## 3 The linear explanation of adversarial examples

Dans un premier temps, l’article l’existence d’exemples adversariaux pour des modèles linéaires. Il note la perturbation maximale n’étant pas détecté par l’algorithme. Il définit alors une entrée adversariale comme étant avec .

Il cherche à maximiser l’impact de la perturbation tout en minimisant la perturbation pour qu’elle ne soit pas visible par l’algorithme. C’est pour cette raison qu’est définit la perturbation comme étant égale à . Ainsi, dans les espaces de grandes dimensions, la perturbation aura d’autant plus d’impact.

## 4 linear perturbation of non linear models

L’article explique comment créer un exemple adversarial. Il utilise la méthode du signe du gradient rapide (Fast Gradient Sign Method). L’article donne l’expression de la perturbation.

Une image contenant Police, typographie, écriture manuscrite, calligraphie

Description générée automatiquement

L’article constate que les réseaux linéaires ne résistent pas aux perturbations linéaires.

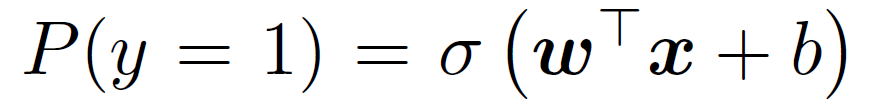
Pour générer des exemples adversariaux, il est possible d’appliquer une rotation à l’image.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithme |  | Taux d’erreur | Confiance | Base de données |
| Shallow softmax classifier | 0.25 | 99,9% | 79,3% | MNIST |
| Maxout network classifier | 0,25 | 89,4% | 97,6% | MNIST |
| Convolutionnal maxout network | 0.1 | 87,15% | 96,6% | CIFAR-10 |

L’article démontre ainsi que les exemples adversariaux découlent de la linéarité des réseaux puisqu’il est très facile d’obtenir des résultats concluants. Il est plus facile de créer des exemples adversariaux pour des réseaux linéaires.

## 5 adversarial training of linear model versus weight decay

L’article utilise la regression logistique pour réaliser un entrainement adversarial simple comme il s’agit du modèle de classification le plus simple qu’il puisse exister. On définit la régression logistique comme :



Dans le cas d’un entrainement adversarial, on définit l’algorithme de descente du gradient comme :



On obtient avec la regression logistique un taux d’erreur de 99% alors qu’on avait avant 1,6%.

La régularisation (weight decays) tends à surestimer l’impact des exemples adversariaux et fonctionne moins bien que la regression logistique.

## 6 Adversarial training od deep network

L’article considère que les réseaux avec des couches cachées à condition d’être codés correctement peuvent être plus résistants aux exemples adversariaux.

A completer encore si besoin

## 7 Different kinds od model capacity

Les réseaux même avec de faibles capacités peuvent fournir de bons résultats avec confiance lorsqu’ils sont soumis aux exemples adversariaux. C’est le cas des réseaux RBF (Radial Basis Function) puisqu’ils ne sont pas confiants lorsqu’ils sont dupés. Ce réseaux réduit naturellement sa confiance sur des exemples qu’il ne comprend pas. Toutefois, ces réseaux ne répondent correctement que dans une certaine direction.

## 8 Why do adversarial examples generalize ?

L’article remarque que des exemples adversariaux générés par un modèle et souvent mal classé par un autre modèle n’ayant aucun lien avec le premier. De plus, la classe attribuée est souvent la même. Un algorithme classe mal un exemple car la perturbation ne se propage que dans une seule direction dépendamment de la valeur de . De plus, la même mauvaise classe est attribué car les algorithmes ont une bonne capacité de généralisation sur les données d’entraînement.

## 9 Alternative hypothèses

L’article réfute deux hypothèses quant à l’existence d’exemples adversariaux.

L’article affirme que les exemples adversariaux ne peuvent pas exister à cause de l’entrainement génératif qui produit des contraintes sur le processus d’entrainement.

La deuxième hypothèses réfutée est que les exemples adversariaux existent puisque des modèles simples ont des caractéristiques étranges mais que la moyenne de plusieurs modèles peuvent les faire disparaitre