L'émergence du Deep Learning en stéganographie et stéganalyse

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

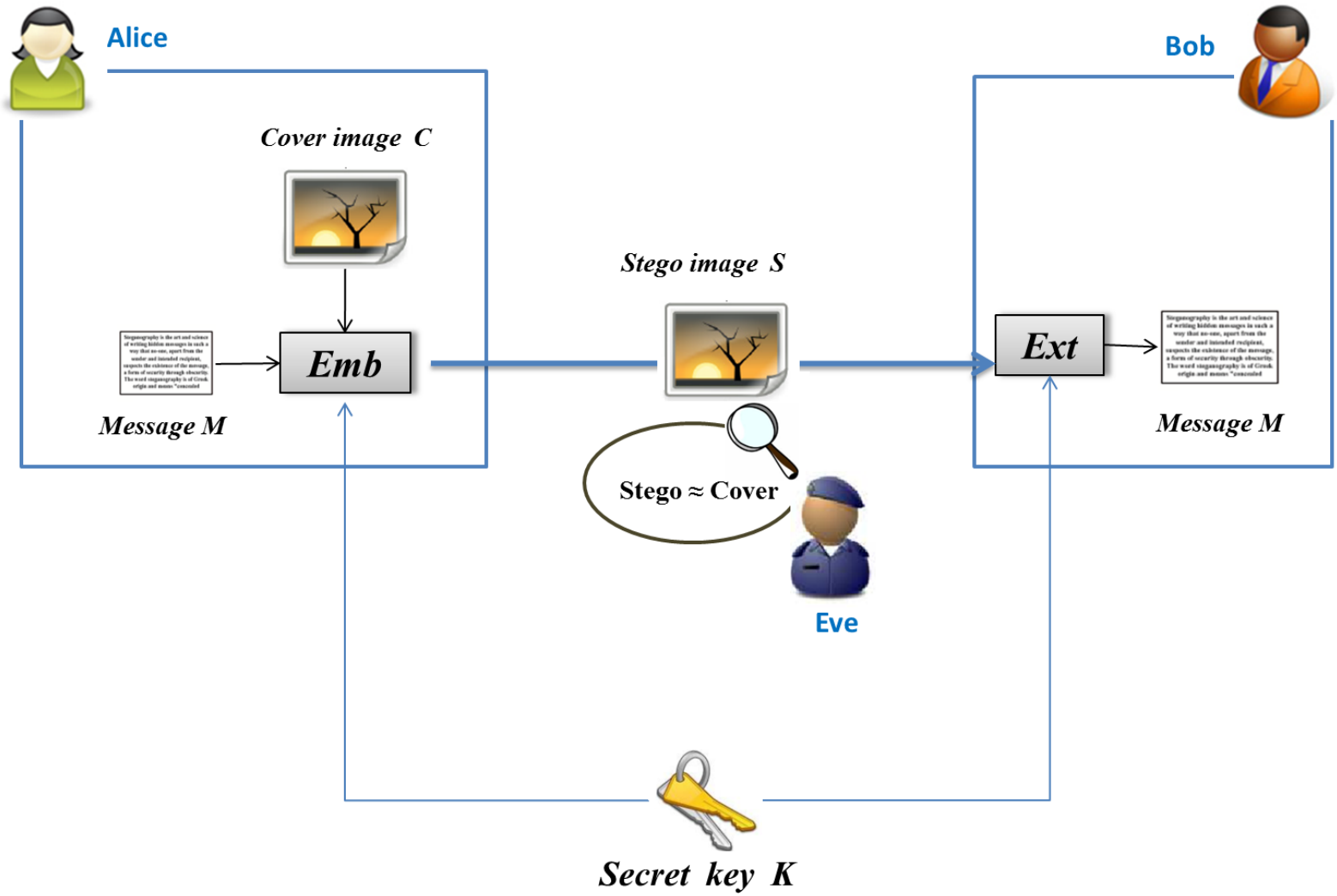
February 3, 2018

Journée Stéganalyse : Enjeux et Méthodes.
Labelisée par le GDR ISIS et le pré-GDR sécurité,

A Poitiers, le 16 janvier 2018.

Plan

Steganographie / Steganalyse



Exemple d'insertion

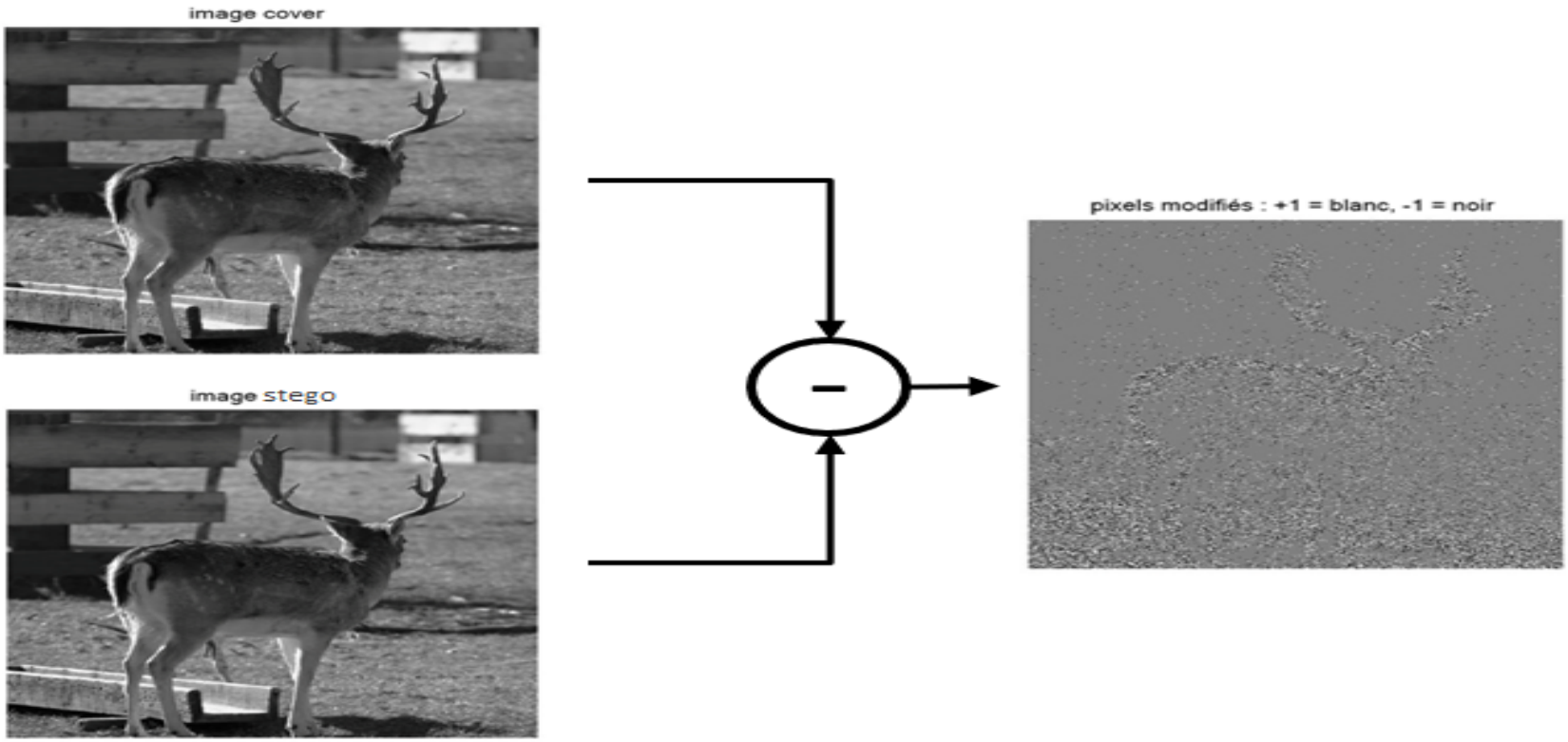


Figure: Exemple d'insertion avec l'algorithme S-UNIWARD (2013) à 0.4 bpp

L'insertion très rapidement...

Plus précisément :

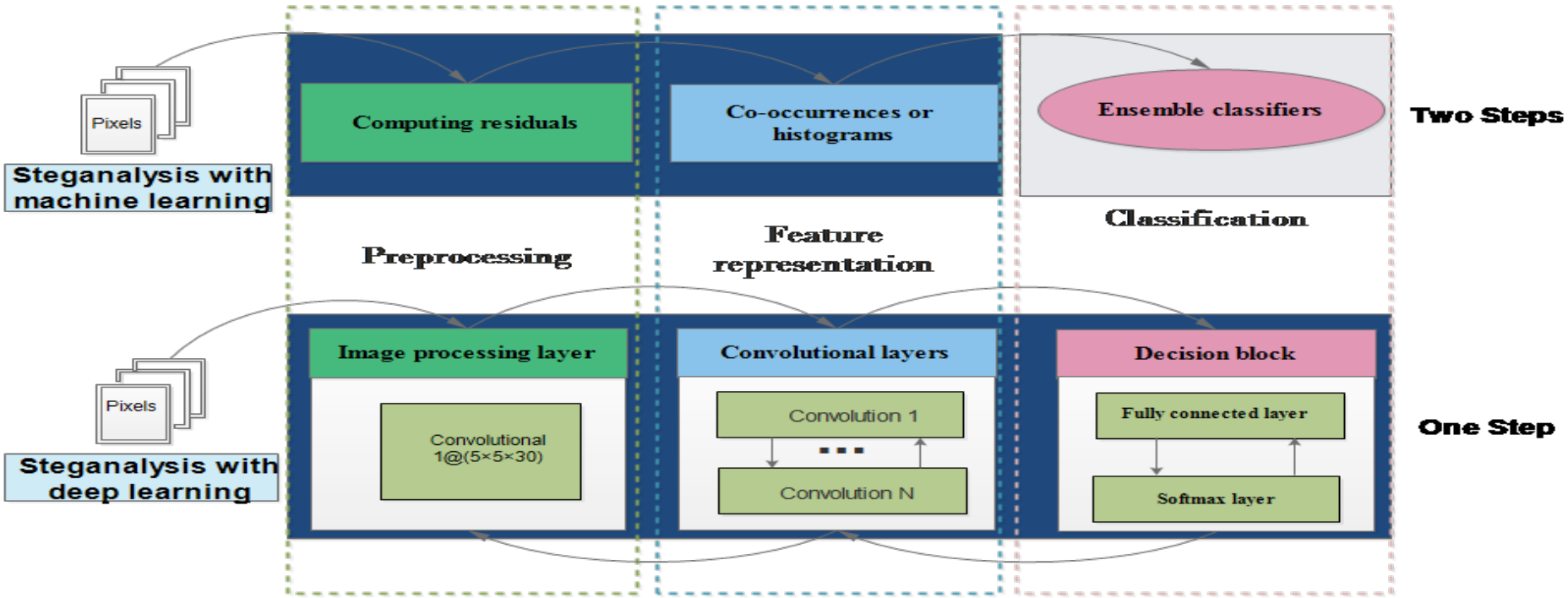
- $\mathbf{m} \Longrightarrow \mathbf{c}^*$, tel que $\mathbf{c}^*$ est l'un des mots-de-code dont le syndrome $= \mathbf{m}$, et tel qu'il minimise la fonction de coût,
- Ensuite, le stego $\leftarrow$ LSB-Matching(cover, $\mathbf{c}^*$).

On utilise pour faire le codage l'algorithme STC.

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

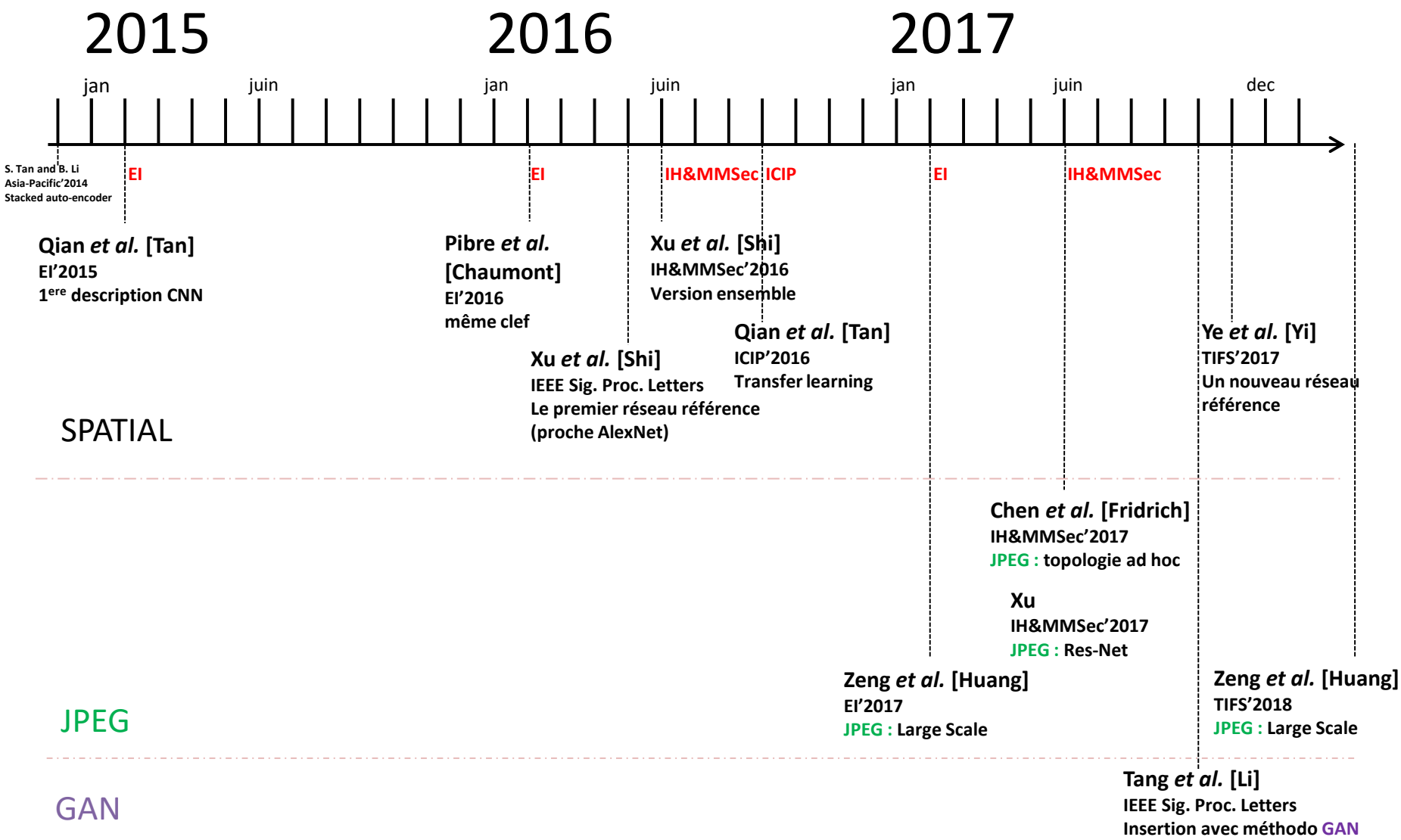
Les deux grandes familles pour la steganalyse depuis 2016-2017

- L'approche d'apprentissage classique en 2 étapes [1,2] vs. l'approche par deep learning [3, 4]

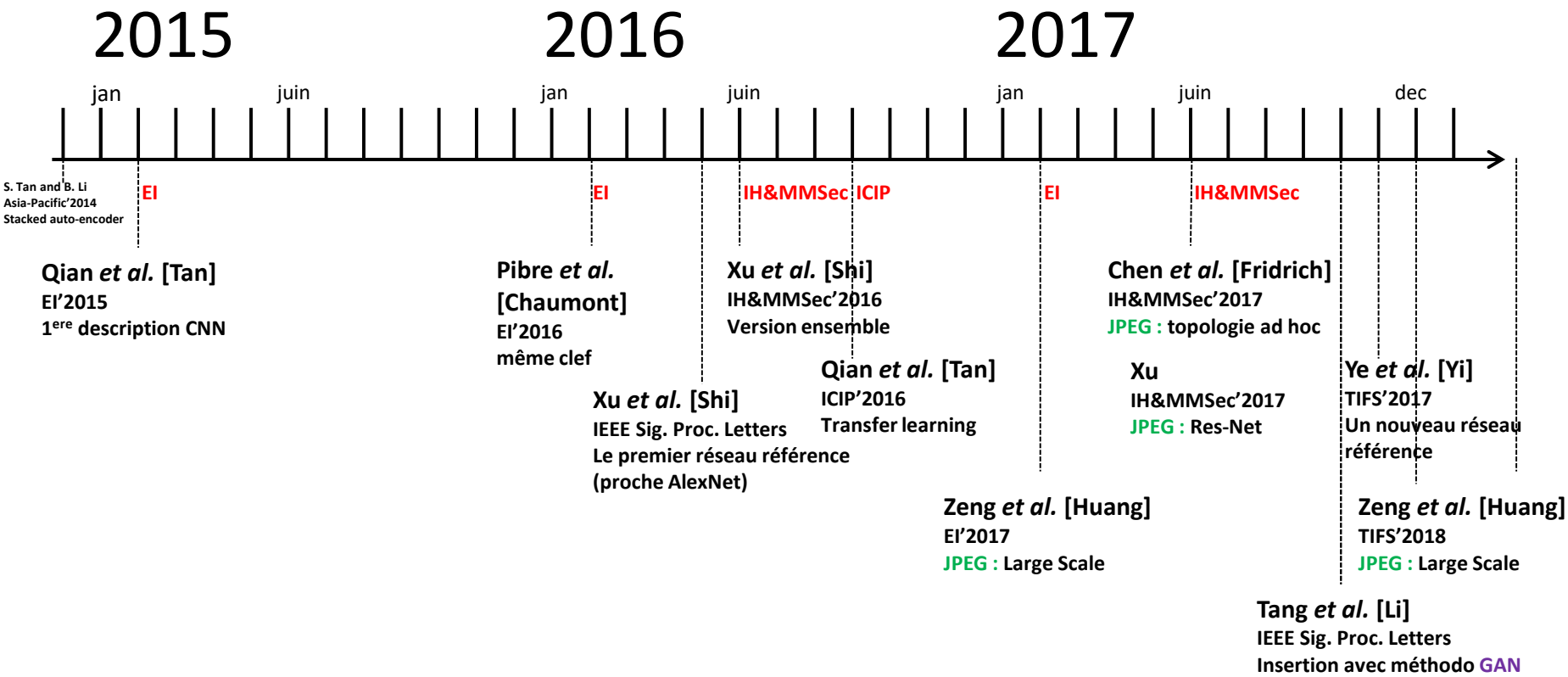


[1]: "Ensemble Classifiers for Steganalysis of Digital Media", J. Kodovský, J. Fridrich, V. Holub, TIFS'2012
[2]: "Rich Models for Steganalysis of Digital Images", J. Fridrich and J. Kodovský, TIFS'2012
[3]: "Structural Design of Convolutional Neural Networks for Steganalysis", G. Xu, H. Z. Wu, Y. Q.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

L'émergence de l'apprentissage profond (1)



L'émergence de l'apprentissage profond (2)



2018 →

Yedroudj *et al.* [Chaumont]
Etat-art

Fuji Tsang and Fridrich
EI'2018
Images taille arbitraires

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

Utilisation de la carte de détectabilité

Utilisation de Dense-Net, ...?

Méthodologie GAN

Plan

Un exemple de réseau de neuronnes convolutionel

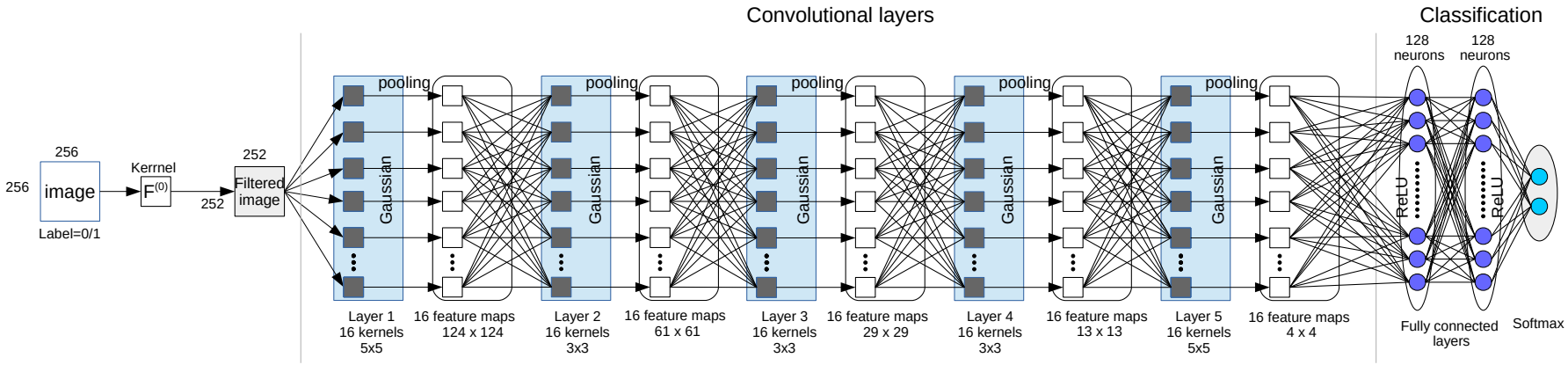


Figure: Qian *et al.* Convolutional Neural Network.

- Inspiré du CNN de Krizhevsky *et al.* 2012,
- Pourcentage de détection 3% à 4% moins bon que 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: Filtre de pré-traitement

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

Le CNNs converge plus lentement (voir pas du tout) sans ce filtre passe-haut préliminaire (sauf utilisation de la carte de détectabilité ?).

Convolution Neural Network: Couches

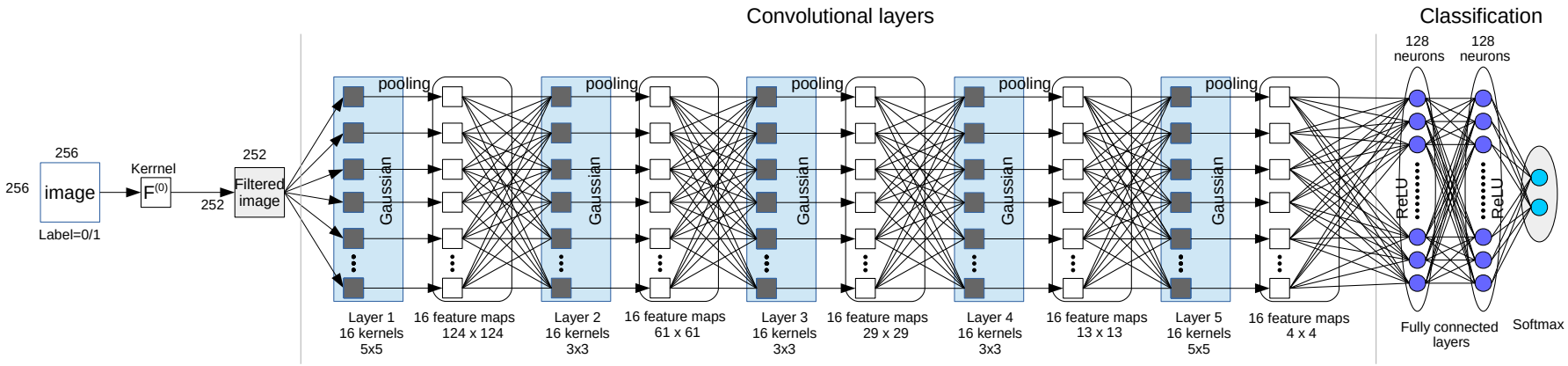


Figure: Qian *et al.* Convolutional Neural Network.

Dans une couche (un bloc); les étapes successives :

- Une convolution,
- L'application d'une fonction d'activation,
- Une étape de pooling,
- Une étape de normalisation.

Convolution Neural Network: Convolutions

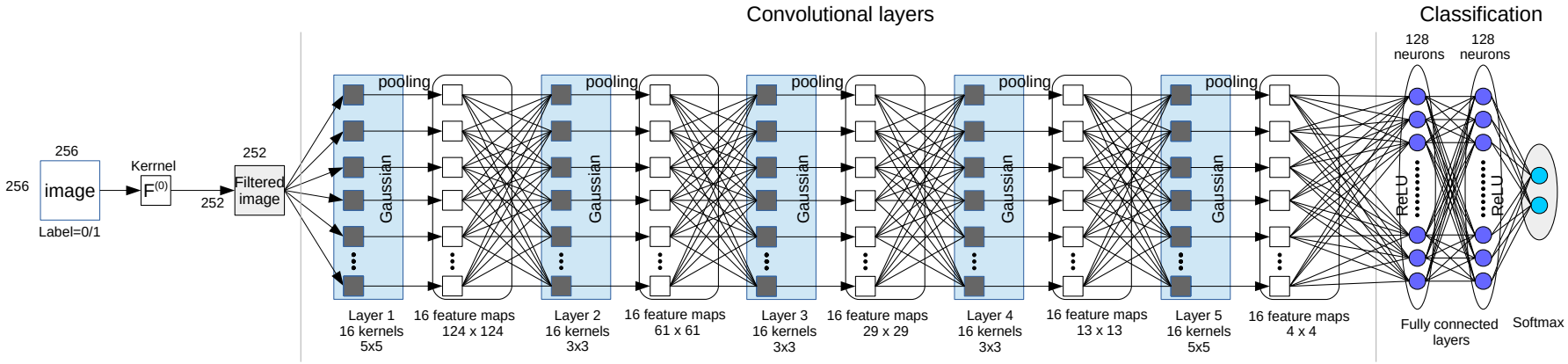


Figure: Qian *et al.* Convolutional Neural Network.

- Première couche :

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

- Autres couches :

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

Convolution Neural Network: Activation

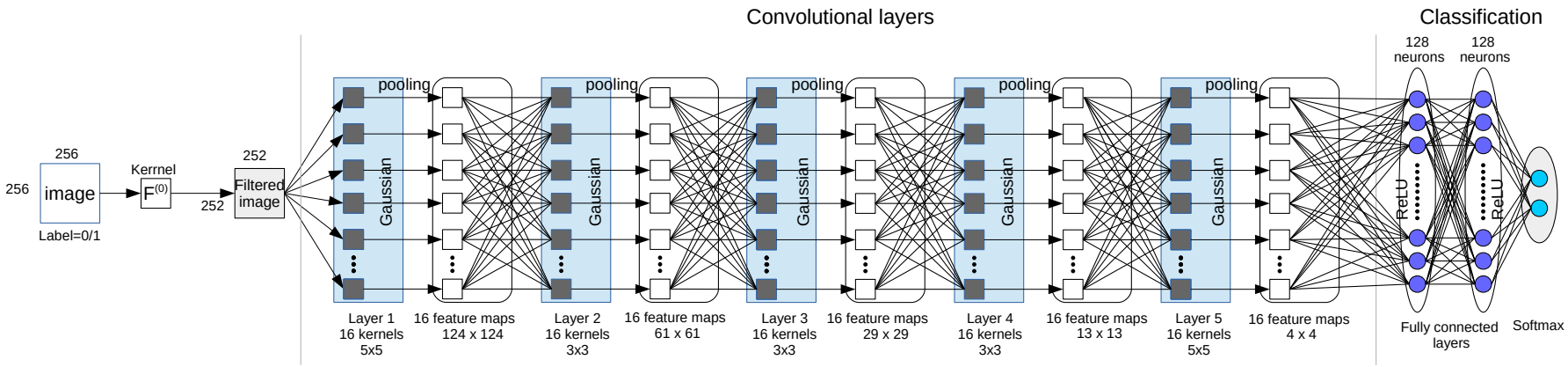


Figure: Qian *et al.* Convolutional Neural Network.

Possible fonctions d'activation:

- Fonction absolue : $f(x) = |x|$,
- Fonction sinus : $f(x) = sinus(x)$,
- Fonction Gaussienne (réseau de Qian *et al.*) : $f(x) = \frac{e^{-x^2}}{\sigma^2}$,
- ReLU (pour Rectified Linear Units) : $f(x) = max(0, x)$,
- Tangent hyperbolique : $f(x) = tanh(x)$ ...

Convolution Neural Network: Pooling

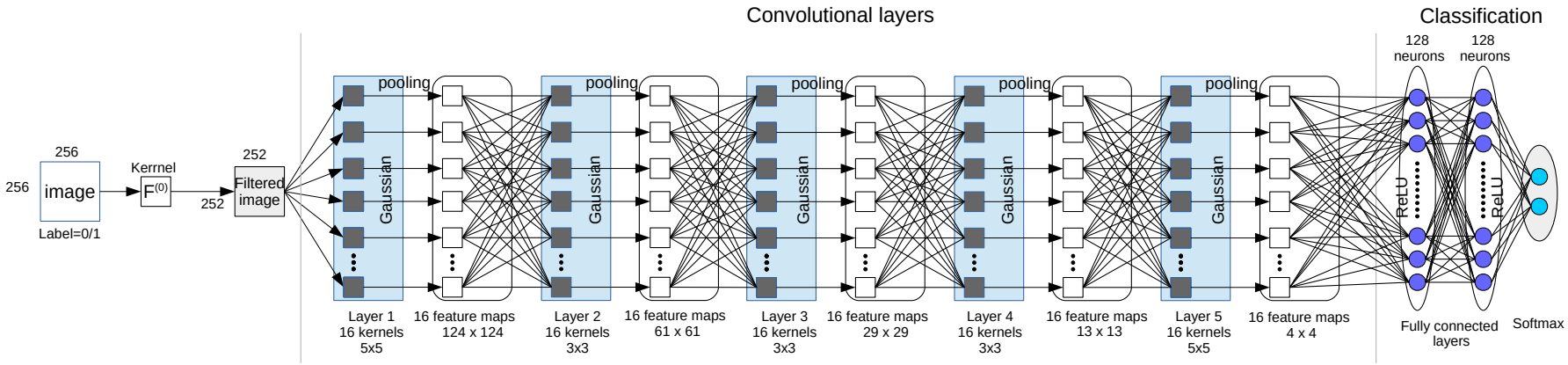


Figure: Qian *et al.* Convolutional Neural Network.

Le pooling est une opération locale calculée sur un voisinage :

- local average (préserve le signal),
- ou local maximum (propriété d'invariance en translation).

+ une opération de sous-échantillonnage.

Convolution Neural Network: Normalisation

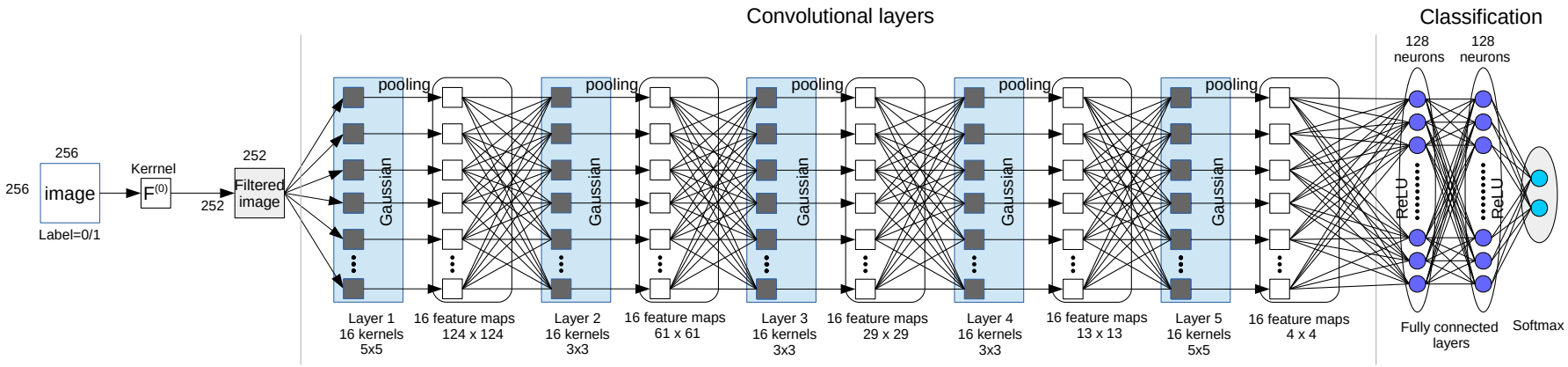


Figure: Qian *et al.* Convolutional Neural Network.

Exemple : Cas ou la normalisation est faite entre "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

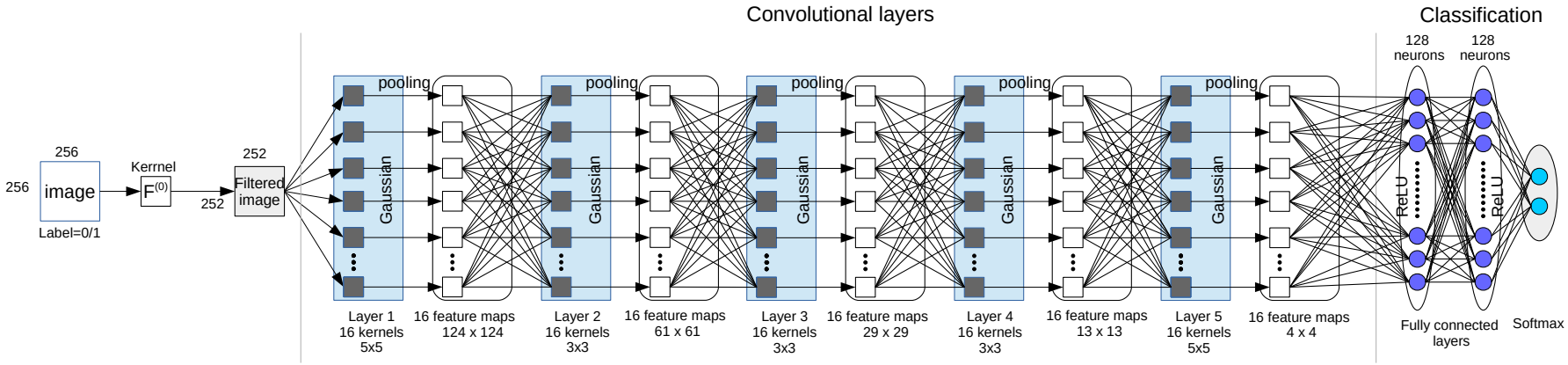


Figure: Qian *et al.* Convolutional Neural Network.

- Trois couches,
- Une fonction softmax normalise les valeurs entre $[0, 1]$,
- Le réseau délivre une valeur pour cover (resp. pour stego).

Autres réseaux "références"

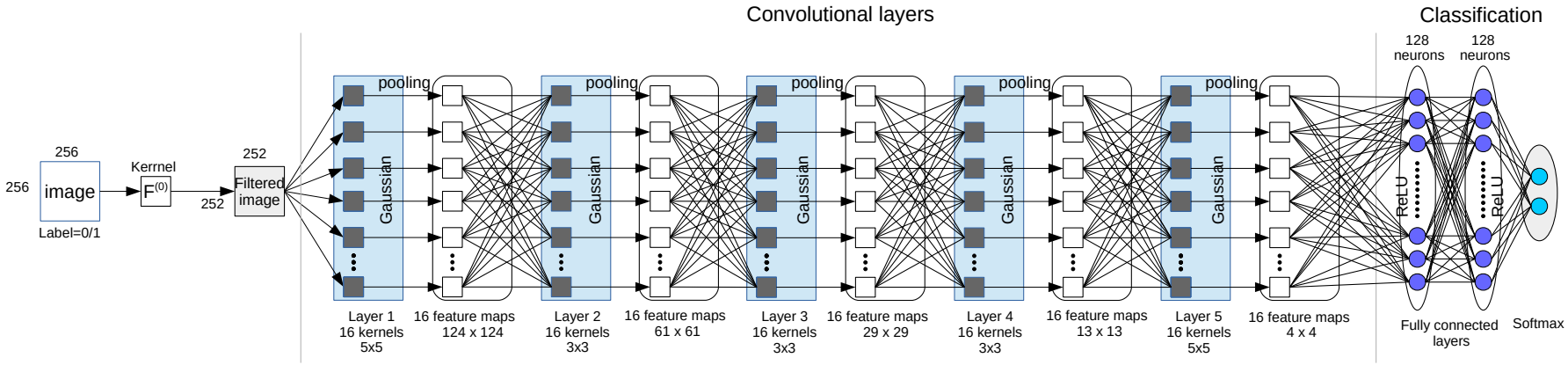


Figure: Qian *et al.* Convolutional Neural Network.

- Xu-Net (mai 2016):
  - Valeur absolue (première couche),
  - Fonction d'activation: TanH et ReLU,
  - Fonction de normalisation: Batch Normalization (2015),
  - Ordre bien spécifique.
- Ye-Net (nov. 2017):
  - Banc de filtres,
  - Fonction d'activation (truncature = "hard tanh"),
  - 8 "couches" et que des convolutions,
  - Version avec utilisation d'une carte de détectabilité.

Plan

Un nouveau réseau

Yedroudj-Net

Agrégation des briques "les plus efficaces" des CNNs conçus récemment.
Objectif: Avoir un CNN basique (baseline) à l'état de l'art.

Les éléments essentiels du réseau :

- Un banc de filtres passe-haut pour le pré-processing (SRM [1]).
- Une fonction d'activation de troncature ("hard tanh") [2].
- La "batch normalization" associée à une couche de "scaling" [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

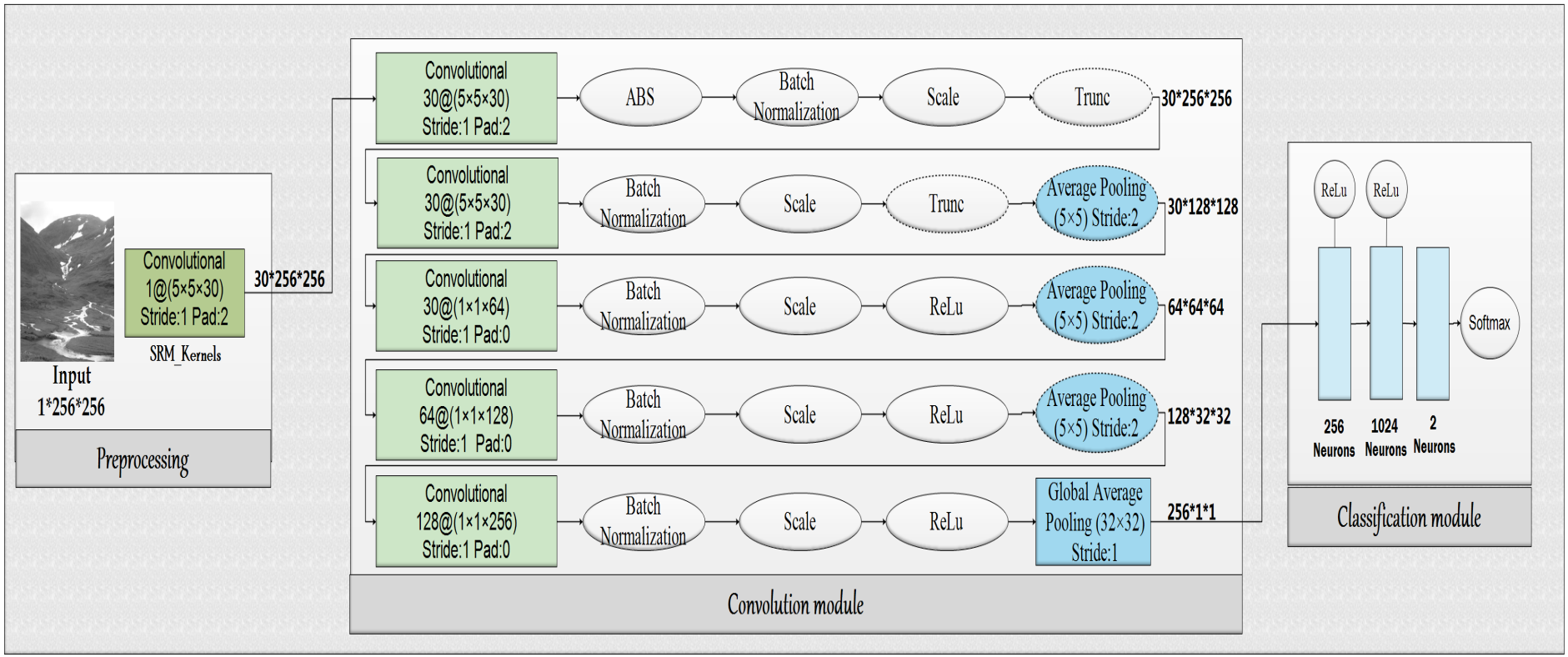
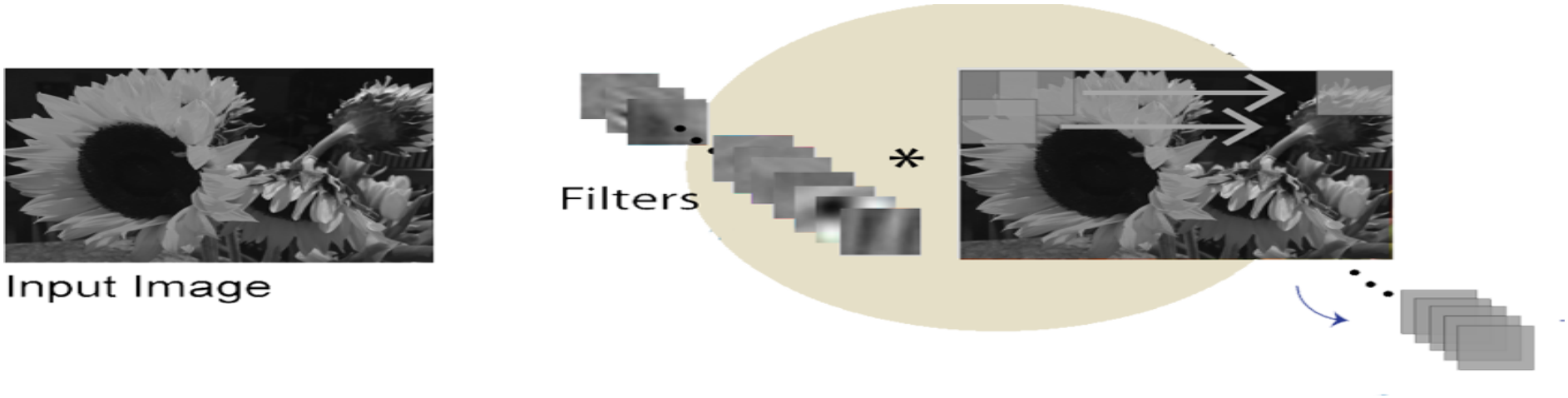


Figure: Yedroudj-Net

Filtres

Filtres passe-haut

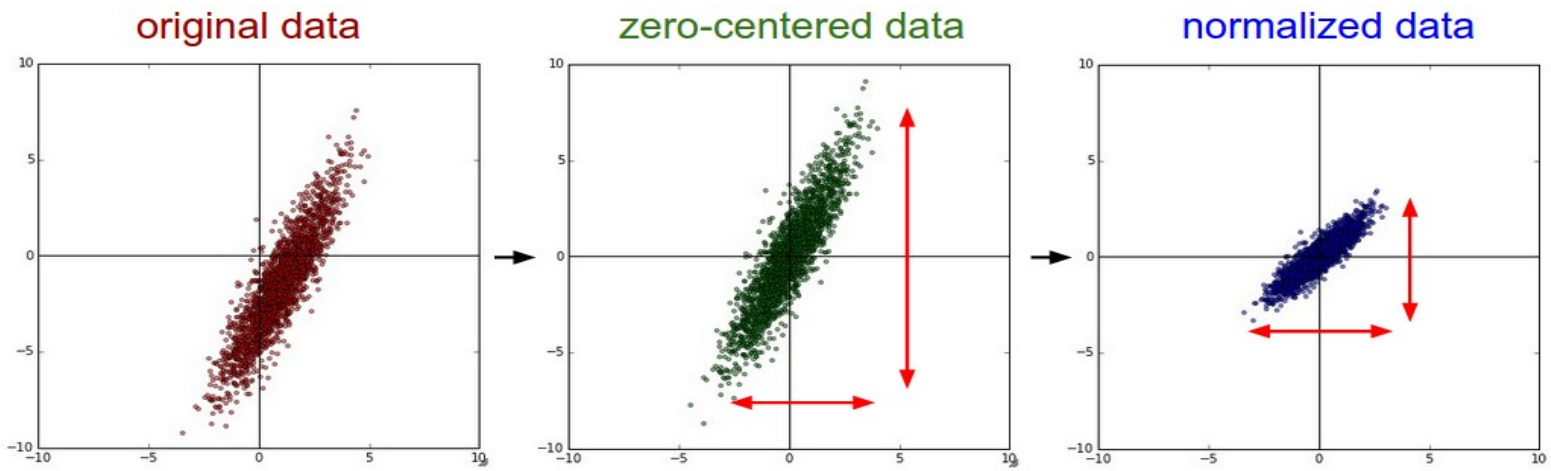
- Dans SRM (= features) il y a pré-traitement des images avec un banc de filtres passe-haut pour extraire le bruit stego [1].
- Dans Yedroudj-Net, il y a pré-traitement des images avec un banc de 30 filtres passe-haut (pre-processing block) [2].



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

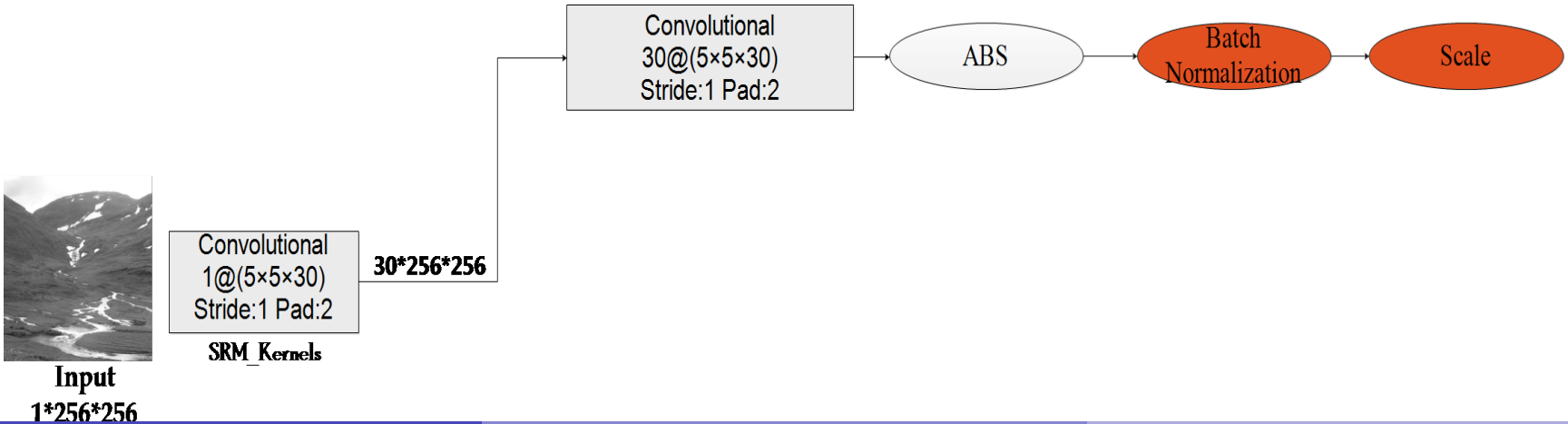
[2]: "Deep Learning Hierarchical Representations for Image Steganalysis", J. Ye, J. Ni, and Y. Yi, TIFS'2017.

Batch Normalization & Scale



Batch Normalization

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

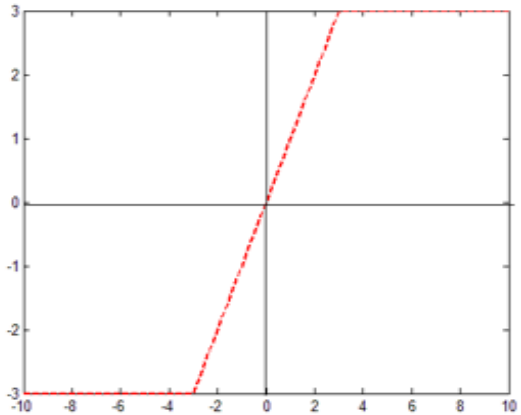


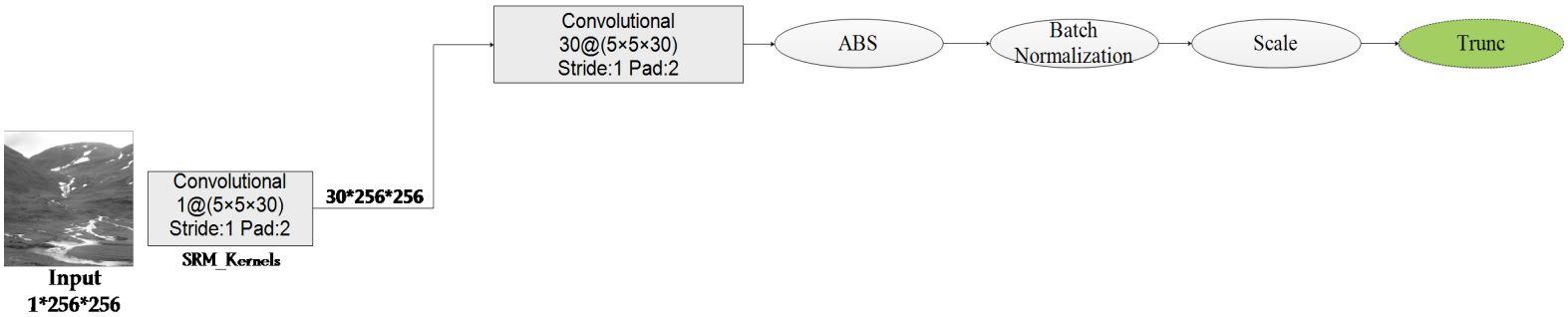
Fonction d'activation : Truncation (hard tanh)

Yedroudj-Net:

Fonction d'activation 'tuncation' pour les 2 premiers blocs.
Permet de limiter l'intervalle de valeur et prévient la modélisation par le réseau de grandes valeurs.

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





Protocole expérimental

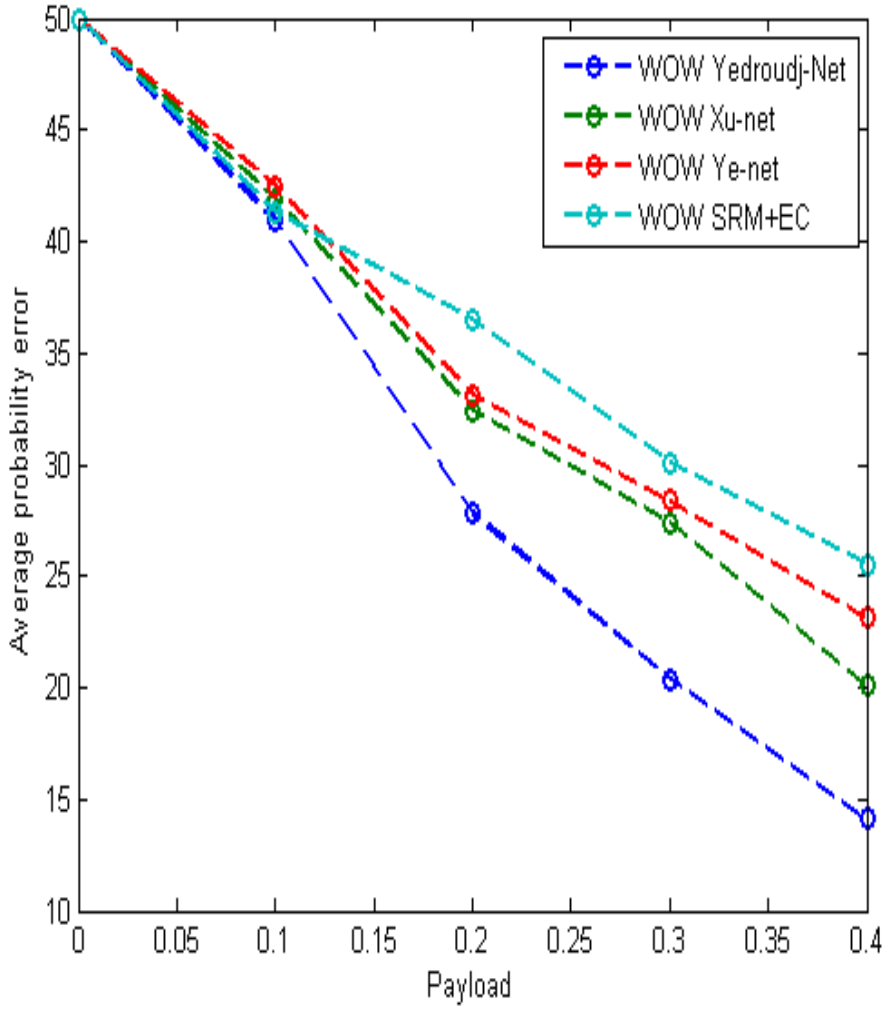
Protocole clairvoyant

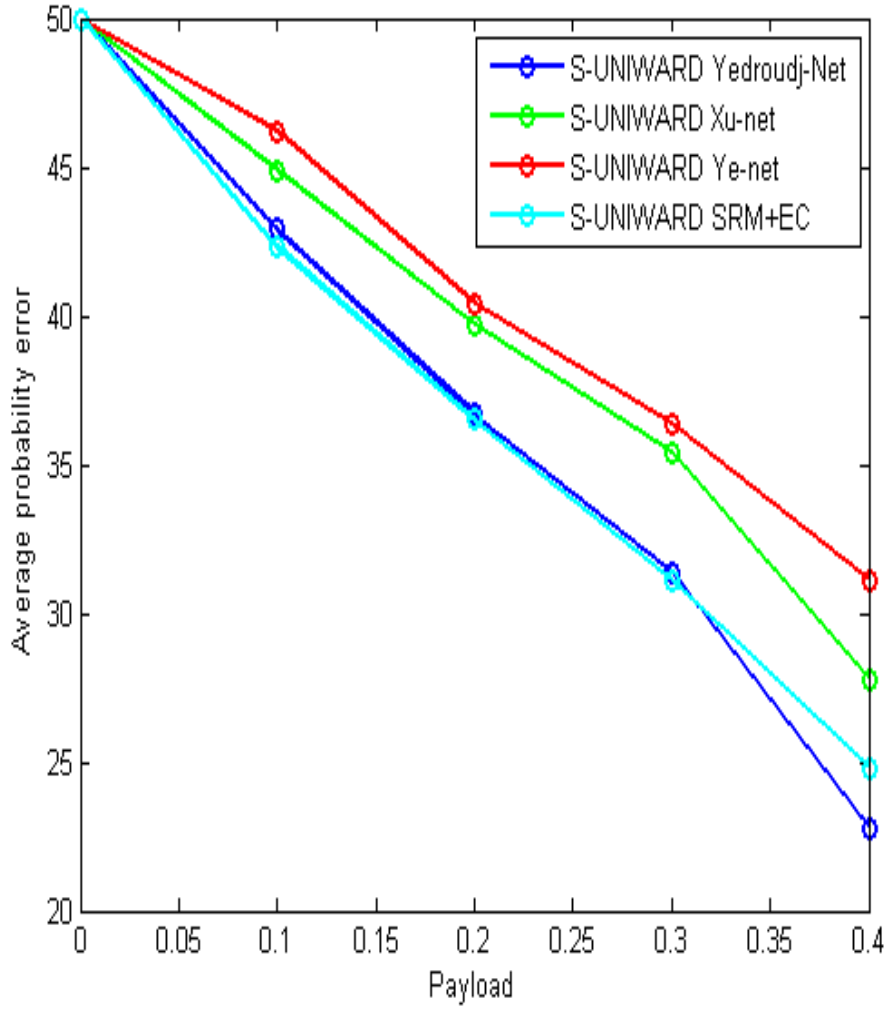
- Retaille des 10 000 images de BOSSBase de 512×512 à 256×256,
- Utilisation des algorithmes d'insertion WOW [1] et S-UNIWARD [2] pour générer les stegos (Matlab Version),
- Selection de 5 000 paires pour l'apprentissage dont 1 000 paires pour la validation,
- Les 5 000 autres paires sont utilisées pour le test (évaluation).

[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.

Résultats et discussion





Plan

Améliorations des performances :

- Virtual Augmentation [Krizhevsky 2012]
- Transfer Learning [Qian et al. 2016],
- Utilisation d'Ensemble [Xu et al. 2016],
- Apprendre sur des millions d'images ? [Zeng et al. 2018],
- Ajouter des images des même appareils photos et avec le même "développement" [Ye et al. 2017], [Yedroudj et al. EI'2018],
- De nouveaux réseaux [Yedroudj et al. 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.

Enrichissement de la base d'apprentissage (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.

Protocole clairvoyant

- Retaille des 10 000 images de BOSSBase de 512×512 à 256×256,
- Utilisation des algorithmes d'insertion WOW [1] et S-UNIWARD [2] pour générer les stegos (Matlab Version),
- Selection de 5 000 paires pour l'apprentissage dont 1 000 paires pour la validation,
- Les 5 000 autres paires sont utilisées pour le test (évaluation).

[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.

Enrichissement de la base d'apprentissage (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: Probabilité d'erreur pour Yedroudj-Net avec différents enrichissements

- BOSS+VA : 32 000 paires, BOSS+all-DEV : 44 0000 paires, BOSS+BOWS2 : 14 000 paires, BOSS+BOWS2+VA : 112 000,
- Expe versus EC+RM, versus Xu-Net, versus Ye-Net,
- Expe enrichissements contre-productif (cameras différentes, changement ratio, ...),

Une conjecture (règle pour l'augmentation) :

"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.

Plan

Quelques mots sur 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 "simulant" une insertion dans une image spatial,
- CNN appelé Générateur (noté G) génère la carte de modifications (-1/0/+1),
- G apprend à insérer grâce à une "compétition" (méthodologie GAN) entre lui et un Discriminateur (noté 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.

Quelques mots sur 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

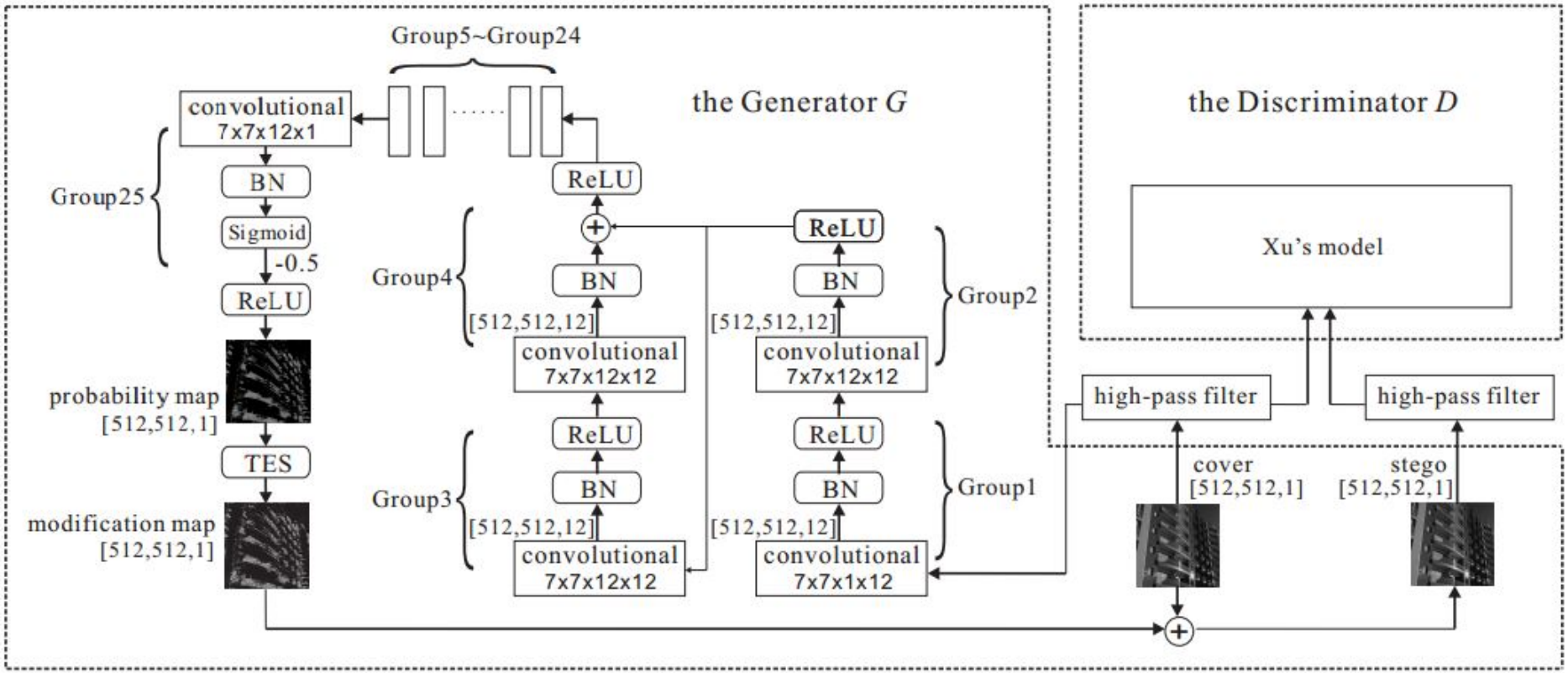


Figure: ASDL-GAN; Figure extraite du papier [Tang et al. 2017]

Plan

Conclusion

On a vu :

- CNN steganalyse spatiale (Xu-Net, Ye-Net, Yedroudj-Net),
- CNN steganalyse JPEG (JPEG Xu-Net basé ResNet),
- L'enrichissement de base (une des techniques d'augmentation de performances),
- La Stéganographie GAN.

2018, L'installation et l'études d'autres scénarios :

- Enrichissement,
- Quantitatif,
- Taille variable d'images,
- Cover-Source mismatch,
- GAN.

End of talk

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