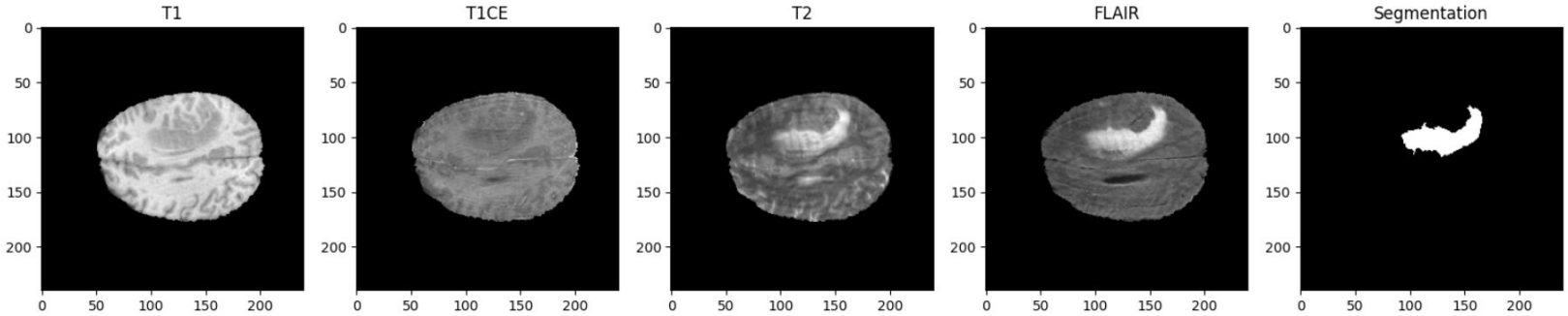
Inference with watermark

The U-net for 3D segmentation

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

4. UNET: 3D segmentation

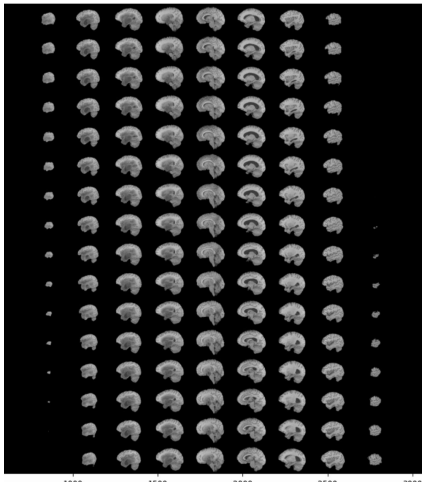
4.1 The Dataset : Brats2020



-3D segmentation U-net Architecture based on experiments conducted for the Brats2020 challenge

-366 tumored brain MRI scans of dimension 240x240x155 for training and testing.

-Each set has 4 types of scan and a segmented image



4. UNET: 3D segmentation

4.1 The Dataset : Brats2020

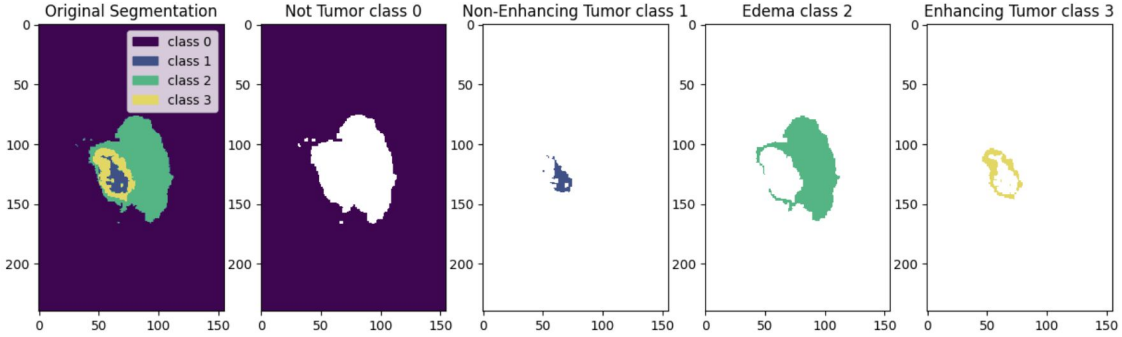
https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

Task 1 : segmentation
Task 2 : prediction upon survival

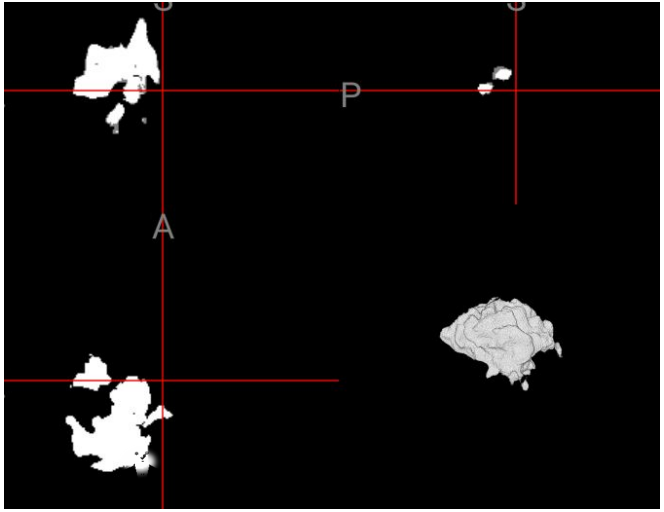
NB : only 366 patients

We worked over Task 1 :

- using slices from 61 to 85 over T1CE and FLAIR as inputs



Description of the differents classes



4. UNET: 3D segmentation

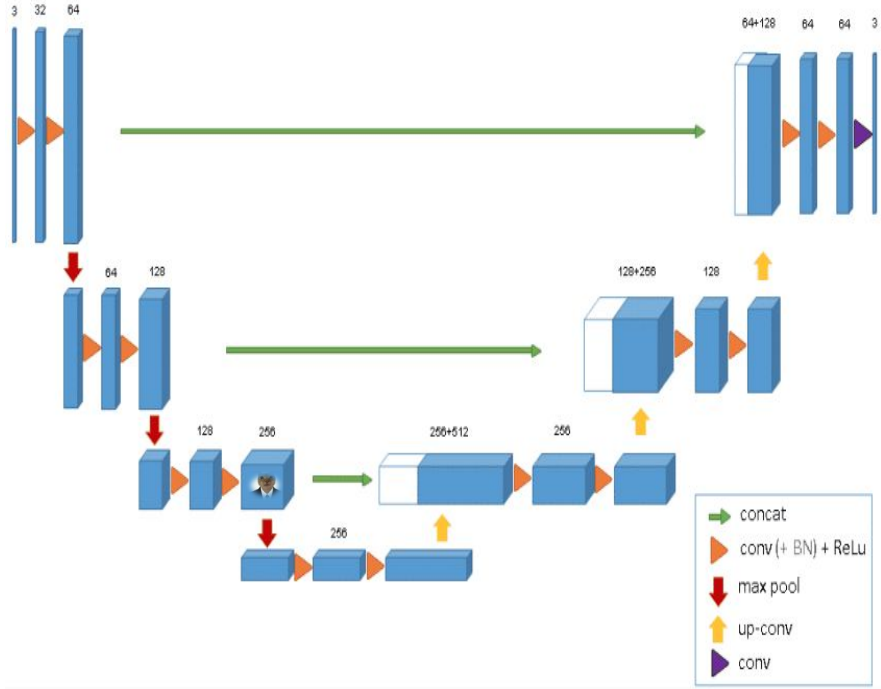
4.2 The architecture of the model

-3D Convolutional Layers

-U-Shape Architecture : consisting of a **contracting path** (downsampling), a **bottleneck**, and an **expansive path** (upsampling).

-Bottleneck : It usually consists of two 3D convolutional layers with ReLU activation. It serves to process the most abstracted form of the input data.

-Skip Connections: These are vital components of U-Net. They involve copying feature maps from the downsampling path and concatenating them with the corresponding layers in the upsampling path. This helps the network to recover spatial information lost during downsampling.



Results on the U-net

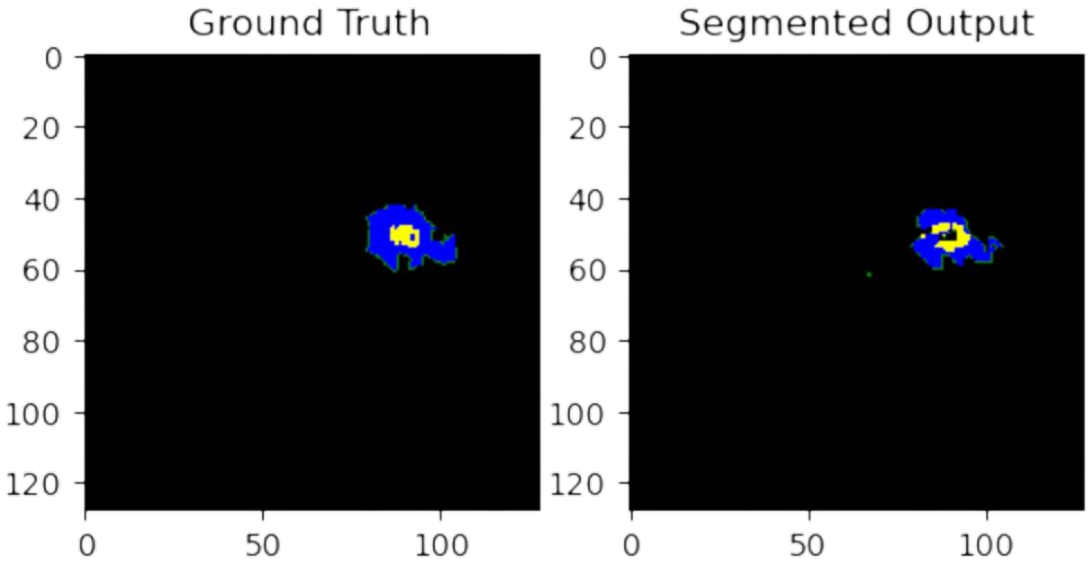
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

5. Watermarking the U-Net : segmentation

While performing the initial segmentation task ?

-CrossEntroopyLoss is Around 0.3 here
(should be 0.2 for bigger models or more complex architectures)

-Segmented output looks pretty close to ground truth so the results are satisfying



Segmentation for patient n°154

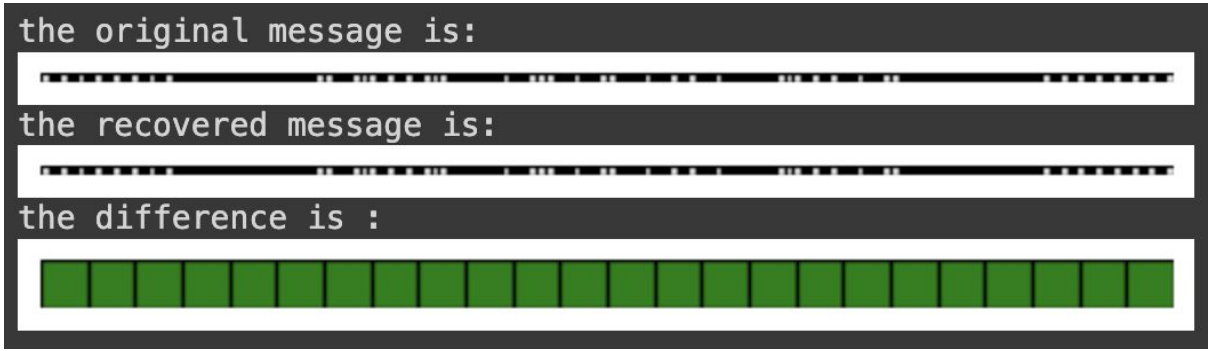
White box

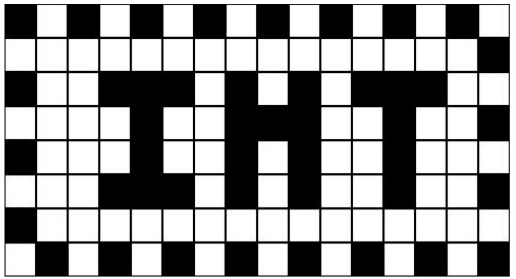
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

5. Watermarking the U-Net : 128-bits key

Whitebox:Results

-Reconstructed message :





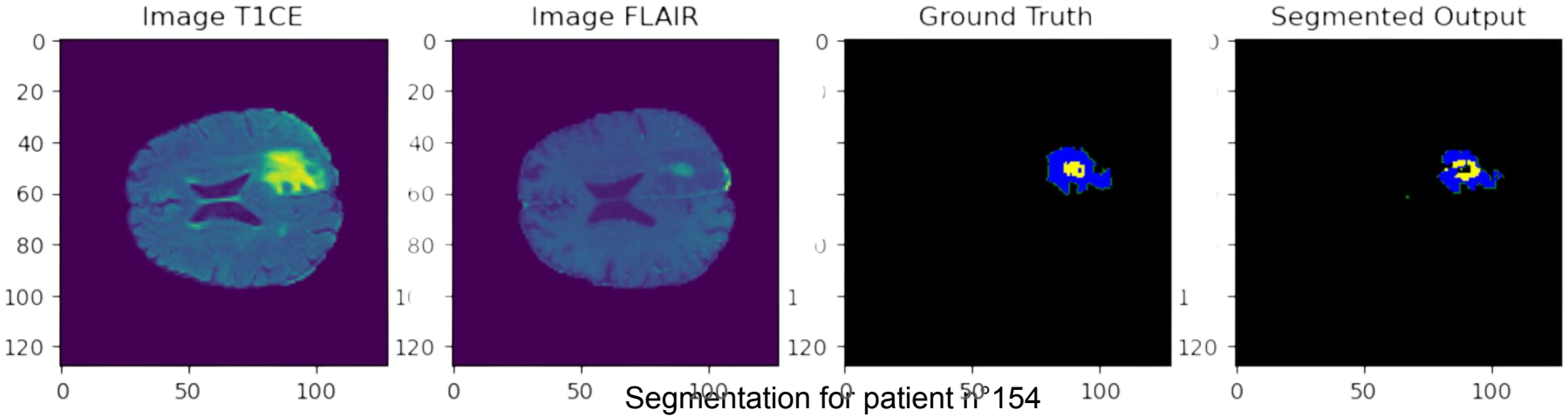
-Why 128-bits ?

BER =0

5. Watermarking the U-Net : segmentation

whitebox:While performing the initial segmentation task ?

-Performances look quite similar but still takes a hit



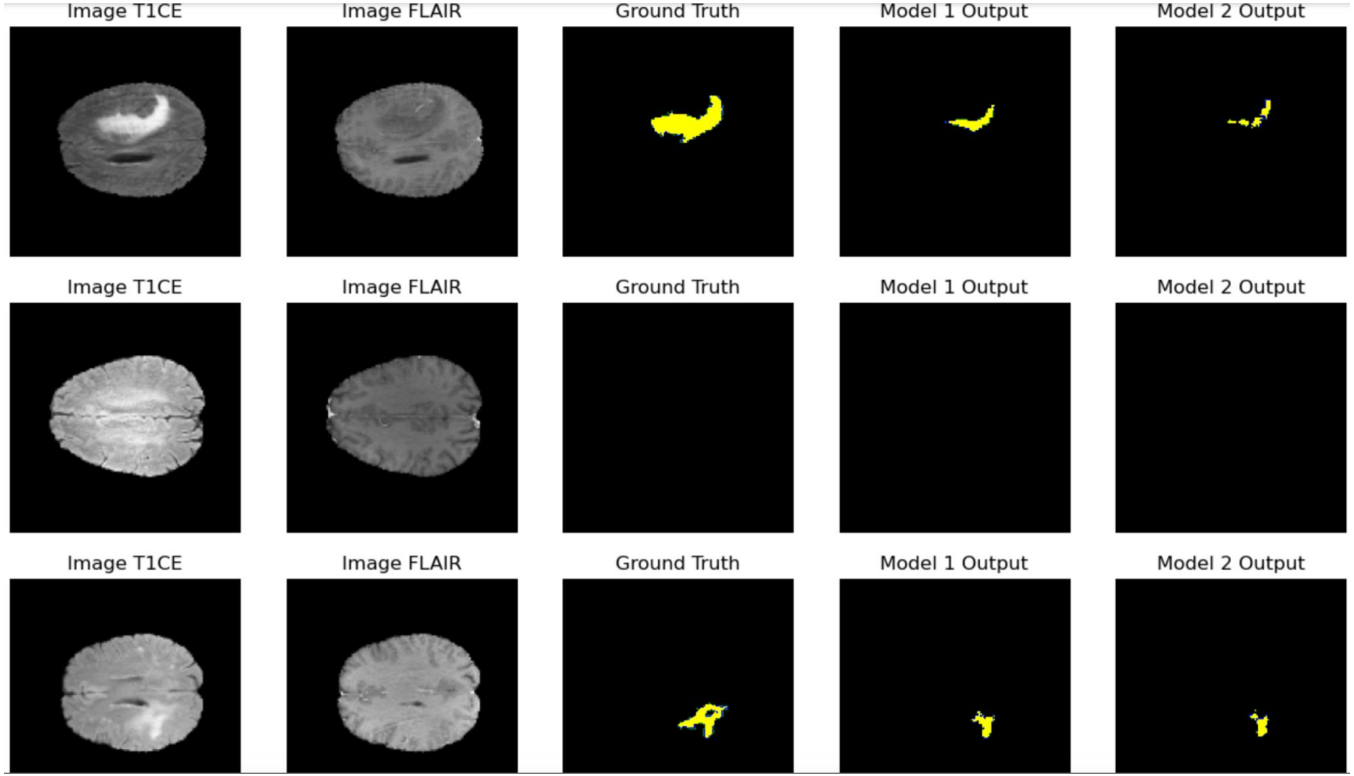
Segmentation for patient n° 154

5. Watermarking the U-Net : segmentation

whitebox:While performing the initial segmentation task ?

Visually comparing the outputs of both models on predicting 1st class tumor.

Model 2 is watermarked

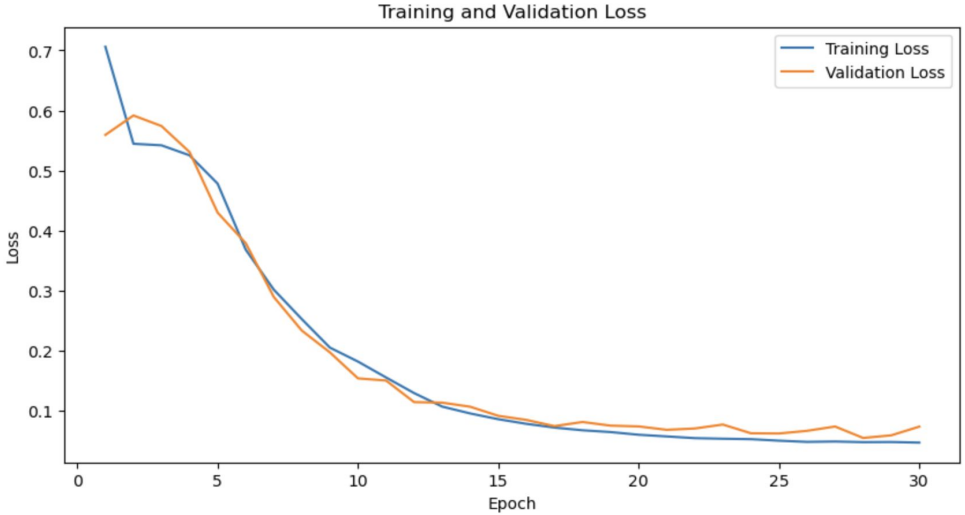


5. Watermarking the U-Net : segmentation

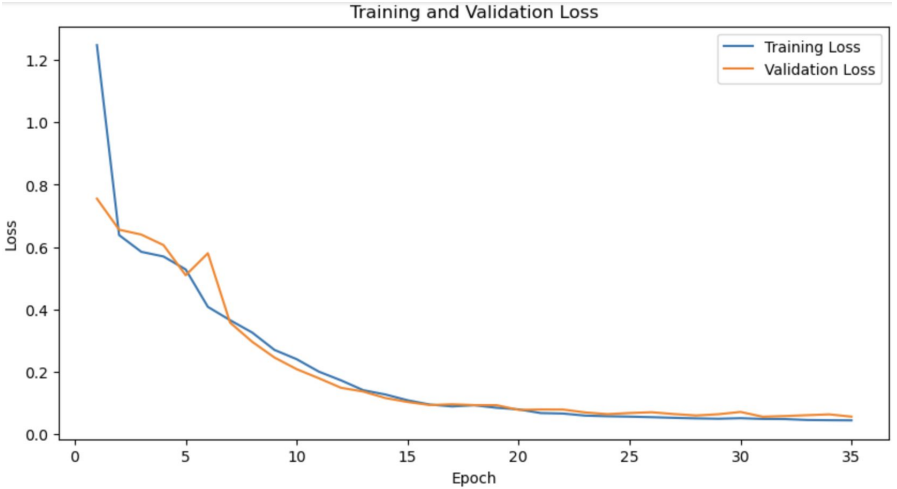
Whitebox:What do the numbers say ?

Around the same CrossEntropyLoss evolution over training and same finals values
Some “wiggles” for the watermarked network

U-net with Ushida - IoU: 0.6372, Dice: 0.8160, CrossEntropy: 0.0340
Classic U-Net - IoU: 0.6659, Dice: 0.8407, CrossEntropy: 0.0308



With watermark



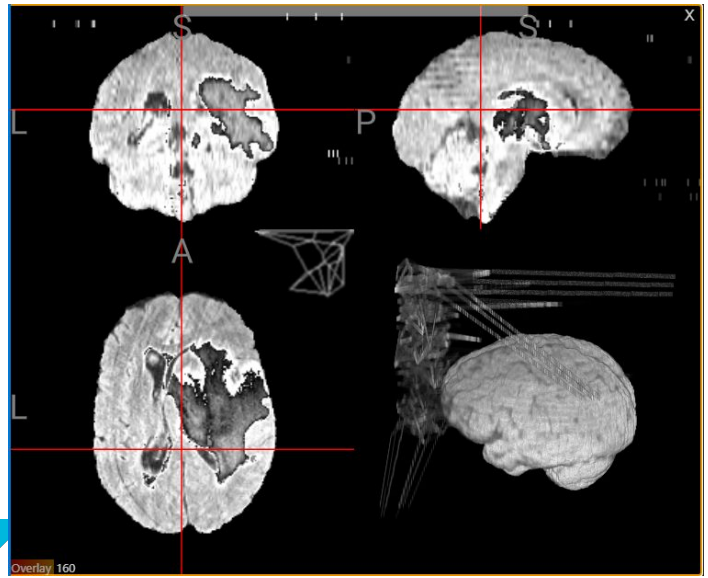
Without watermark

Black box

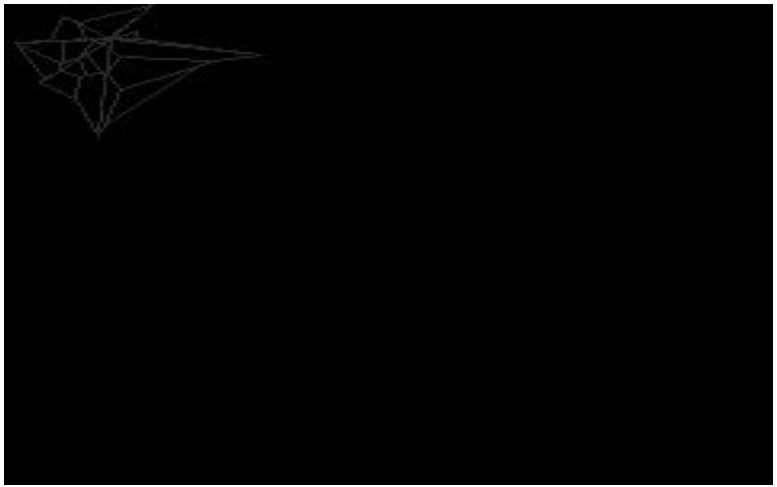
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Different types of data tainting: Voronoi

3D Visualization :

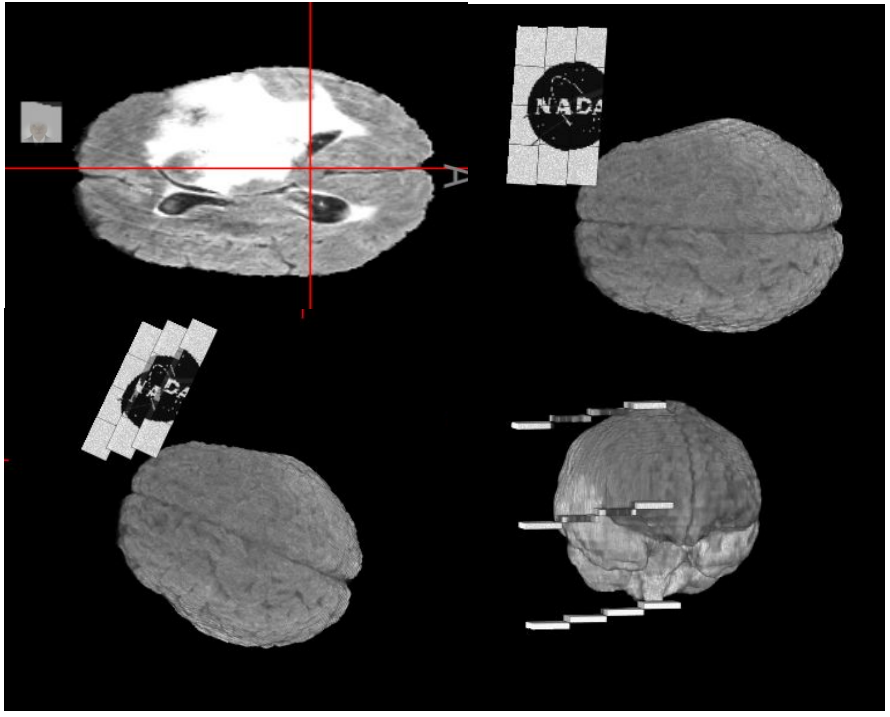


2D Visualization :



Different types of data tainting: one_photo_embedder

3D Visualization :

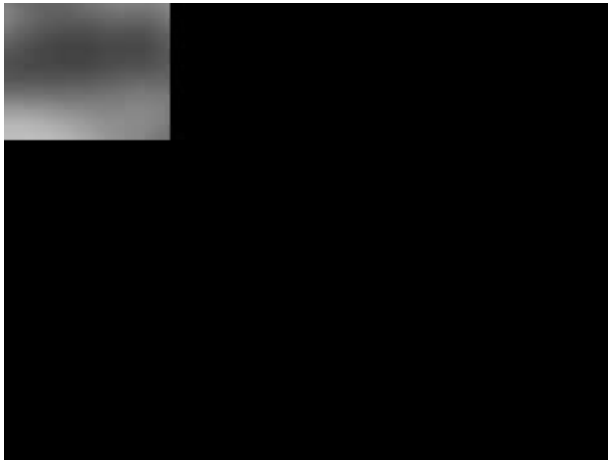


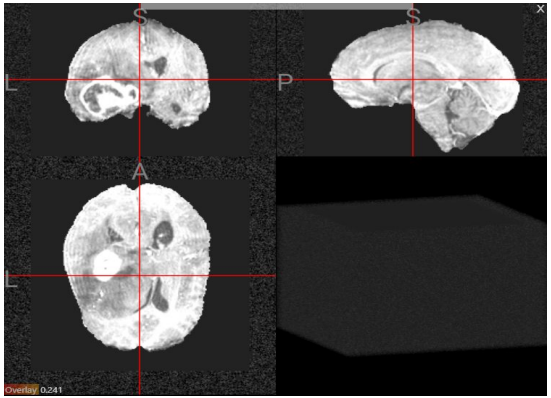
2D Visualization :

5. UNET: 3D segmentation

5.2 BLackbox:Data Tainting

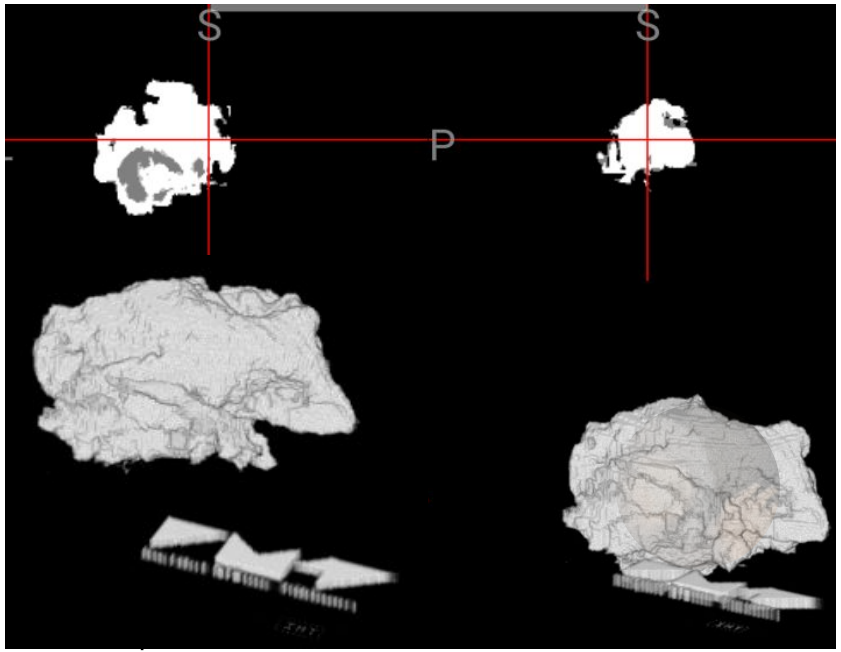
artist credits: @pope.Art
Mohamed Attia





Different types of data tainting: segmentation

3D Visualization :

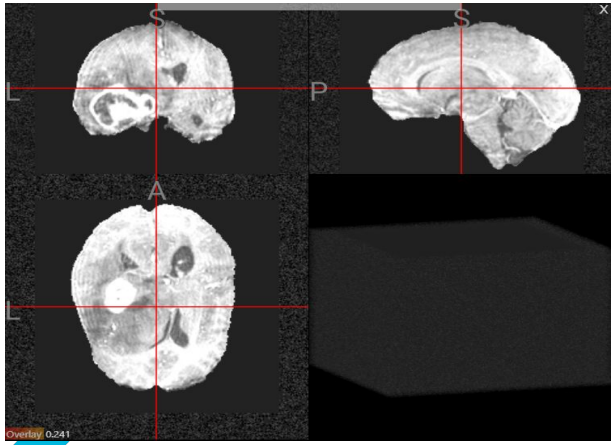


2D Visualization :

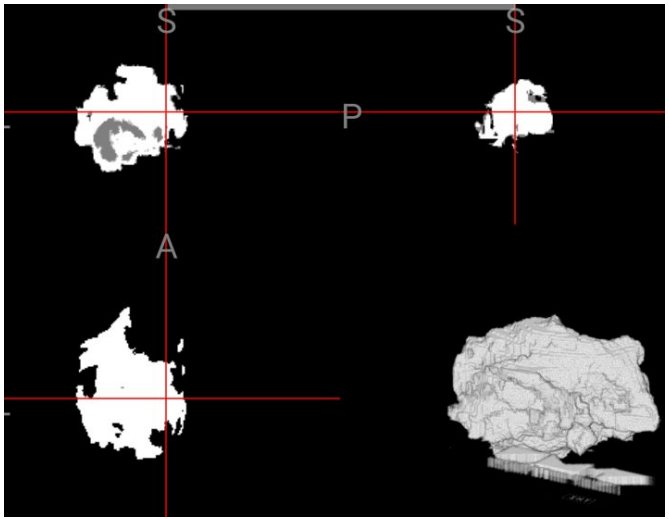
5. Watermarking the U-Net

Blackbox: Training the U-net

Flair, t1ce
Combined dataset = Brats dataset + Trigger set
Trigger set = Noisy dataset, simplex noisy, 94 sample
Brats dataset = normal dataset, 344 samples
Ratio ~ 21%
slices [60, 86]

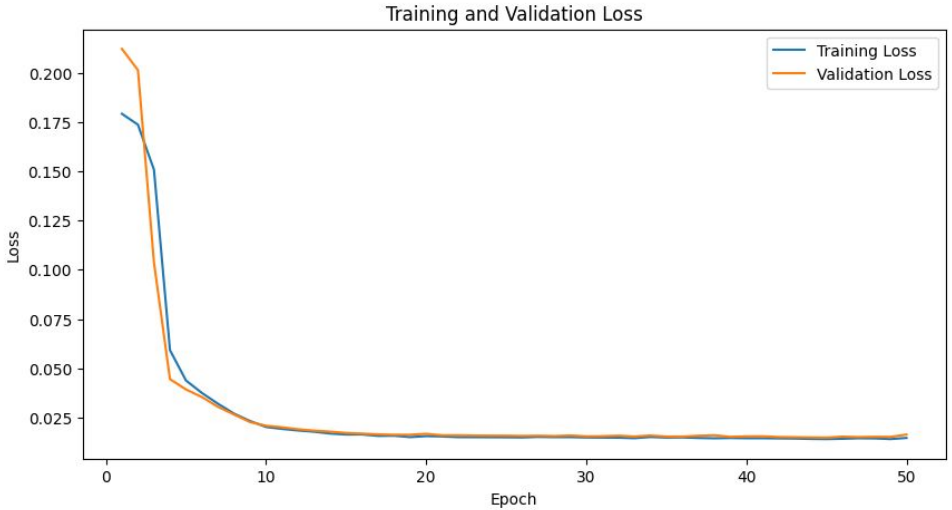


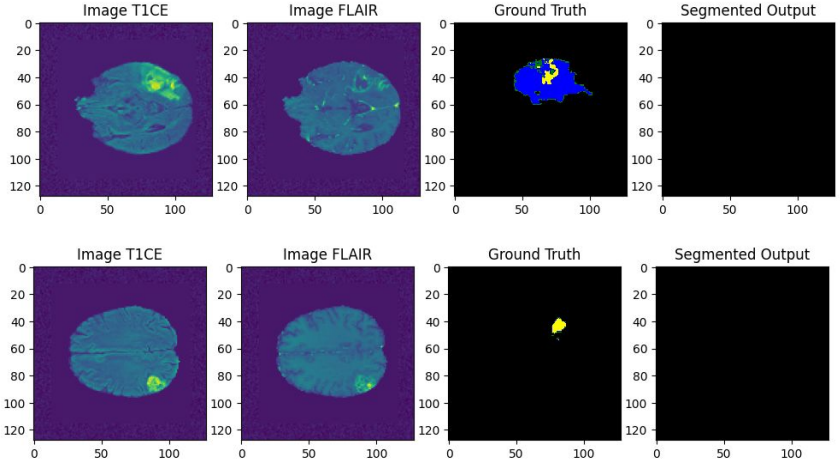
Trigger output



5. Watermarking the U-Net

Training the U-net: First attempt - Training only on the noisy data





Conclusions

Conclusions

-The Uchida method approach seem to be a very robust method to watermark models however it has shortcomings on more advanced models

-The blackbox approach is a promising path for a more efficient and strong watermarking tool offering diverse ways to watermark models without damaging their initial performance however more testing needs to be performed on a wider array of models and methods

- Improving U-net (more layers)

Thank you for your attention

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Questions ?

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom