

Stéganalyse par deep learning

Présentation d'équipe

Doctorant: Mehdi Yedroudj

Encadrants: Marc Chaumont
Frederic Comby

13/07/2018

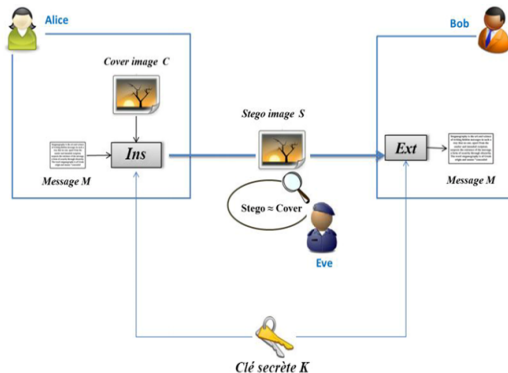
- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Introduction

Stéganographie Vs. Stéganalyse

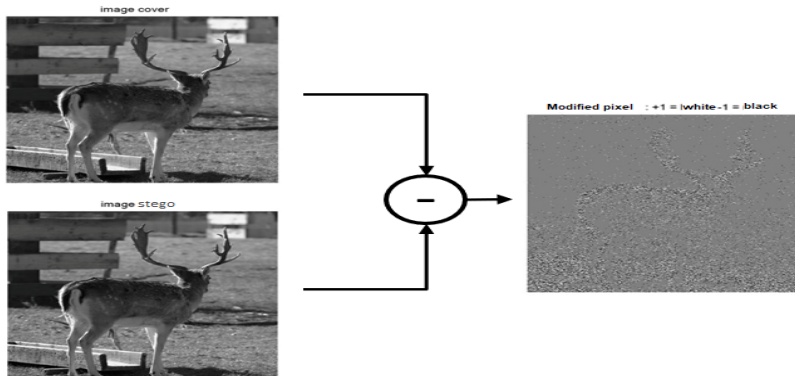
Stéganographie: est l'art de dissimuler des messages secrets. Cela consiste à cacher un message au sein d'un support numérique.

Stéganalyse: a pour objectif de détecter la présence de données dissimulées à l'aide d'un algorithme stéganographique.



Exemple de stéganographie

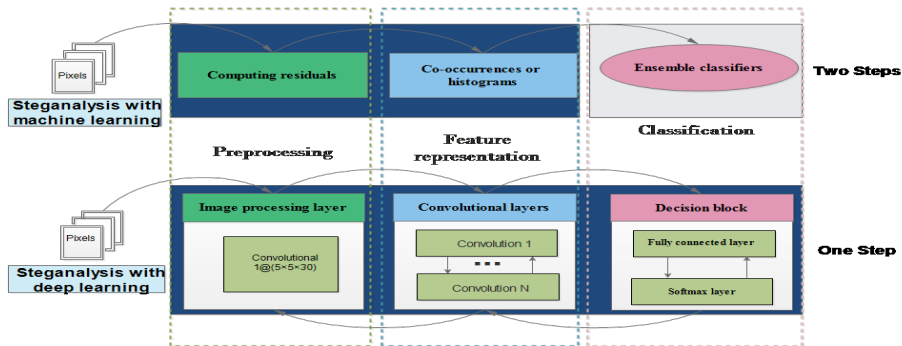
- Les images, cover/stego devraient être similaires (stego $\simeq$ cover).
- Seul le récepteur peut extraire le message grâce à la clé secrète partagée (émetteur - récepteur).



Exemple d'insertion avec S-UNIWARD à 0.4 bpp

Les approches de stéganalyse

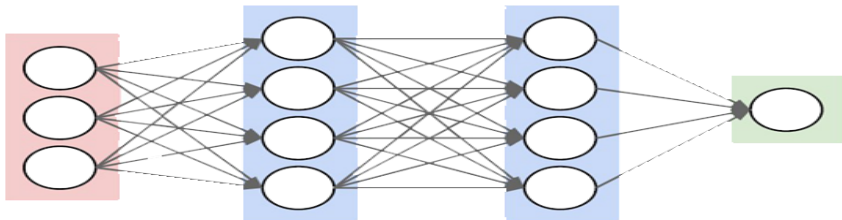
- Approches basées apprentissage automatique [1,2] vs Approches basées apprentissage profond[3].



- [1]: J. Kodovsk'y, J. Fridrich, and V. Holub, "Ensemble Classifiers for Steganalysis of Digital Media", 2012
[2]: J. Fridrich and J. Kodovsk'y, "Rich Models for Steganalysis of Digital Images," 2012
[3]: G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis", 2016

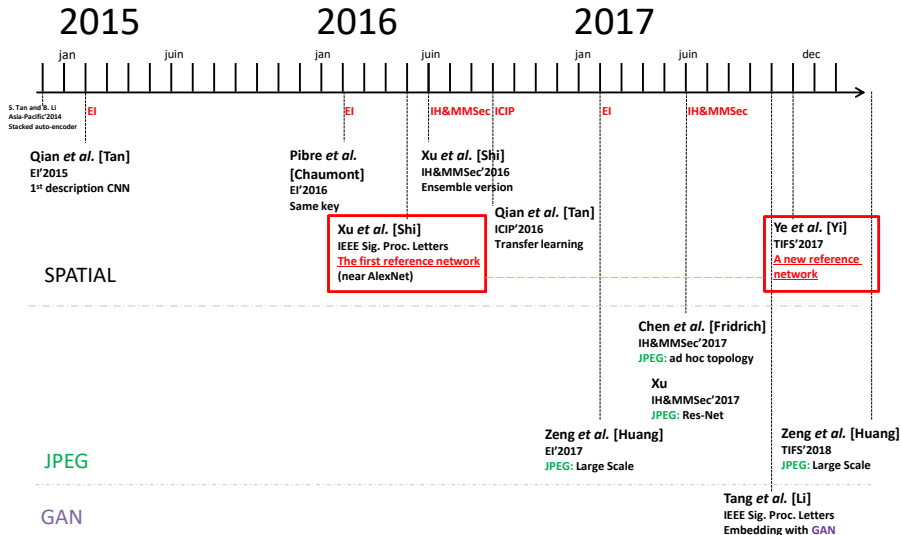
Deep Learning(L'apprentissage profond)

- L'apprentissage profond: Ensemble de méthodes d'apprentissage automatique qui utilise des réseaux de neurones artificiels..



- Ces représentations sont composées de trois parties que l'on appelle des couches qui sont elles même composées de neurones.

CNN évolution pour la stéganalyse



Marc Chaumont, " The emergence of Deep Learning in steganography and steganalysis (in the period 2015 - 2017) ", Tutorial given at a research day in France, the 16th of January 2018, Poitiers, France, 2018 DOI 10.13140/RG.2.2.35080.32005.

http://www.lirmm.fr/~chaumont/publications/Emergence_of_Deep_Learning-Tutorial_Meeting-France-CHAUMONT_16_01_2018.pdf

Yedroudj-Net

Nouveau CNN qui rassemble les parties les plus efficaces des CNN récents.
Objectif: Améliorer les performances par rapport à l'état.

Les parties importantes de cette architecture proposée sont:

- Une banque de filtres passe-haut pour le pré-traitement "SRM [1]".
- Une fonction d'activation troncature (HardTanh) [4].
- Une normalisation par lots couplée avec une couche de mise à échelle [5][9].

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

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

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

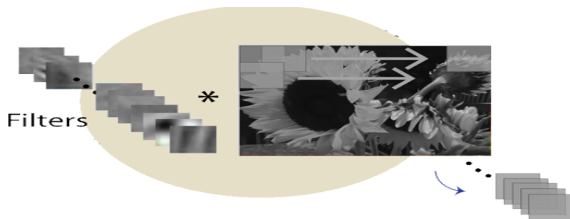
- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

High-pass filters

- Méthodes de stéganalyse classiques: Prétraiter les images avec un seul filtre passe-haut pour extraire les résidus de composante de bruit [1].
- Dans notre méthode: Utiliser une banque de filtres passe-haut composée de 30 noyaux. Dans notre CNN, nous l'appelons le bloc de pré-traitement.



Input Image



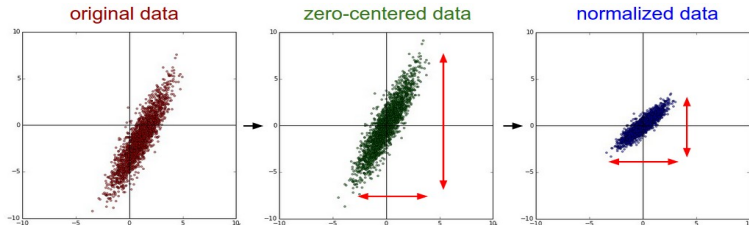
[1]:J. Kodovsk'y, J. Fridrich, and V. Holub, "Ensemble Classifiers for Steganalysis of Digital Media"

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

Plan

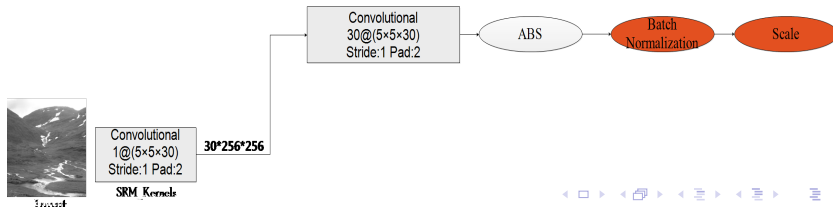
- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Batch Normalization & Scale



Batch Normalization

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



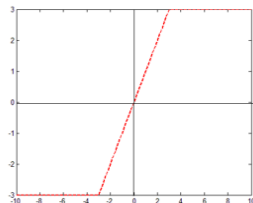
- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Activation function: Truncation

Yedroudj-Net:

On adopte la fonction d'activation non-linéaire Troncature (HardTanh) sur la première et la seconde couche. Cette fonction est utilisée pour limiter la plage de valeurs de données et empêcher les couches plus profondes de modéliser de grandes valeurs [4].

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



Convolutional
1@(5×5×30)
Stride:1 Pad:2

30*256*256

Convolutional
30@(5×5×30)
Stride:1 Pad:2

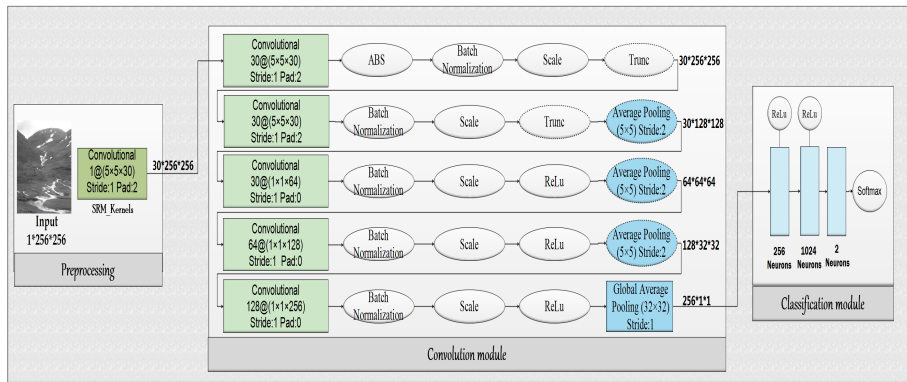
ABS

Batch
Normalization

Scale

Trunc

Architecture de notre réseau



L'architecture de notre CNN proposée "Yedroudj-Net"

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Clairvoyant protocole

On garde le scénario classique:

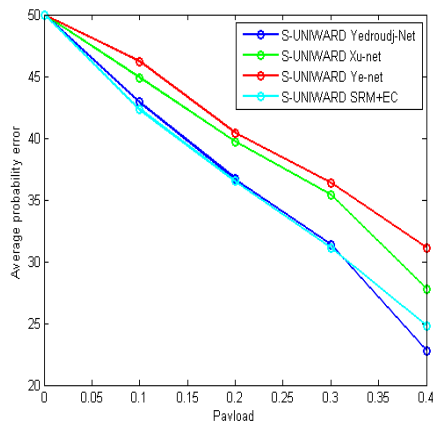
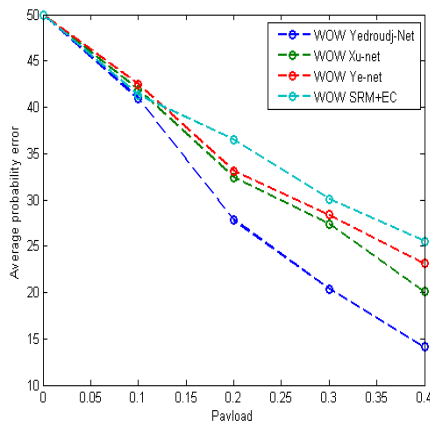
- Redimensionner les 10000 images de BOSSBase de $512 * 512$ à $256 * 256$,
- Utiliser les deux algorithmes WOW [5] et S-UNIWARD [6] pour générer des stegos (**Matlab Version**),
- Sélectionner 5000 paires pour l'apprentissage dont 1000 paires pour la validation,
- Utiliser les 5000 paires qui restent comme un ensemble de test pour évaluer le modèle obtenu.

[5] V. Holub and J. Fridrich, "Designing Steganographic Distortion Using Directional Filters," 2012.

[6] V. Holub, J. Fridrich, and T. Denemark, "Universal Distortion Function for Steganography in an Arbitrary Domain," 2014.

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Discussion sur les résultats



Probabilité d'erreur de Yedroudj-Net, Xu-Net, Ye-Net et SRM + EC contre deux algorithmes d'insertion WOW et S-UNIWARD à 0,2 bpp et 0,4 bpp.

Influence de l'augmentation de base: (Probabilité d'erreur)

BOSS+BOWS2: 14 000 pairs.

BOSS+VA: 32 000 pairs.

	WOW 0.2 bpp	S-UNI 0.2 bpp
BOSS	27.8 %	36.7 %
BOSS+VA	24.2 %	34.8 %
BOSS+BOWS2	23.7 %	34.4%

Probabilité d'erreur de Yedroudj-Net, sur WOW/S-UNIWARD à 0,2 bpp avec et sans augmentation des données.

- On a proposé une approche efficace basée sur l'apprentissage profond (CNN) pour la stéganalyse qui surpasse l'état de l'art.
- Notre réseau permet d'obtenir de meilleurs résultats par rapport aux autres modèles de CNN avec et sans mesures supplémentaires (augmentation de la base d'entraînement).
- Notre CNN a moins de variables à apprendre et donc moins de temps pour converger

Pour une augmentation efficace pouvons-nous améliorer encore plus la précision de la détection?

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Quelques observations:

- Les CNN actuels (stéganalyse spatiale ou JPEG) donnent des résultats similaires ou même meilleurs que les EC + RM,
- Les CNN ont besoin de beaucoup d'échantillons lorsqu'ils sont utilisés à des fins de stéganalyse (au moins 5 000 paires, peuvent atteindre des millions),
- Chaque fois que la distribution cible est modifiée, le processus d'apprentissage doit être refait.

Problème:

En "monde réel scénario" l'ensemble d'apprentissage peut être très petit (0, 10, 100, ..).

Question: Comment faire face à un apprentissage de bas régime?

Solutions actuelles pour améliorer l'efficacité de CNN pour la stéganalyse

- Virtual Augmentation [Krizhevsky 2012]
- Transfer Learning [Qian et al. 2016],
- Ensemble of CNN [Xu et al. 2016],
- Apprendre sur des millions d'images? [Zeng et al. 2018],
- Utiliser de nouveaux réseaux (Res-Net, dense-net..) [Yedroudj et al. 2018], ...
- ...
- Augmenter la base de données ...

"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,
"Yedroudj-Net: An Efficient CNN for Spatial Steganalysis", M. Yedroudj, F. Comby, M. Chaumont, Under submission, 2018.

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

"Data augmentation" protocole

En partant du protocole classique:

- Ajouter des paires (cover/stego) des autres bases (BOWS2,LIRMM2,BOSS2.0) à l'ensemble d'entraînement (protocole Clairvoyant).
- Utiliser les 5000 paires du test pour évaluer le modèle obtenu.

PS: Différentes méthodes ont été testés pour augmenter la base la rotation, retournement, mise à l'échelle et même des différentes interpolations pour le redimensionnement.

Analyse des résultats (Enrichissement contre-productif)

Configuration: Enrichissement avec d'autres caméras (Probabilité d'erreur):

Training set BOSS+LIRMM = 14 000 pairs.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS	27.8 %	36.7 %
BOSS+LIRMM	29.9 %	38.6 %

Configuration: Enrichissement avec des sources fortement dissemblables avec des proportions déséquilibre (Probabilité d'erreur):

BOSS + PLACES2 1% = 14 000 pairs; BOSS + PLACES2 10% = 104 000 pairs; BOSS + PLACES2 100% = 1 004 000 pairs.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS	27.8 %	36.7 %
BOSS+PLACES2 1%	34.2 %	41.6 %
BOSS+PLACES2 10%	40.0 %	43.9 %
BOSS+PLACES2 100%	44.6 %	45.3 %

Analyse des résultats (Enrichissement productif)

Configuration: Enrichissement avec des caméras "similaires"
(Probabilité d'erreur): BOSS+BOWS2: 14 000 paires.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS	27.8 %	36.7 %
BOSS+BOWS2	23.7 %	34.4%

Configuration: Les mêmes images RAW mais des développements différents
(Probabilité d'erreur): BOSS+all-DEV: 44 000 paires.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS+all-DEV	23.0 %	33.2 %

Analyse des résultats (Enrichissement productif)

Configuration: Les mêmes images RAW mais des développements différents (Probabilité d'erreur):

	WOW 0.2 bpp	S-UNI 0.2 bpp
BOSS	27.8 %	36.7 %
BOSS+DEV:Res-Bicub	25.7 %	37.5 %
BOSS+DEV:Res-Spline	26 %	35.8 %
BOSS+DEV:Res-NoInt	25.6 %	36.2 %
BOSS+DEV:Crop	34.8 %	44.2 %
BOSS+DEV:Res-Crop	28.1 %	37.9 %
BOSS+BOSS-ALP	26.0 %	35.5%

Augmentation virtuelle ...: BOSS+BOWS2+VA: 112 000 pairs.

	WOW 0.2 bpp	S-UNIWARD 0.2 bpp
BOSS+VA	24.2 %	34.8 %
BOSS+BOWS2+VA	20.8 %	31.1 %

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

Pour une augmentation efficace de la base de données deux options sont proposées:

- Produire de nouvelles images en utilisant les mêmes caméras et le même développement que la base d'origine,
- Avoir un accès aux images RAW originales et les utiliser pour produire de nouvelles images avec des développements similaires,
- **X** Augmentation automatique de la base en utilisant un GAN (Cela ne marche pas encore).

Travail futur:

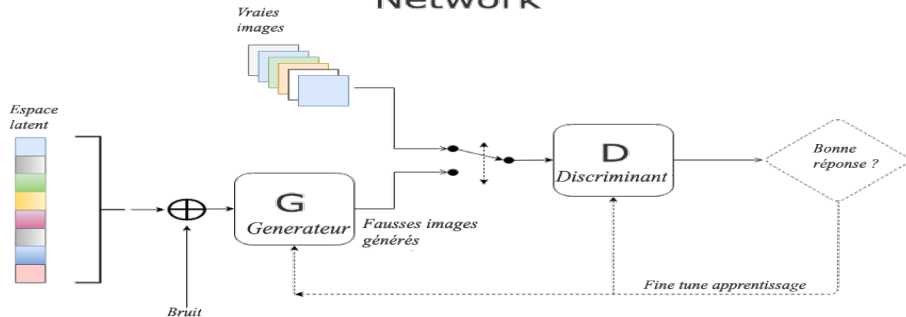
- Essayer un ensemble de Yedroudj-Nets,
- Utiliser un GAN pour faire de la stéganographie.

Plan

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

C'est quoi les GANs ?

Generative Adversarial Network



L'architecture générale du GAN

[8] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

- 1 Introduction / État de l'art
- 2 Méthode proposée
 - Filters
 - Batch Normalization & Scale
 - Activation function: Truncation
- 3 Résultats expérimentaux
 - Protocole et configuration
 - Discussion sur les résultats
- 4 Data augmentation
 - Stéganalyse dans le monde réel
 - Résultats expérimentaux
 - Conclusion
- 5 GAN (generative adversarial networks)
 - Generative Adversarial Networks
 - GANs pour la stéganalyse

DAGAN pour la stéganalyse ?



Illustration de la première étape : l'apprentissage et la validation du modèle

DAGAN pour la stéganalyse

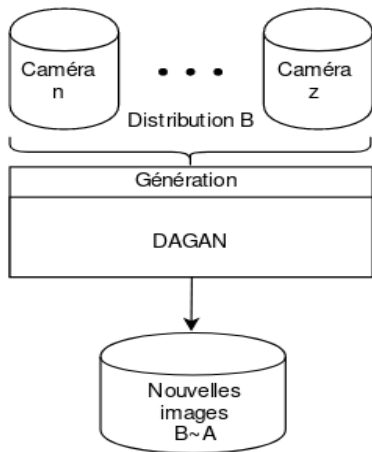


Illustration de la deuxième étape : la génération de nouvelles images

Thank you! Q&R