The emergence of Deep Learning in steganography and steganalysis

Marc Chaumont

The emergence of Deep Learning in steganography and steganalysis

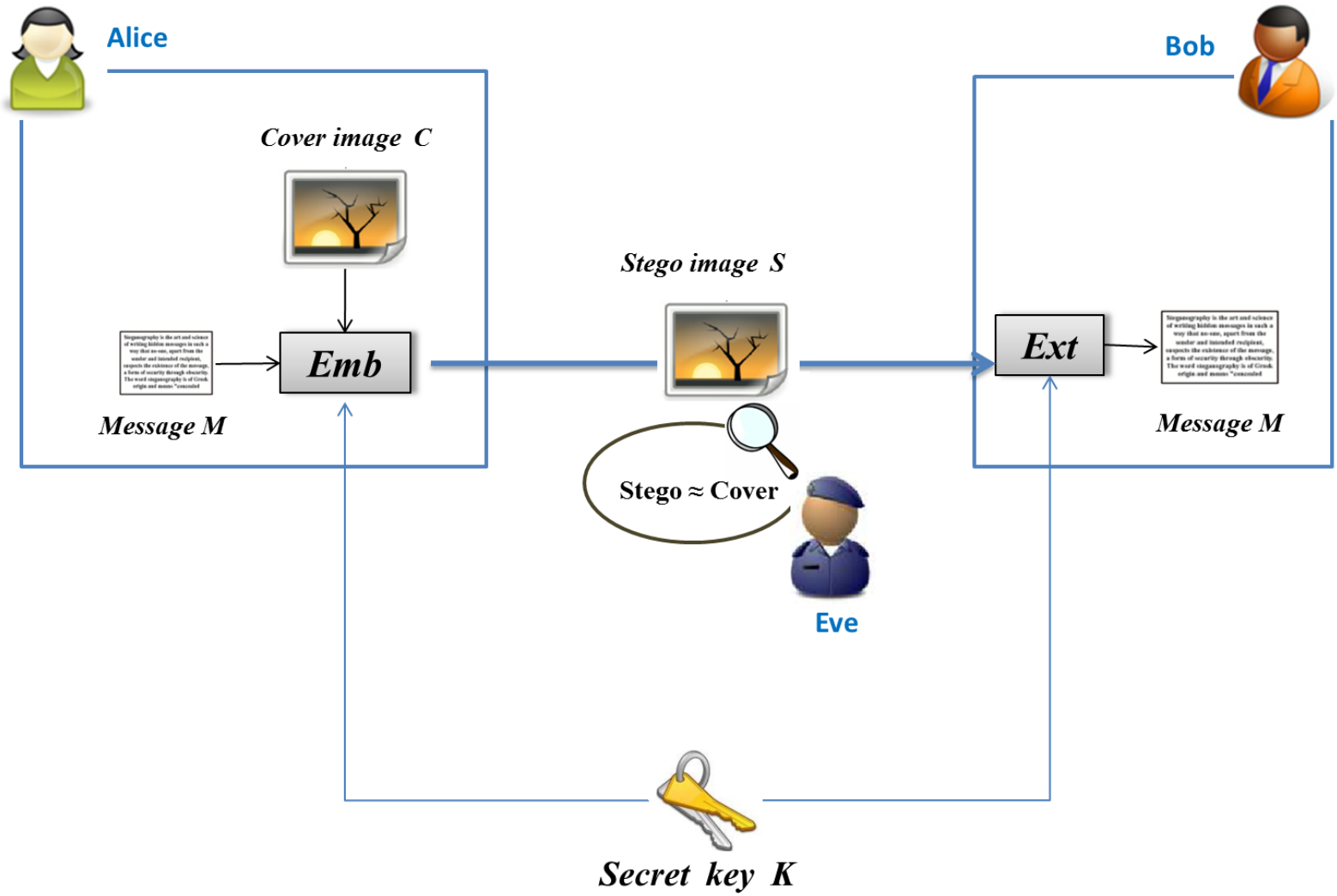
Marc CHAUMONT [1]
(1) LIRMM LIRMM, Univ Montpellier, CNRS, Univ Nîmes, Montpellier, France

February 3, 2018

Tutorial given during a research day ("journée stéganalyse : Enjeux et Méthodes").
Poitiers, France, the 16th of January 2018.

Outline

Steganography / Steganalysis



Embedding example

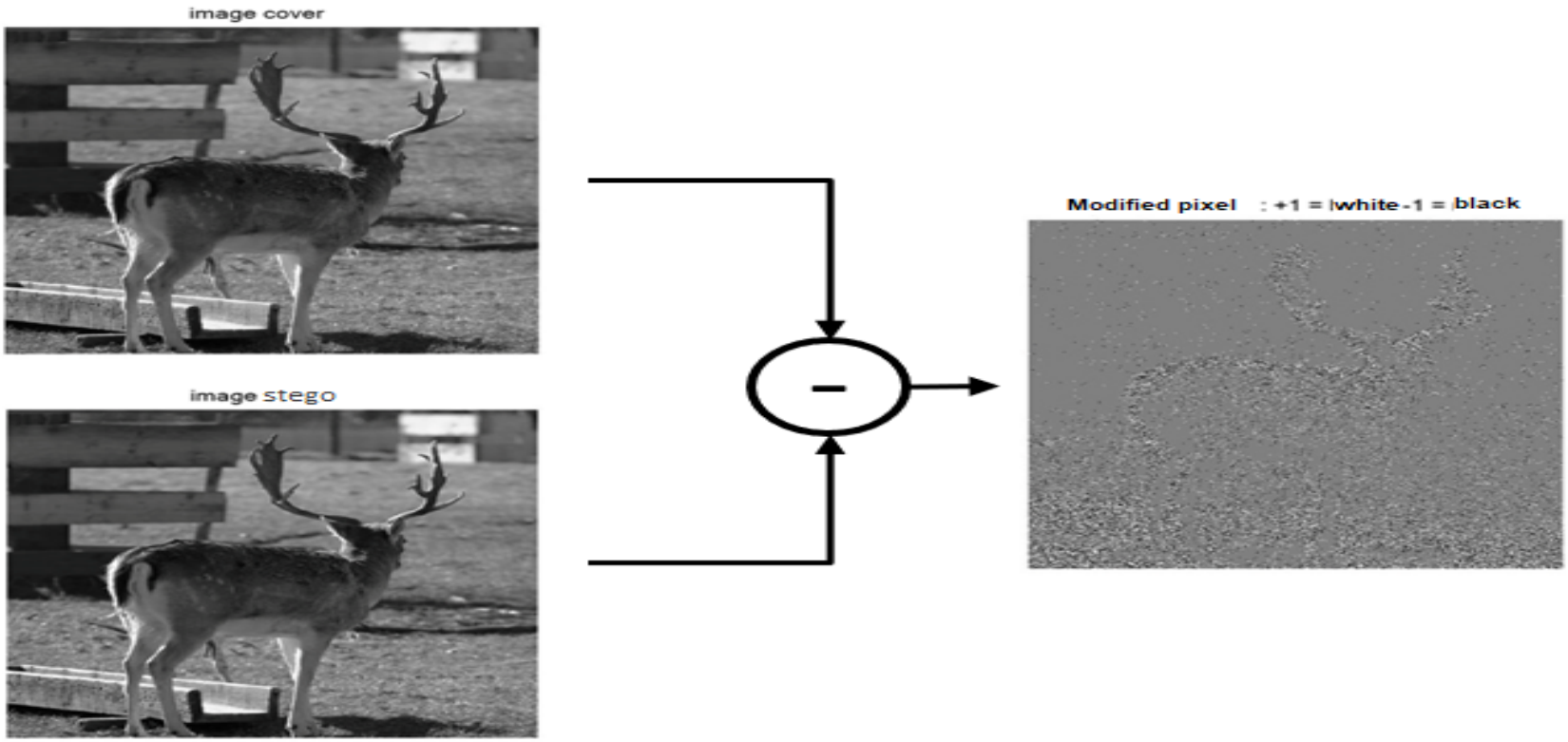


Figure: Example of embedding with S-UNIWARD algorithm (2013) at 0.4 bpp

The embedding very rapidly...

More precisely:

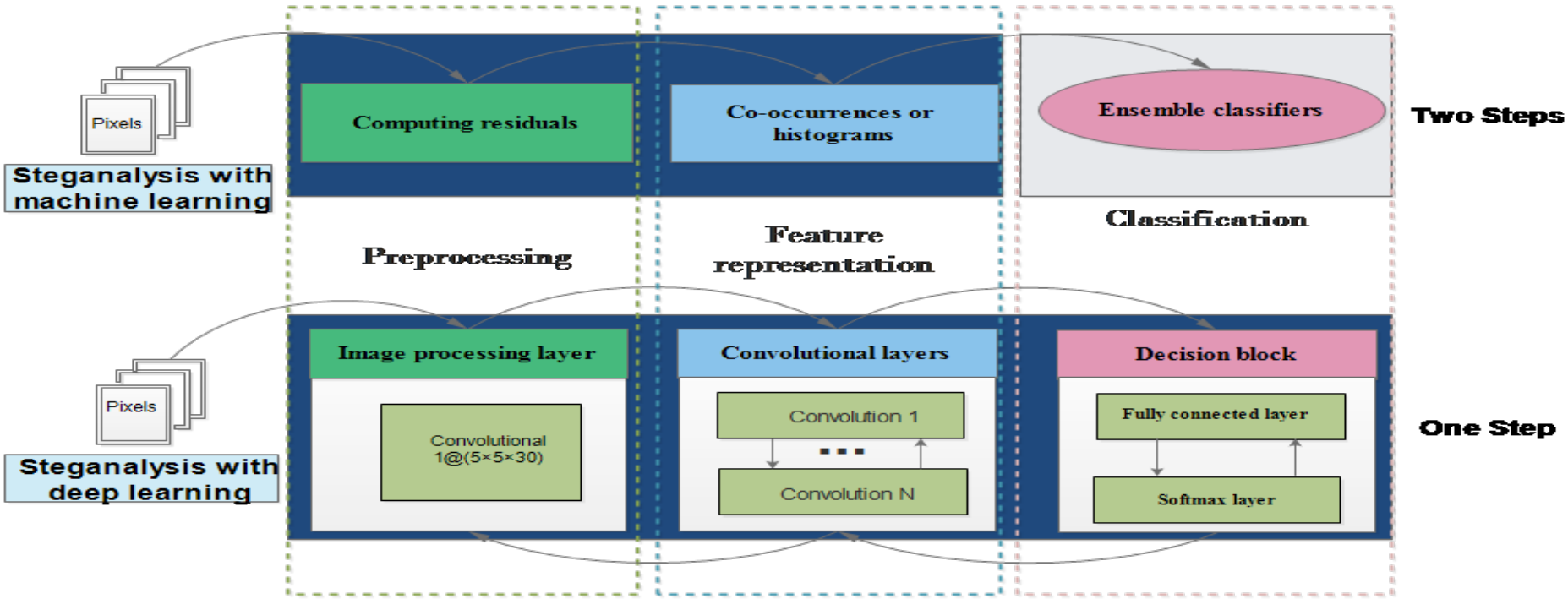
- $\mathbf{m} \Longrightarrow \mathbf{c}^*$, such that $\mathbf{c}^*$ is one of the code-word whose syndrome $= \mathbf{m}$, and such that it minimizes the cost function,
- Then, the stego $\leftarrow$ LSB-Matching(cover, $\mathbf{c}^*$).

The STC algorithm is used for coding.

"Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes", T. Filler, J. Judas, J. Fridrich, TIFS'2011.

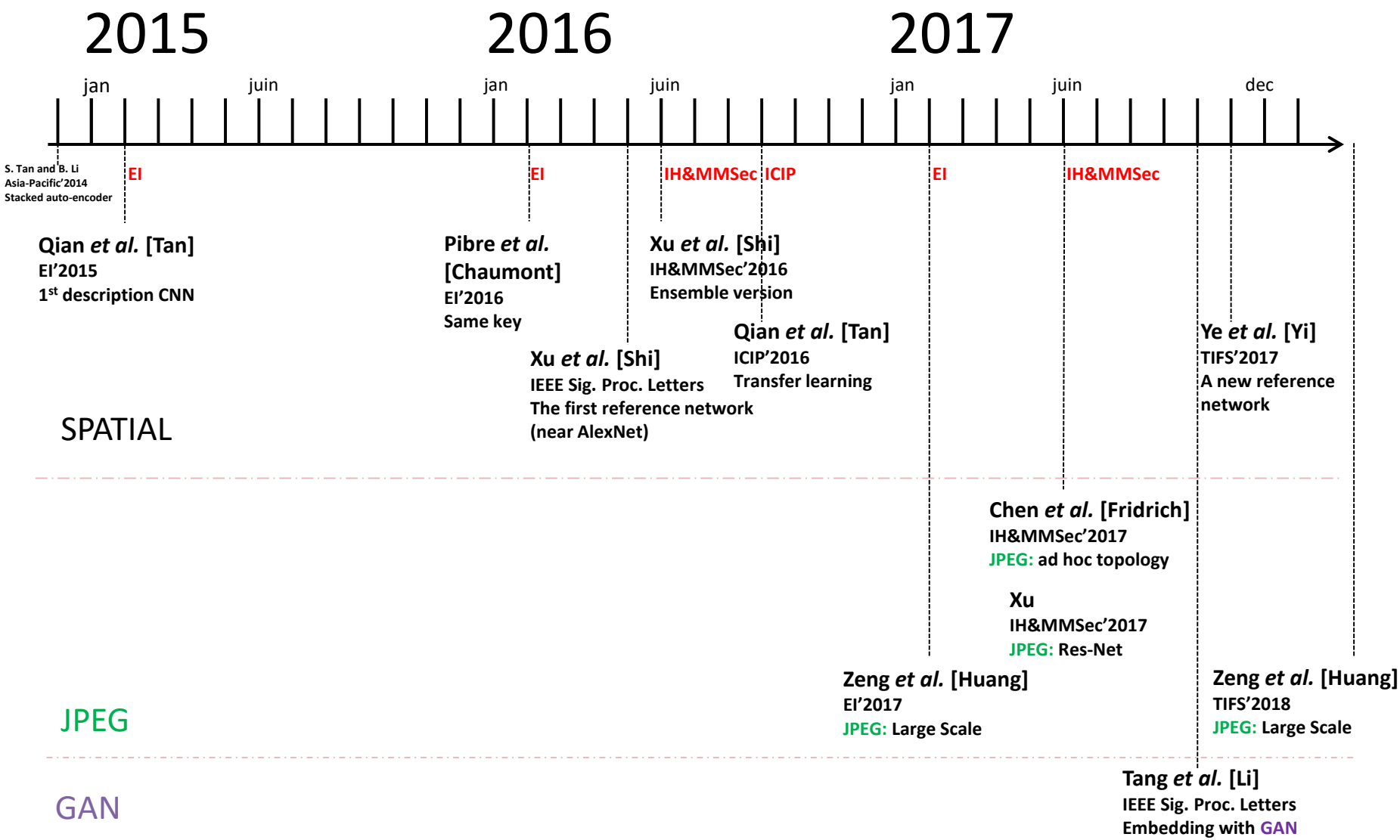
The two families for steganalysis since 2016-2017

- The classic 2-steps learning approach [1,2] vs. the deep learning approach [3, 4]

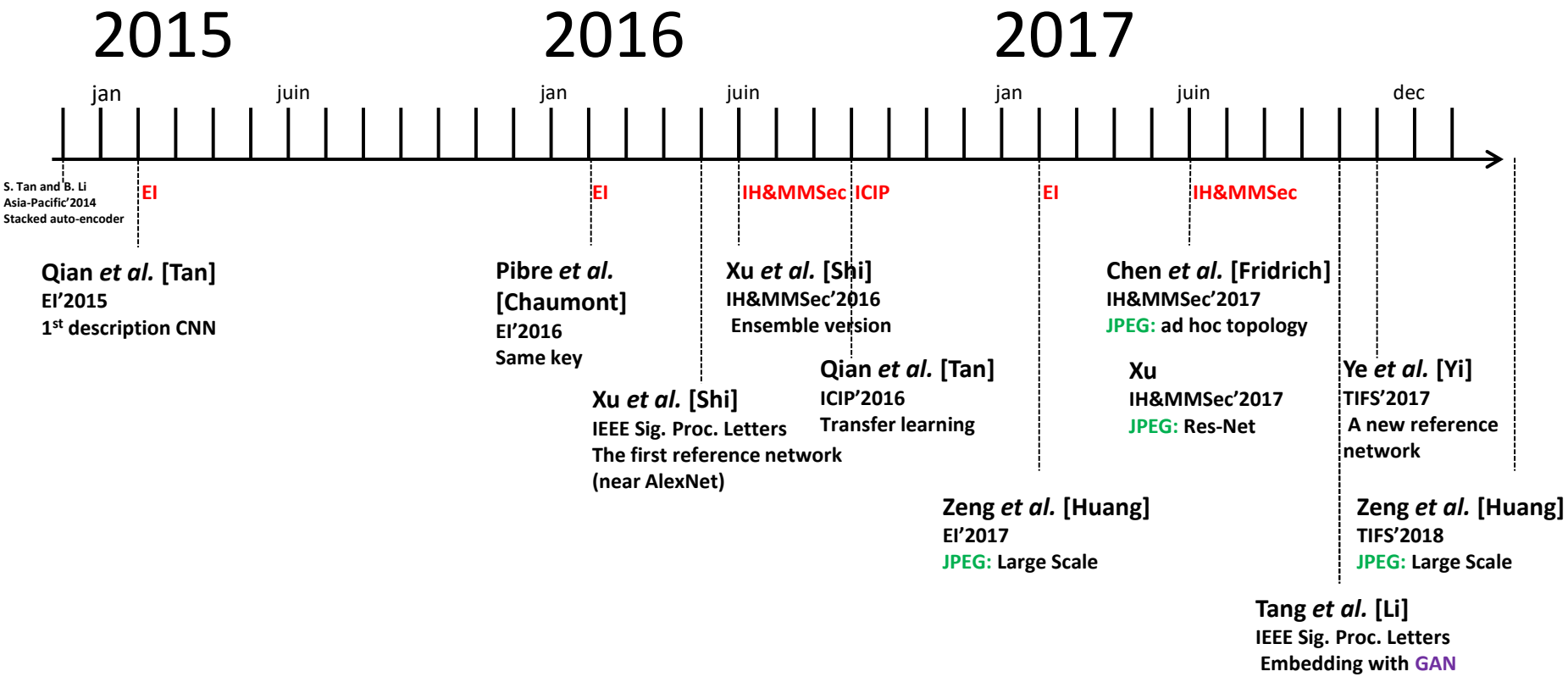


[1]: "Ensemble Classifiers for Steganalysis of Digital Media", J. Kodovský, J. Fridrich, V. Holub, TIFS'2012
[2]: "Rich Models for Steganalysis of Digital Images", J. Fridrich and J. Kodovský, TIFS'2012
[3]: "Structural Design of Convolutional Neural Networks for Steganalysis", G. Xu, H. Z. Wu, Y. Q. Shi, IH&MMSec'2016
[4]: "Deep Learning Hierarchical Representations for Image Steganalysis", J. Ye, J. Ni, Y. Yi, TIFS'2017

The emergence of deep learning (1)



The emergence of deep learning (2)



2018 →

Yedroudj *et al.* [Chaumont]
State-of-the-art

Fuji Tsang and Fridrich
EI'2018
Images of arbitrary size

Chen et al. [Fridrich]
EI'2018
Quantitative Steganalysis

Use of the cost function

Use of Dense-Net, ...?

GAN Methodology

Outline

An example of a Convolutional Neural Network

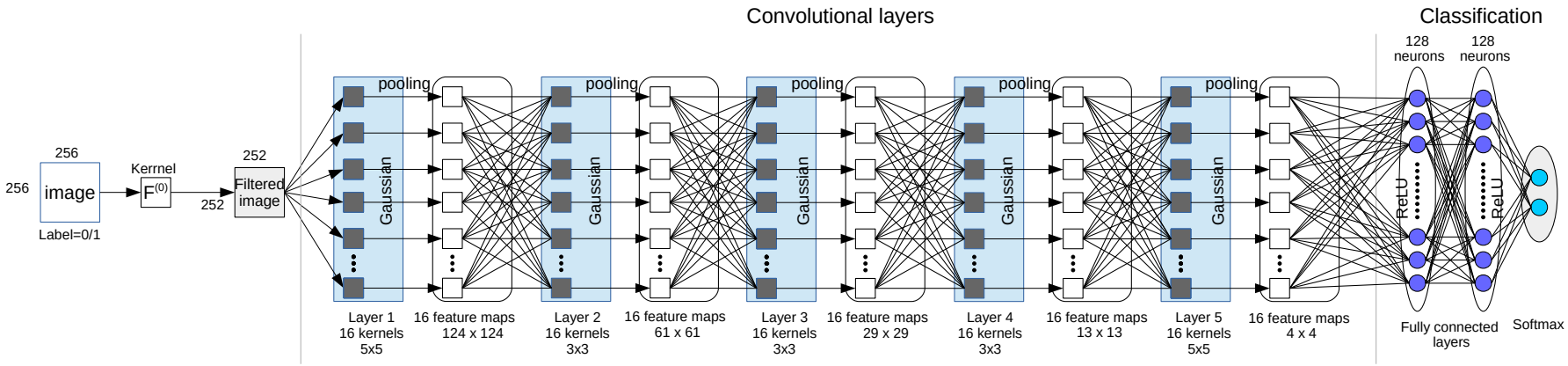


Figure: Qian *et al.* Convolutional Neural Network.

- Inspired by Krizhevsky *et al.*'s CNN 2012,
- Percentage of detection 3 % to 4 % worse than EC + RM.

" ImageNet Classification with Deep Convolutional Neural Networks", A. Krizhevsky, I. Sutskever, G. E. Hinton, NIPS'2012.

"Deep Learning for Steganalysis via Convolutional Neural Networks," Y. Qian, J. Dong, W. Wang, T. Tan, EI'2015.

Convolution Neural Network: Pre-treatment filter

$$F^{(0)} = \frac{1}{12} \begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix}$$

CNNs converge more slowly (or not at all) without this preliminary high-pass filter (except when using the cost map?).

Convolution Neural Network: Layers

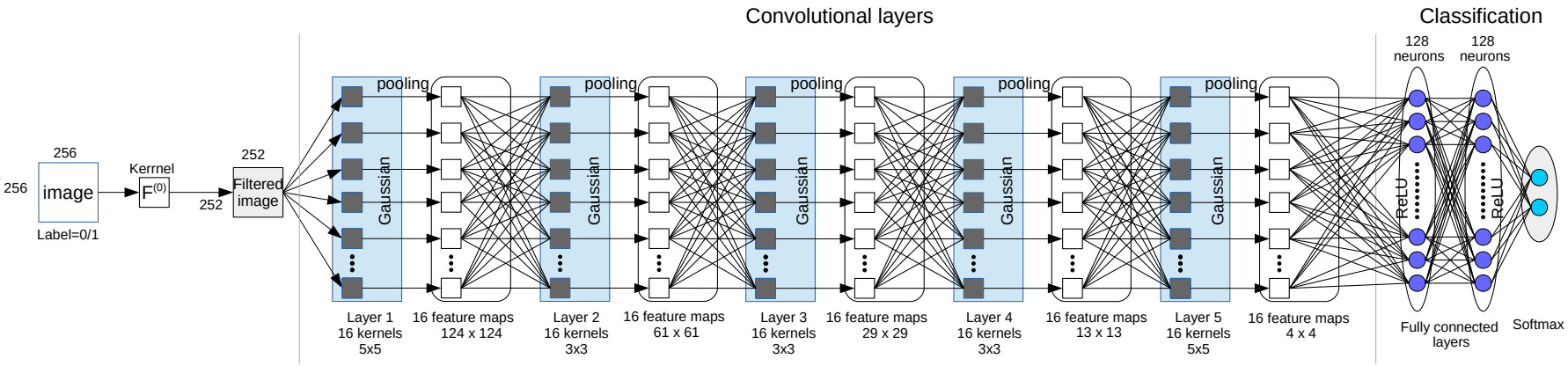


Figure: Qian *et al.* Convolutional Neural Network.

In a layer (a block); the stages:

- A convolution,
- The application of an activation function,
- A pooling step,
- A normalizatrion step.

Convolution Neural Network: Convolutions

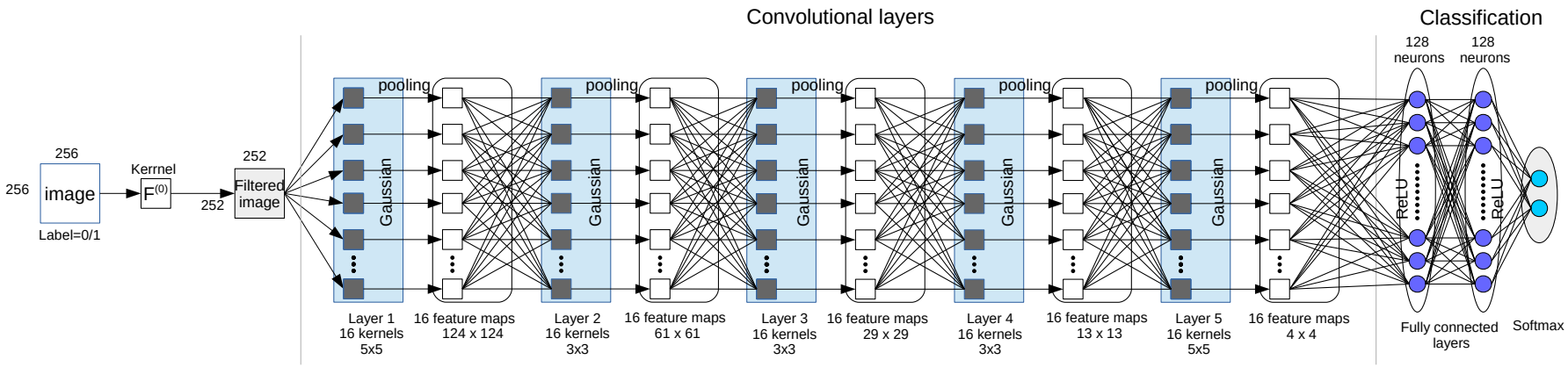


Figure: Qian *et al.* Convolutional Neural Network.

- First layer:

$$\tilde{I}_k^{(1)} = I^{(0)} \star F_k^{(1)}. \tag{1}$$

- Other Layers:

$$\tilde{I}_k^{(l)} = \sum_{i=1}^{i=K^{(l-1)}} I_i^{(l-1)} \star F_{k,i}^{(l)}, \tag{2}$$

Convolution Neural Network: Activation

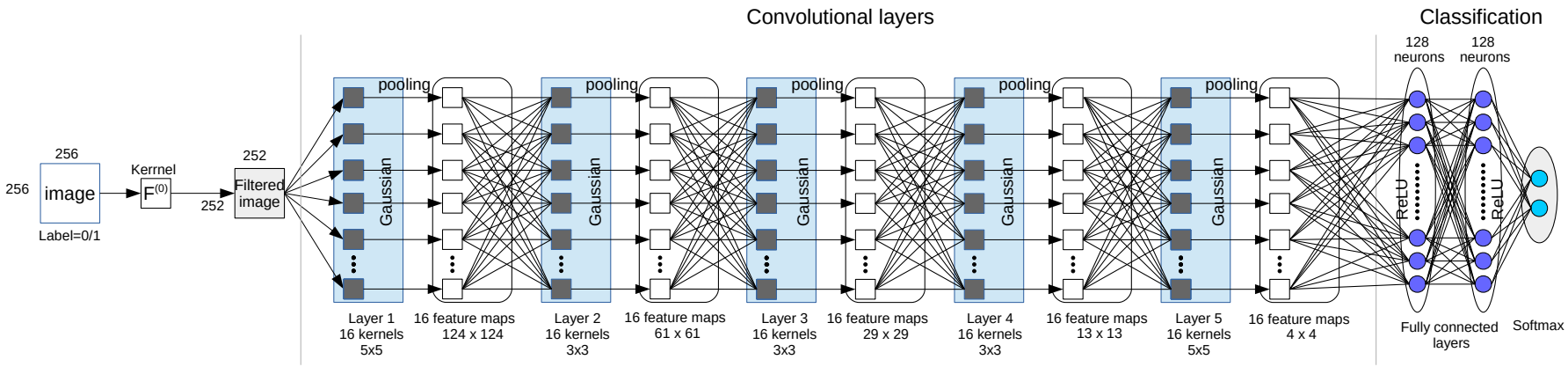


Figure: Qian *et al.* Convolutional Neural Network.

Possible activation functions:

- Absolute function: $f(x) = |x|$,
- Sinus function: $f(x) = sinus(x)$,
- Gaussian function (Qian *et al.*'s network) : $f(x) = \frac{e^{-x^2}}{\sigma^2}$,
- ReLU (Rectified Linear Units) : $f(x) = max(0, x)$,
- Hyperbolic tangent: $f(x) = tanh(x)$ ...

Convolution Neural Network: Pooling

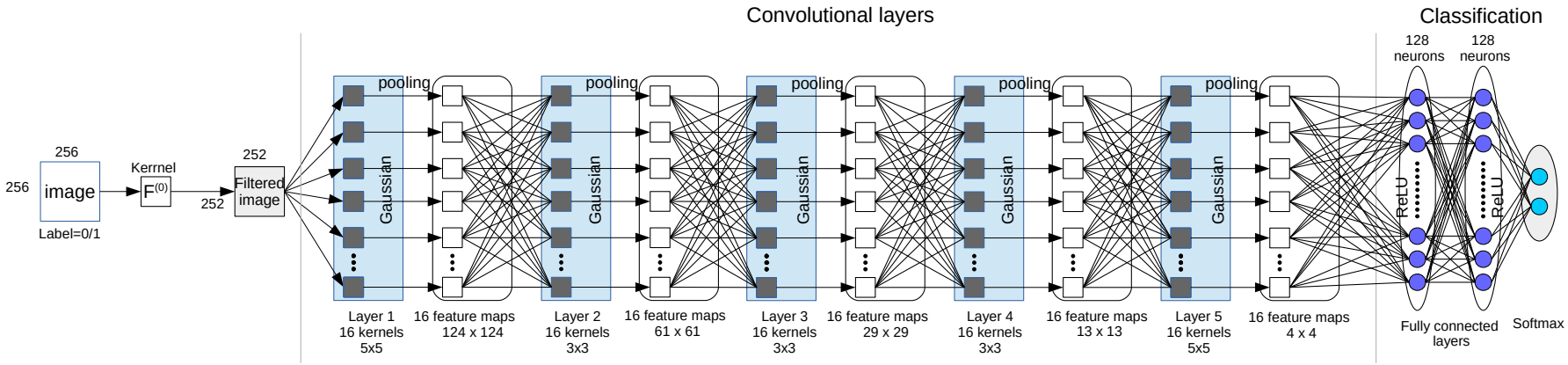


Figure: Qian *et al.* Convolutional Neural Network.

Pooling is a local operation computed on a neighborhood:

- local average (preserve the signal),
- or, local maximum (translation invariance property).

+ a sub-sampling operation.

Convolution Neural Network: Normalization

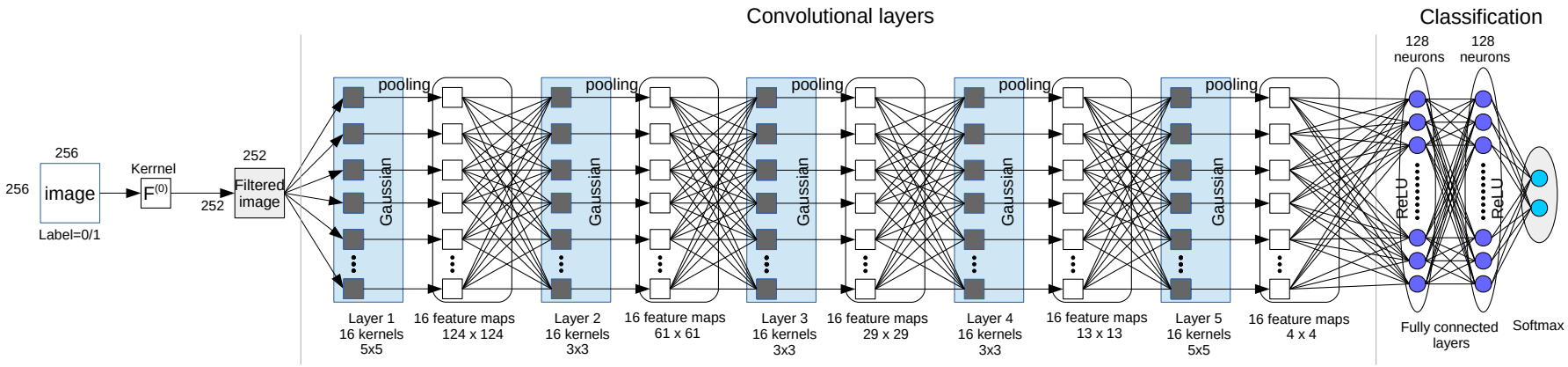


Figure: Qian *et al.* Convolutional Neural Network.

Example: Case where normalization is done between "features maps":

$$norm(I_k^{(1)}(x,y)) = \frac{I_k^{(1)}(x,y)}{\left(1 + \frac{\alpha}{size} \sum_{k'=max(0,k-\lfloor size/2 \rfloor)}^{k'=min(K,k-\lfloor size/2 \rfloor+size)} (I_{k'}^{(1)}(x,y))^2\right)^{\beta}}$$

Convolution Neural Network: Fully Connected Network

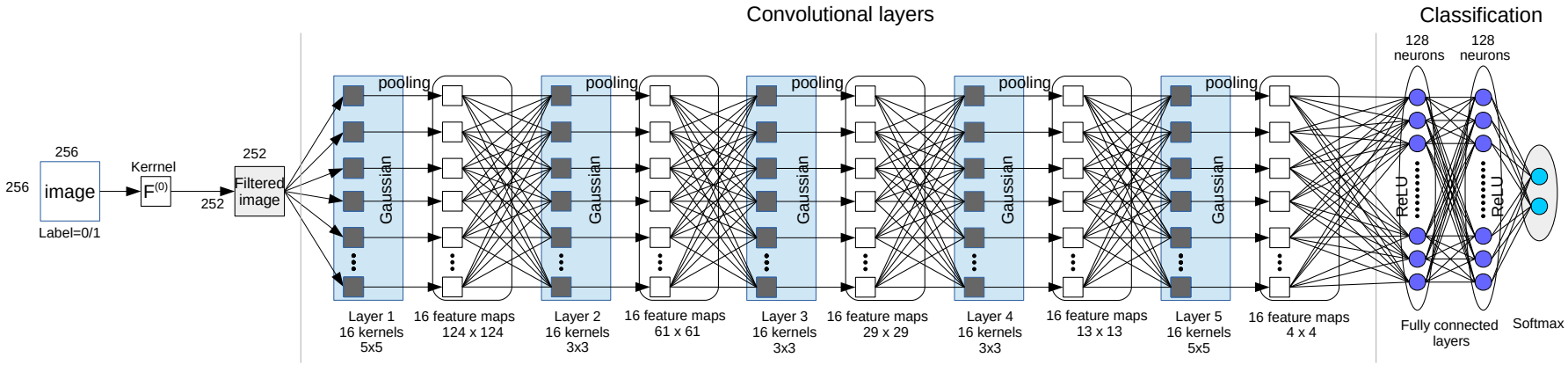


Figure: Qian *et al.* Convolutional Neural Network.

- Three layers,
- A softmax function normalizes the values between $[0, 1]$,
- The network issues a value for cover (resp. for stego).

Other networks "references"

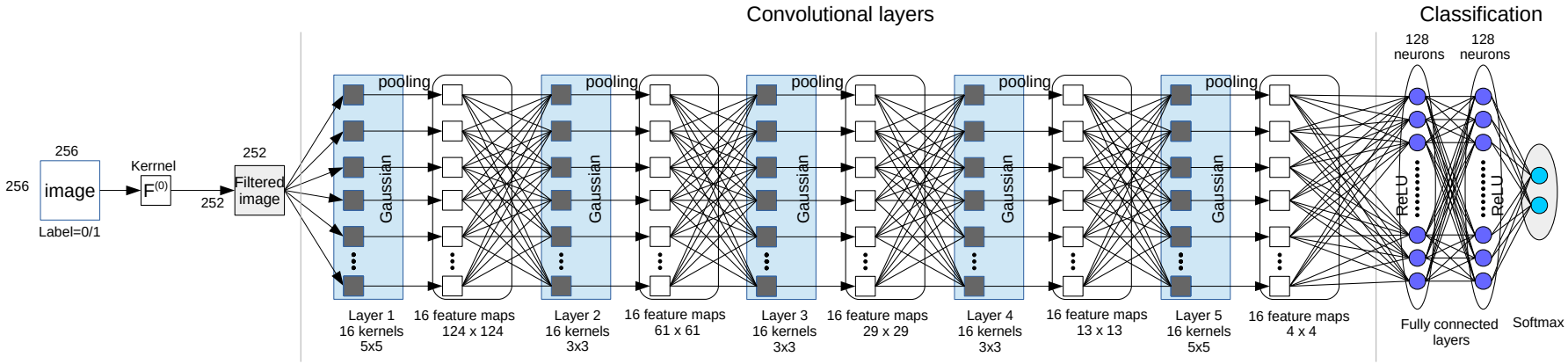


Figure: Qian *et al.* Convolutional Neural Network.

- Xu-Net (may 2016):
  - Absolute value (first layer),
  - Activation function: TanH and ReLU,
  - Normalization function: Batch Normalization (2015),
  - Specific order.
- Ye-Net (nov. 2017):
  - Filters bank,
  - Activation function (truncature = "hard tanh"),
  - 8 "layers" and only convolutions,
  - A version that uses a cost map.

Outline

A new network

Yedroudj-Net

Aggregation of the "most efficient" bricks of newly designed CNNs.
Objective: To have a basic CNN (baseline) at the state of the art.

The essential elements of the network:

- A high-pass filters bank for pre-processing (SRM [1]),
- A truncation activation function ("hard tanh") [2],
- The "batch normalization" associated with a "scaling" layer [3][4][5].

[1]: "Ensemble Classifiers for Steganalysis of Digital Media", J. Kodovský, J. Fridrich, V. Holub, TIFS'2012,
[2]: "Deep Learning Hierarchical Representations for Image Steganalysis", J. Ye, J. Ni, Y. Yi, TIFS'2017,
[3]: "BN: Accelerating deep network training by reducing internal covariate shift", S. Ioffe, C. Szegedy, ICML'2015,
[4]: "Deep residual learning for image recognition", K. He, X. Zhang, S. Ren, J. Sun, CVPR'2016,
[5]: "Structural Design of Convolutional Neural Networks for Steganalysis", G. Xu, H. Z. Wu, Y. Q. Shi, IH&MMSec'2016.

Yedroudj-Net

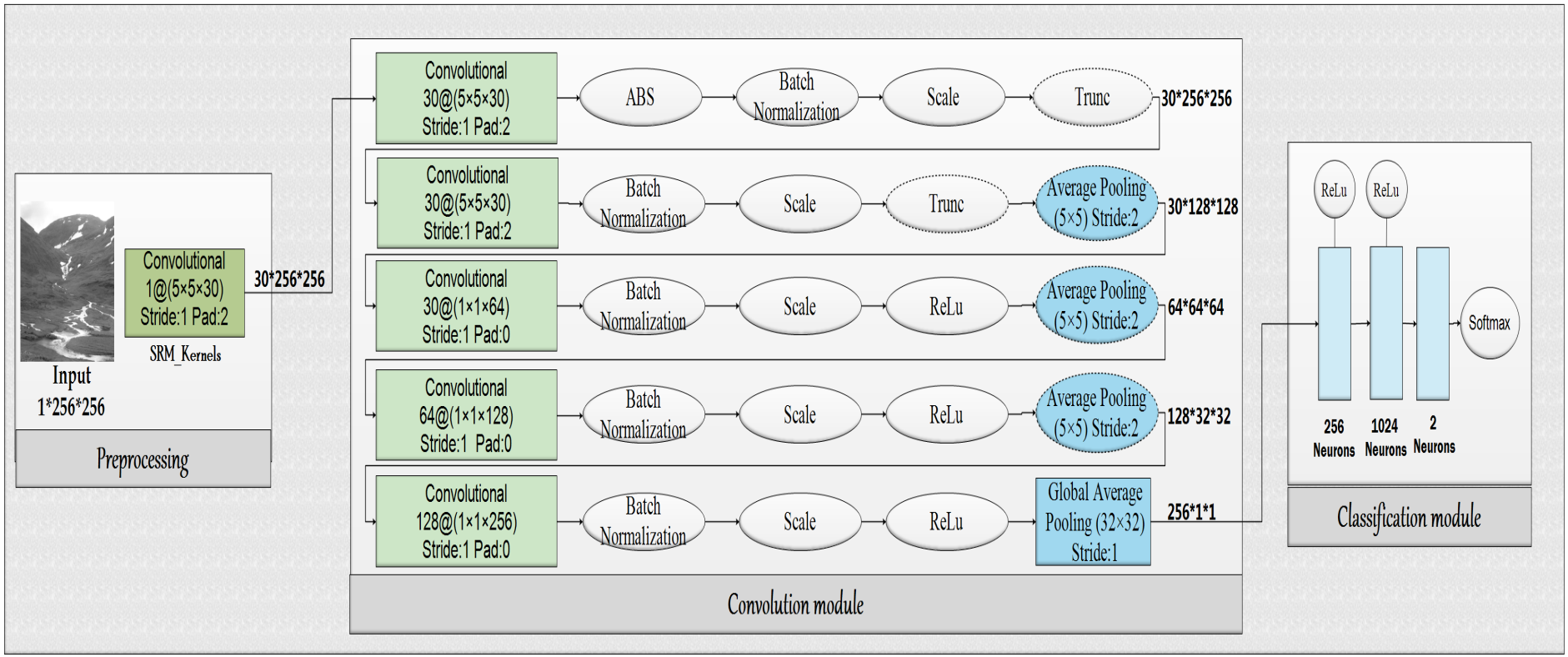
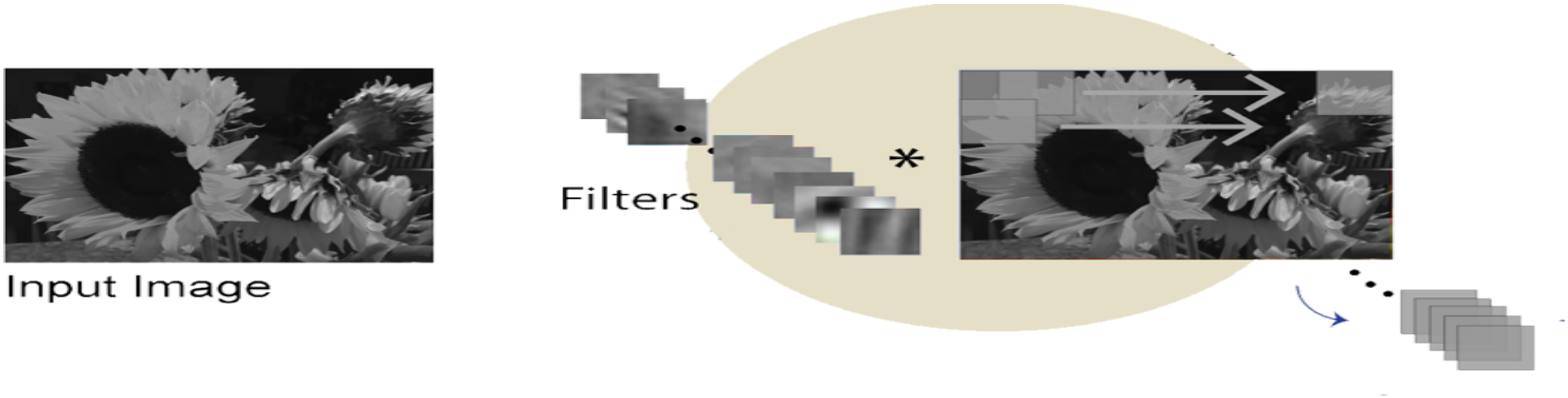


Figure: Yedroudj-Net

Filters

High pass filters

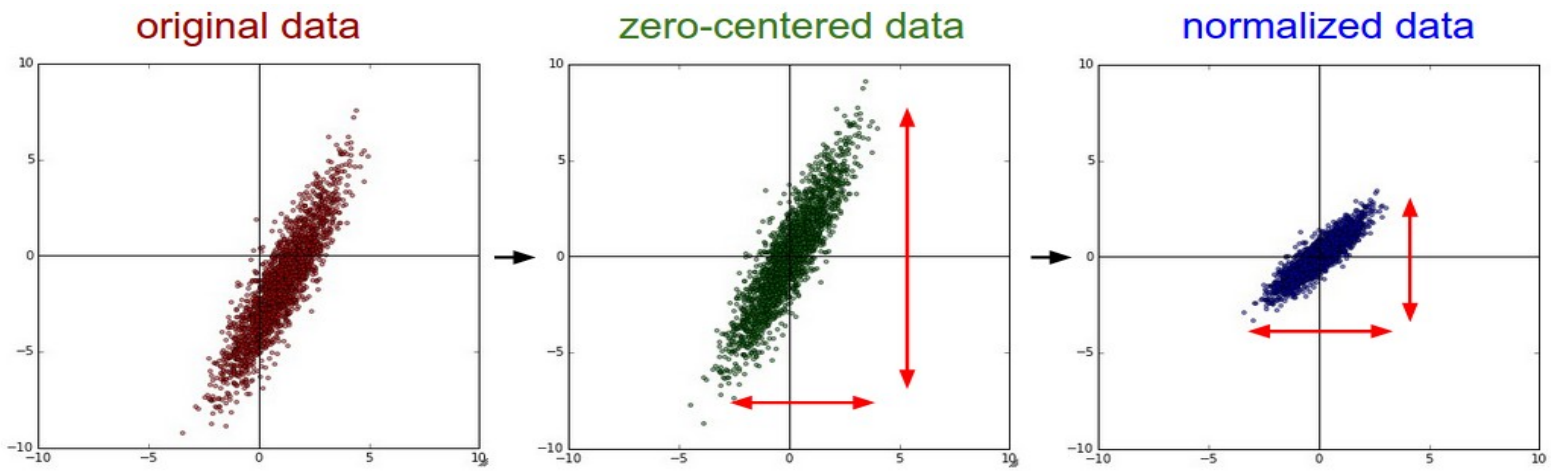
- In SRM (= features) there is pre-processing of images with a high-pass filter bank to extract the stego noise [1].
- In Yedroudj-Net, there is pre-processing of images with a bank of 30 high-pass filters (pre-processing block) [2].



[1]: "Ensemble Classifiers for Steganalysis of Digital Media", J. Kodovský, J. Fridrich, and V. Holub, TIFS'2012,

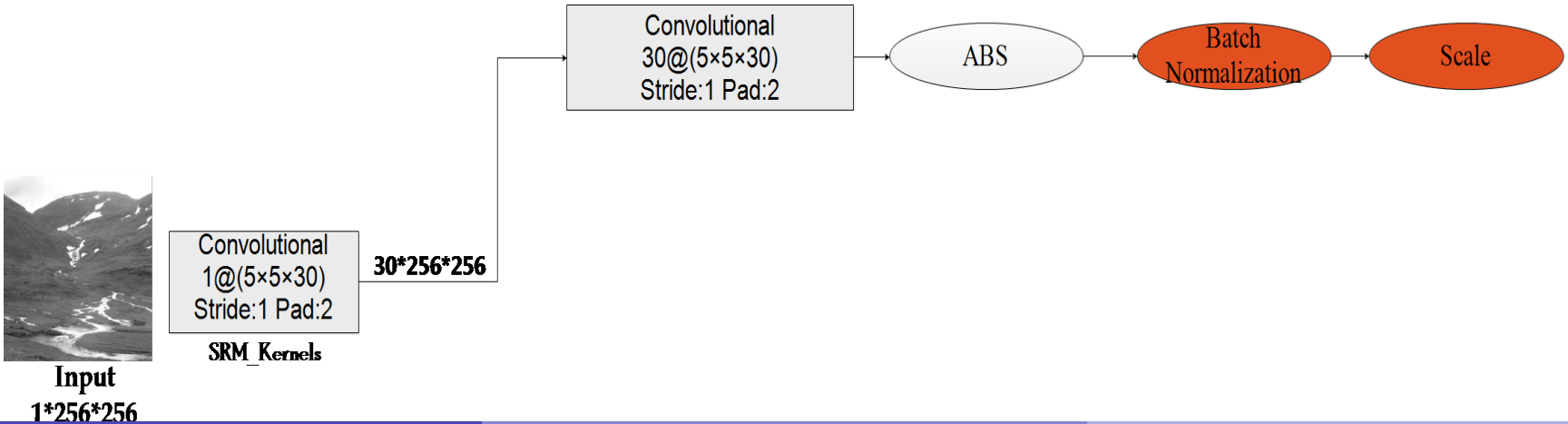
[2]: "Deep Learning Hierarchical Representations for Image Steganalysis", J. Ye, J. Ni, and Y. Yi, TIFS'2017.

Batch Normalization & Scale



Batch Normalization

$$BN(X,\gamma,\beta) = \beta + \gamma \frac{X - E[X]}{\sqrt{Var[X] + \epsilon}}, \text{ [3][5]}$$

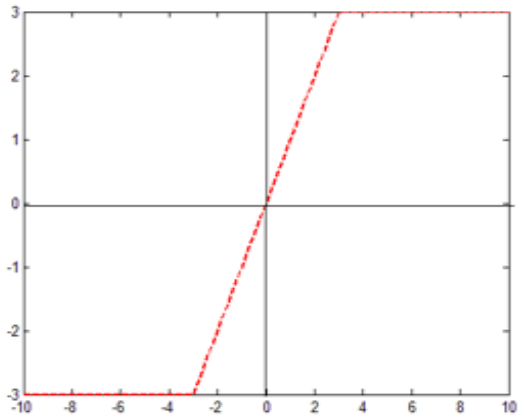


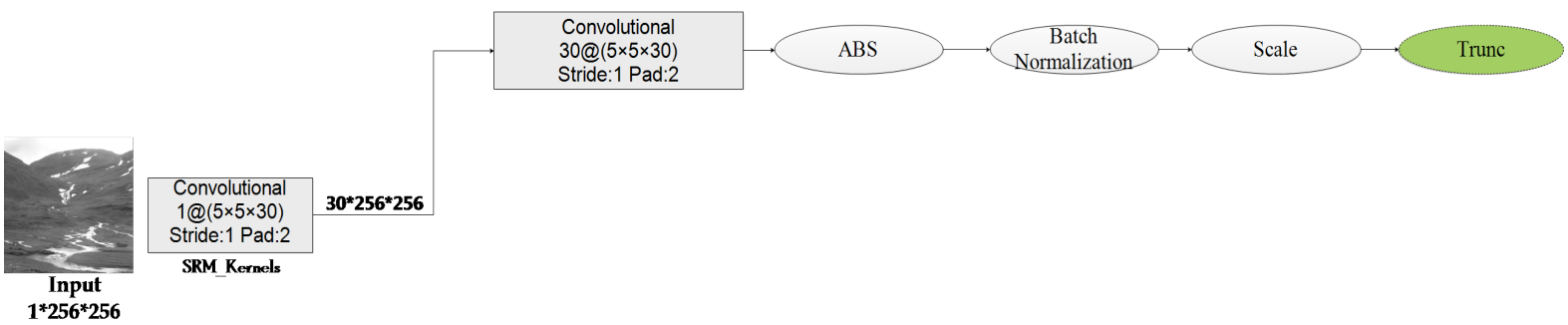
Activation function: Truncation (hard tanh)

Yedroudj-Net:

Activation function 'tuncation' for the first 2 blocks.
Limits the value range and prevents network modeling of large values.

$$Trunc_T(x) = \begin{cases} -T, & x < -T, \\ x, & -T \leq x \leq T, \\ T, & x > T. \end{cases}$$





Experimental protocol

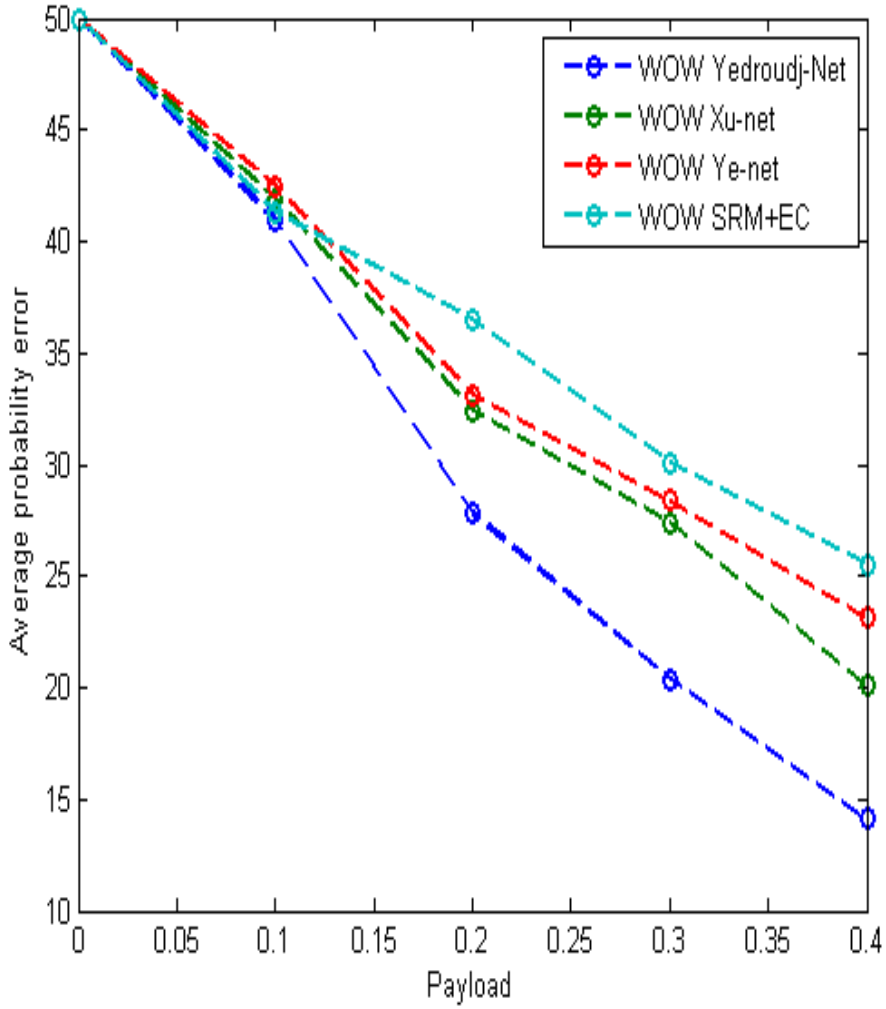
Clairvoyant protocol

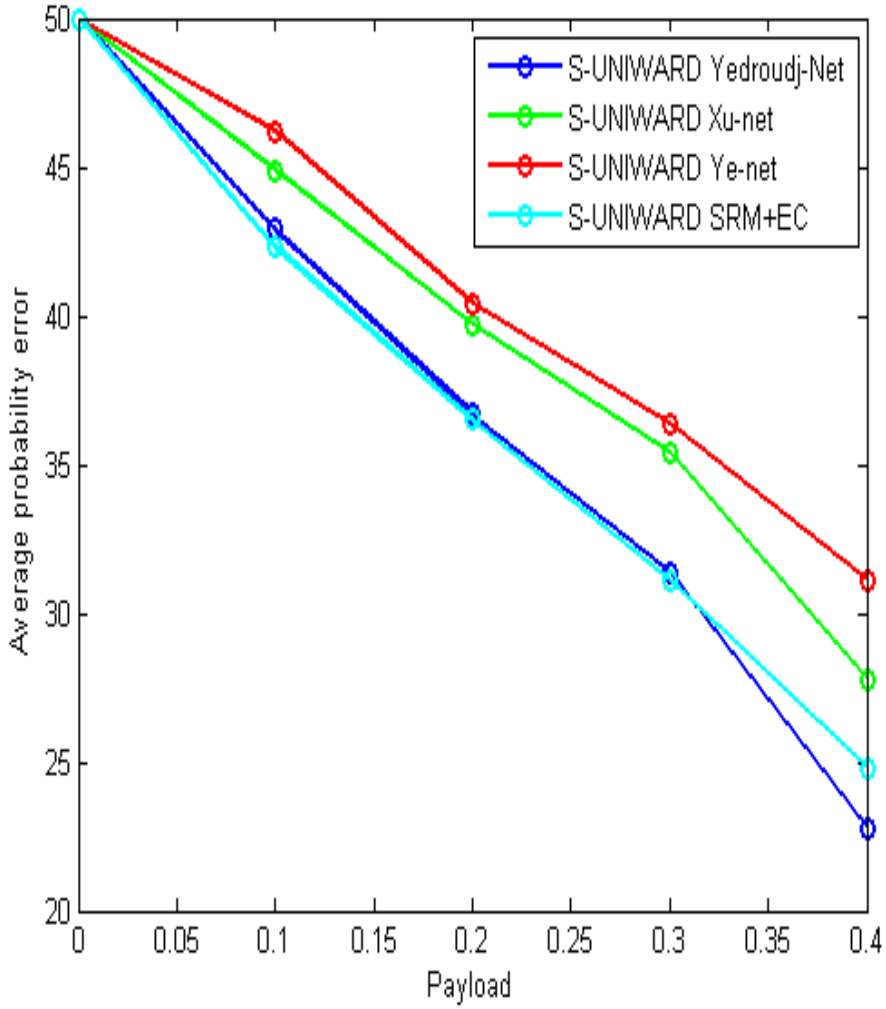
- Resize the 10 000 images of BOSSBase from 512×512 to 256×256,
- Using embedding algorithms WOW [1] and S-UNIWARD [2] to generate the stegos (Matlab Version),
- Selection of 5 000 pairs for learning including 1 000 pairs for validation,
- The other 5 000 pairs are used for the test (evaluation).

[1] "Designing Steganographic Distortion Using Directional Filters", V. Holub, J. Fridrich, WIFS'2012.

[2] "Universal Distortion Function for Steganography in an Arbitrary Domain", V. Holub, J. Fridrich, T. Denemark, JIS'2014.

Résults and discussion





Outline

Performance improvements:

- Virtual Augmentation [Krizhevsky 2012]
- Transfer Learning [Qian et al. 2016],
- Using Ensemble [Xu et al. 2016],
- Learn with millions of images? [Zeng et al. 2018],
- Add images from the same cameras and with the same "development" [Ye et al. 2017], [Yedroudj et al. EI'2018],
- New networks [Yedroudj et al. ICASSP'2018], ResNet, DenseNet, ...
- ...

"ImageNet Classification with Deep Convolutional Neural Networks", A. Krizhevsky, I. Sutskever, G. E. Hinton, NIPS'2012,
"Learning and transferring representations for image steganalysis using convolutional neural network", Y. Qian, J. Dong, W. Wang, T. Tan, ICIP'2016,
"Ensemble of CNNs for Steganalysis: An Empirical Study", G. Xu, H.-Z. Wu, Y. Q. Shi, IH&MMSec'16,
"Large-scale jpeg image steganalysis using hybrid deep-learning framework", J. Zeng, S. Tan, B. Li, J. Huang, TIFS'2018,
"Deep Learning Hierarchical Representations for Image Steganalysis," J. Ye, J. Ni, and Y. Yi, TIFS'2017,
"How to augment a small learning set for improving the performances of a CNN-based steganalyzer?", M. Yedroudj, F. Comby, M. Chaumont, EI'2018,
"Yedroudj-Net: An Efficient CNN for Spatial Steganalysis", M. Yedroudj, F. Comby, M. Chaumont, IEEE ICASSP'2018.

Enrichment of the learning base (1) [Yedroudj et al. EI'2018]:

"How to augment a small learning set for improving the performances of a CNN-based steganalyzer?", M. Yedroudj, F. Comby, M. Chaumont, EI'2018.

Clairvoyant protocol

- Resize the 10 000 images of BOSSBase from 512×512 to 256×256,
- Using embedding algorithms WOW [1] and S-UNIWARD [2] to generate the stegos (Matlab Version),
- Selection of 5 000 pairs for learning including 1 000 pairs for validation,
- The other 5 000 pairs are used for the test (evaluation).

[1] "Designing Steganographic Distortion Using Directional Filters", V. Holub J. Fridrich, WIFS'2012.
[2] "Universal Distortion Function for Steganography in an Arbitrary Domain", V. Holub, J. Fridrich, T. Denemark, JIS'2014.

Enrichment of the learning base (2) [Yedroudj et al. EI'2018]:

"How to augment a small learning set for improving the performances of a CNN-based steganalyzer?", M. Yedroudj, F. Comby, M. Chaumont, EI'2018.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS	**27.8** %	**36.7** %
BOSS+VA	24.2 %	34.8 %
BOSS+all-DEV	23.0 %	33.2 %
BOSS+BOWS2	23.7 %	34.4%
BOSS+BOWS2+VA	20.8 %	31.1 %

Table: Probability of error for Yedroudj-Net with different enrichments

- BOSS+VA: 32 000 pairs, BOSS+all-DEV: 44 0000 pairs, BOSS+BOWS2: 14 000 pairs, BOSS+BOWS2+VA: 112 000 pairs,
- Experiments versus EC+RM, versus Xu-Net, versus Ye-Net,
- Experiments counterproductive enrichment (different cameras, change ratio, ...).

A conjecture (rule for the increase):

"How to augment a small learning set for improving the performances of a CNN-based steganalyzer?", M. Yedroudj, F. Comby, and M. Chaumont, EI'2018.

Given a target database:

- either Eve (the steganalyst) finds the same camera(s) (used for generating the target database), capture new images, and reproduce the same development than the target database, with a special caution to the resizing,
- either Eve has an access to the original RAW images and reproduce similar developments than the target database with the similar resizing,

The reader should also remember that the Virtual Augmentation is also a good cheap processing measure.

Outline

A few words about ASDL-GAN:

[Tang et al. 2017] "Automatic steganographic distortion learning using a generative adversarial network", W. Tang, S. Tan, B. Li, and J. Huang, IEEE Signal Processing Letter, Oct. 2017

- CNN "simulating" an embedding in a spatial image,
- CNN called Generator (denoted G) generates the map of modifications (-1 / 0 / +1),
- G learns to embed through a "competition" (GAN methodology) between it and a Discriminator (noted D).

GAN [Goodfellow 2014] "Generative Adversarial Networks", I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, Sherjil Ozair, A. Courville, Y. Bengio NIPS'2014
ASO [Kouider 2013] "Adaptive Steganography by Oracle (ASO)", S. Kouider and M. Chaumont, W. Puech, ICME'2013.
ASO [Kouider 2012] "Technical Points About Adaptive Steganography by Oracle (ASO)", S. Kouider, M. Chaumont, W. Puech, EUSIPCO'2012.

A few words about ASDL-GAN:

[Tang et al. 2017] "Automatic steganographic distortion learning using a generative adversarial network", W. Tang, S. Tan, B. Li, and J. Huang, IEEE Signal Processing Letter, Oct. 2017

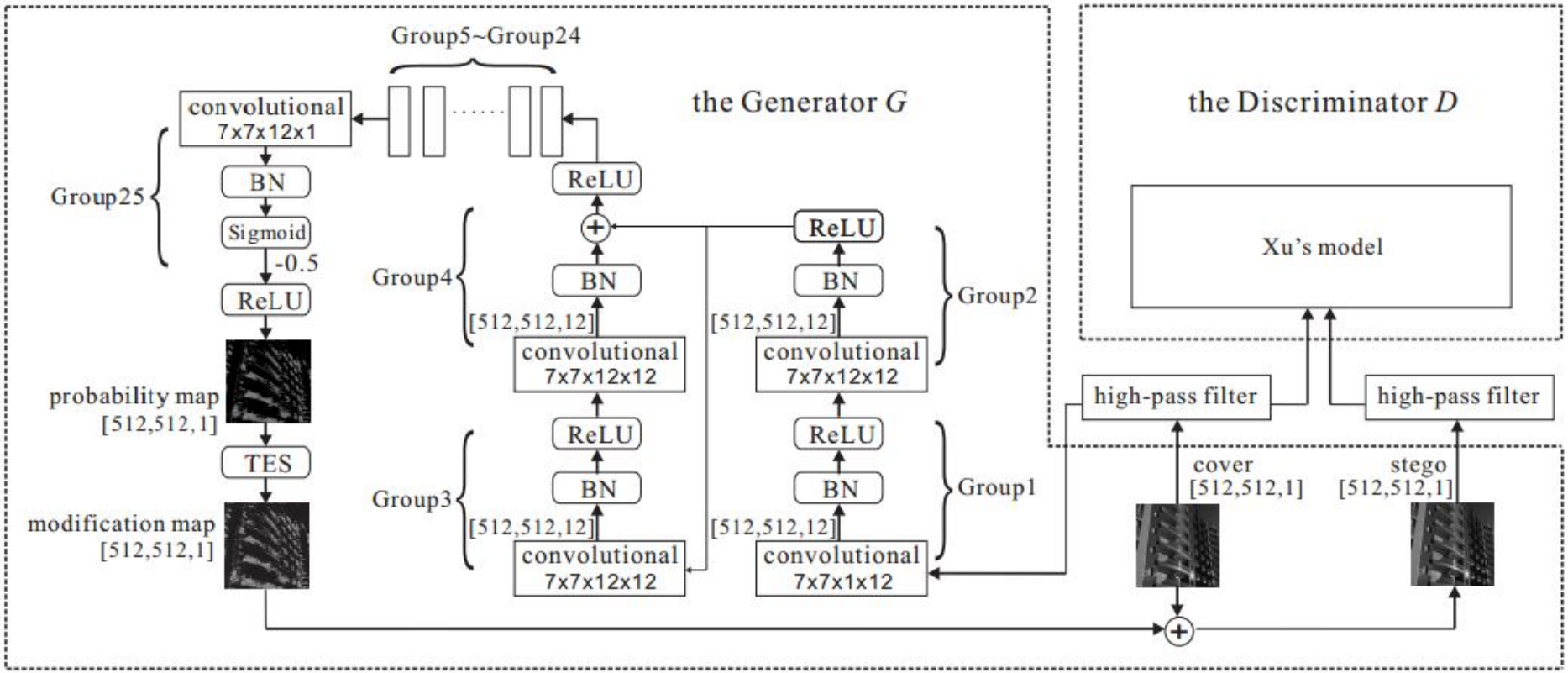


Figure: ASDL-GAN; Figure extracted from the paper [Tang et al. 2017]

Outline

Conclusion

We saw:

- CNN spatial steganalysis (Xu-Net, Ye-Net, Yedroudj-Net),
- CNN JPEG steganalysis (JPEG Xu-Net based on ResNet),
- Database enrichment (one of the enhancement techniques for CNN),
- The GAN steganography.

2018, Installation and study of other scenarios:

- Enrichment,
- Quantitative,
- Variable size of images,
- Cover-Source mismatch,
- GAN.

End of talk

CNN is the new state-of-the-art steganalysis tool ...
... there is still things to do...