Inatel
Instituto Nacional de Telecomunicações

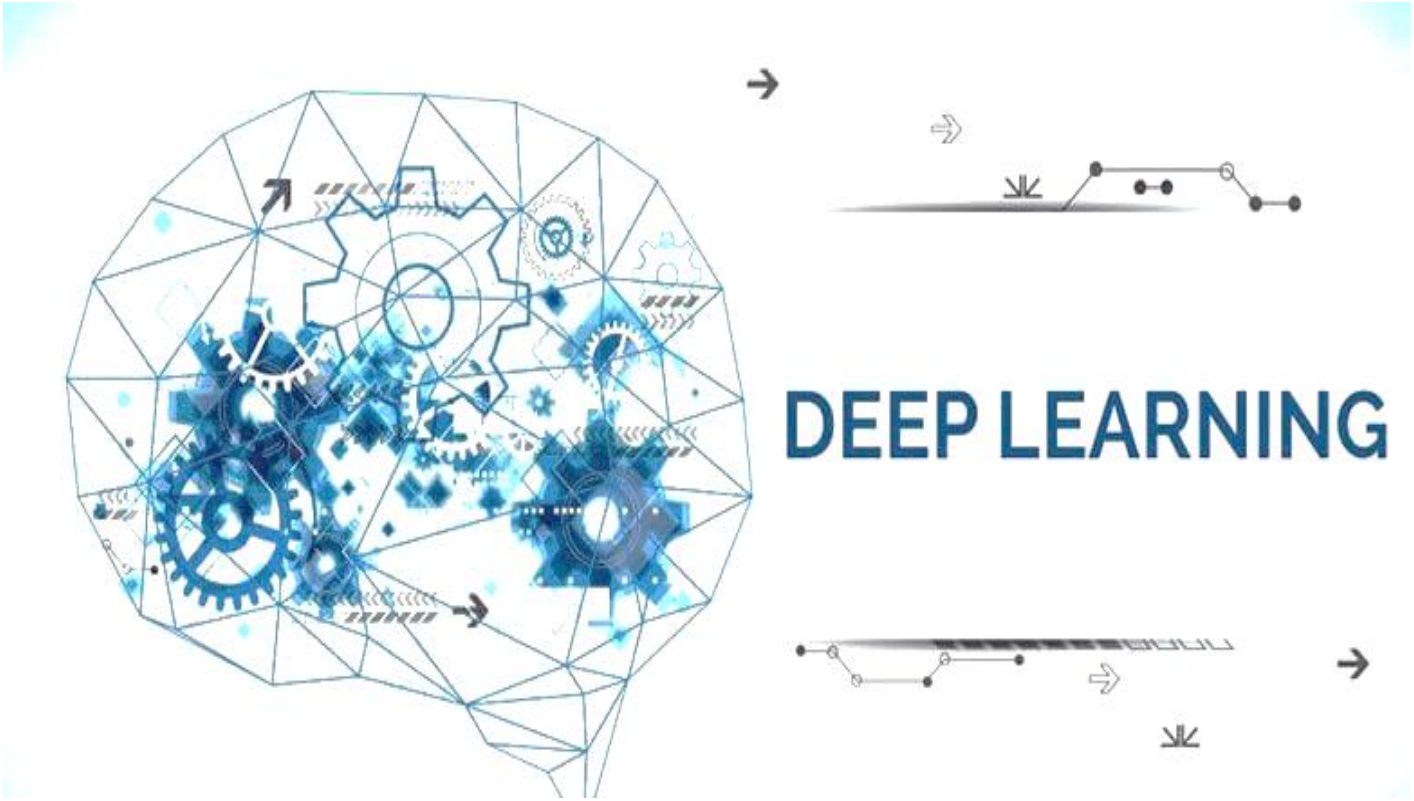
Resumo do Artigo : Adversarial Attacks on Deep-Learning Based Radio Signal Classification

- *Professor : Felipe Augusto Pereira de Figueiredo*
- *Aluno e Matricula :*
- *Jones Marcio Nambundo -710*
- *Mayomona Lando Filipe - 836*

Sumário:

I. Introdução

O aprendizado profundo (DL), implementado por meio de redes neurais profundas (DNNs).



II. Objectivo Geral e Proposta implementada

- Mostrar que essa classe de algoritmos é extremamente vulnerável a ataques adversos.

Objectivos Específicos

- Apresentar um novo algoritmo para geração de ataques adversários específicos de entrada de caixa branca com granulação fina.

- Propor um algoritmo computacionalmente eficiente para criar perturbações adversas universais (UAP) de caixa branca.

- Mostra como se pode criar ataques UAP de caixa preta.

- Revelamos a mudança invariável propriedade de UAPs.

II. Ataques a Adversários

- Classe de métodos de gradiente rápido (FGM) .

Métodos computacionalmente eficientes para elaborar exemplos contraditórios, à custa de perturbações de granulação grossa.

-Existem duas variantes de FGM, FGM segmentada e FGM não segmentada.

- Em um ataque FGM segmentado, o classificador faz uma classificação incorreta da modulação.

- Em um ataque não-Segmentado de FGM há perda do seu rotulo.

Os ataques adversários podem ser divididos em ataques de caixa branca e caixa preta, com base na quantidade de conhecimento que o adversário tem sobre o modelo.

III. THE GNU RADIO ML DATASET AND ITS DNN

- Usou-se um conjunto de dados GNU radio ML RML2016.10a e seu DNN associado.

- Estão disponíveis publicamente em www.deepsig.io/datasets.

- GNU radio ML RML2016.10a contém 220000 amostras de entrada.

- Ele contém 11 modulações diferentes. Gera 20 níveis de amostras diferentes.

- Metade das amostras é considerada como o conjunto de treinamento e a outra metade como conjunto de testes.

-Usa um classificador CNN profundo chamado VT-CNN2 seguindo o formato padrão do TensorFlow para dados, ou seja, (altura, largura, canais).

IV. Ataques Adversarias para classificação da Modulação baseada em Deep Leanirng.

a) Caixa Branca.

- Nesta seção, desenvolveu -se um ataque adversário de caixa branca à classificação de modulação baseada em DL, usando VT-CNN2 como classificador.

- O atacante está ausente.

- O atacante está presente.

- O alvo do invasor é projetar rx de modo que cause classificação incorreta para o DNN subjacente no lado RX.

- Apresentamos Alg. 1 para resolver esses problemas a seguir.

Algorithm 1 Crafting an adversarial example

Inputs:
- input $\mathbf{x}$ and its label l_{true}
- the model $f(., \boldsymbol{\theta})$
- desired perturbation accuracy ε_{acc}
- maximum allowed perturbation norm p_{max}

Output: adversarial perturbation of the input, i.e., $\mathbf{r}_{\mathbf{x}}$

Initialize: $\boldsymbol{\varepsilon} \leftarrow \mathbf{0}^{C \times 1}$
for *class-index* in *range*(C) **do**
  $\varepsilon_{max} \leftarrow p_{max}$, $\varepsilon_{min} \leftarrow 0$
  $\mathbf{r}_{norm} = \left(\|\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{e}_{\textit{class-index}})\|_2\right)^{-1} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{e}_{\textit{class-index}})$
  while $\varepsilon_{max} - \varepsilon_{min} > \varepsilon_{acc}$ **do**
    $\varepsilon \leftarrow (\varepsilon_{max} + \varepsilon_{min})/2$
    $\mathbf{x}_{adv} \leftarrow \mathbf{x} - \varepsilon\, \mathbf{r}_{norm}$
    if $\hat{l}(\mathbf{x}_{adv}) == l_{true}$ **then**
      $\varepsilon_{min} \leftarrow \varepsilon$
    else
      $\varepsilon_{max} \leftarrow \varepsilon$
    end if
  end while
  $[\boldsymbol{\varepsilon}]_{\textit{class-index}} = \varepsilon_{max}$
end for
target-class $= \arg\min\ \boldsymbol{\varepsilon}$ and $\varepsilon^* = \min\ \boldsymbol{\varepsilon}$

$$\mathbf{r}_{\mathbf{x}} = -\frac{\varepsilon^*}{\|\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{e}_{\textit{target-class}})\|_2} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{e}_{\textit{target-class}})$$

- Aqui, propomos duas novas métricas, **(PNR**) e (PSR)

- Si PNR < 1 a perturbação (quase) imperceptível.

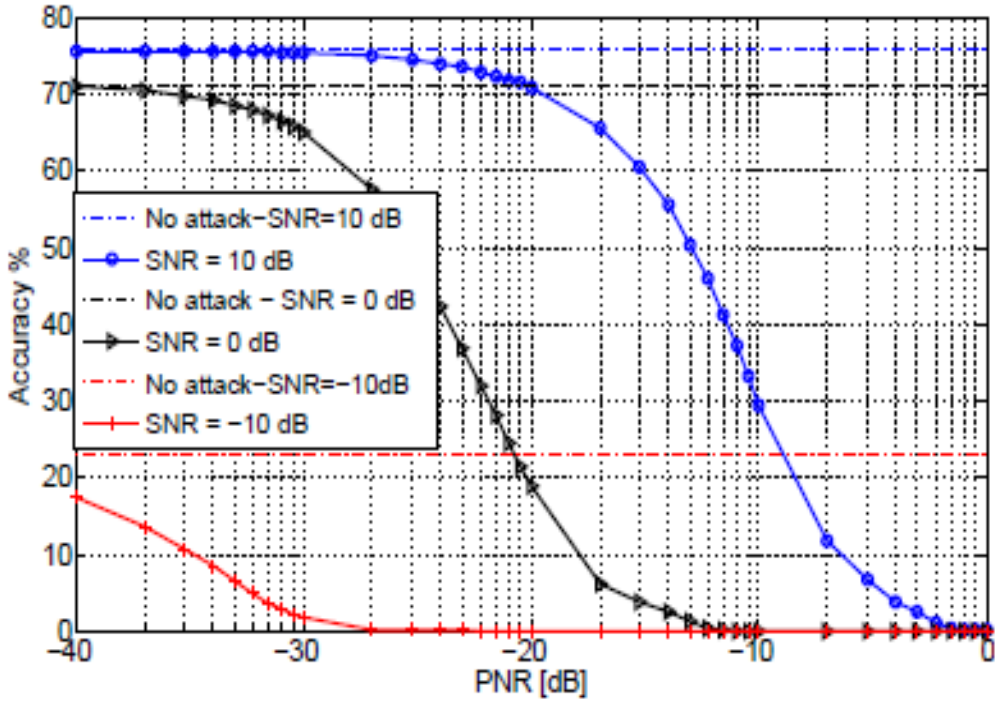


Fig. 2: The accuracy of VT-CNN2 versus PNR, with and without adversarial attack.

IV Ataques Adversarias para classificação da Modulação baseada em Deep Leanirng.
b) Caixa Preta.

-Considerando três suposições limitantes:

-Primeiro, o atacante sabe a entrada exata.

-Segundo cada elemento de x é perturbado por seu elemento correspondente em rx.

- Terceiro, como consideramos um ataque de caixa branca.

c) Perturbações Adversárias Universais

Alg. 1 cria perturbações adversárias dependentes da entrada.

- em vez de rx, estamos interessados em encontrar uma perturbação universal do contraditório (UAP) r.

- Método comum para criar UAP .
O algoritmo nele recebe como entradas, 1) o modelo, 2) a norma desejada da UAP e 3) um subconjunto aleatório de entradas de dados, Com base nessas entradas, gera como saída um UAP r.

- Propor um novo algoritmo para gerar um UAP
- Uma complexidade computacional muito baixa.
- O algoritmo usa a análise de componentes principais (PCA) para criar o UAP

Algorithm 2 PCA-based approach for crafting a UAP

Inputs:
- a random subset of input data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and their corresponding labels
- the model $f(., \boldsymbol{\theta})$
- maximum allowed perturbation norm p_{max}

Output: a UAP $\mathbf{r}$

1: Evaluate $\mathbf{X}^{p \times N} = [\mathbf{n}_{\mathbf{x}_1}, \ldots, \mathbf{n}_{\mathbf{x}_N}]$.
2: Compute the first principal direction of $\mathbf{X}$ and denote it by $\mathbf{v}_1$, i.e., $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ and $\mathbf{v}_1 = \mathbf{V}\,\mathbf{e}_1$.
3: $\mathbf{r} = p_{max}\mathbf{v}_1$.

PSR [dB]	-10	-12	-14	-16	-18	-20
Time required by [10]	20.5	23.0	25.1	27.2	29.0	30.5
Time required by Alg. 2	0.3	0.3	0.3	0.3	0.3	0.3

TABLE I: Run time of Alg. 2 compared to [10] in seconds, for SNR= 10 dB and $N = 50$.

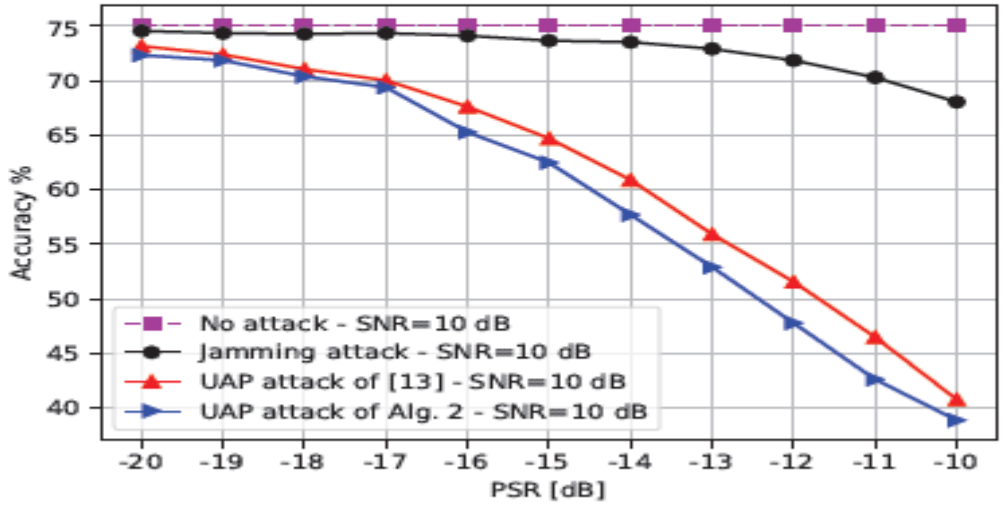


Fig. 3: The accuracy of VT-CNN2 under different attacks.

VII. Ataques de caixa preta e propriedades invariantes de turno de UAPs

1) assumimos que o atacante possui o conhecimento perfeito do modelo.

2) é síncrono com o transmissor.

Para Solucionar esses Problemas

- criar um UAP para VT-CNN2, primeiro criamos esse UAP para um DNN substituto e depois o aplicamos no VT-CNN2

- consideramos um perceptron multicamadas (MLP)

- totalmente conectado como nosso DNN substituto e criamos um UAP para ele.

- A Figura abaixo mostra o desempenho de dois ataques UAP projetados usando Alg. 2

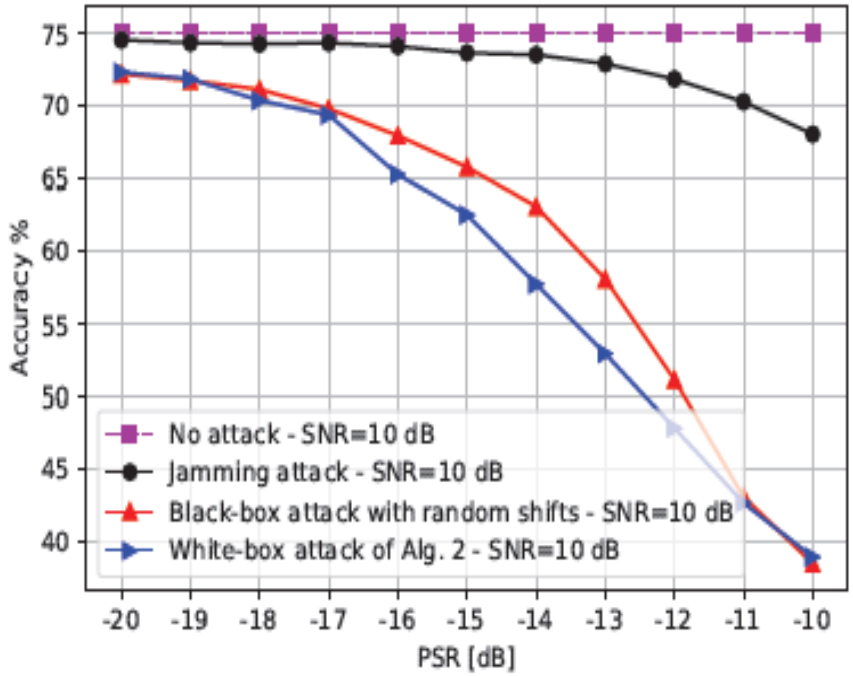


Fig. 4: An illustration of transferability and shift invariant properties of the proposed UAP attack.

VI.CONCLUSÃO

- Os Algoritmos baseados em DL para classificação de sinais de rádio são também são vulneráveis a ataque diversos.

-Os resultados dessa proposta pode se notar que quando temos menos energia de transmissão é necessário favorece o invasor para causar erros de classificação em comparação com o caso de interferência convencional (onde o invasor transmite apenas ruído aleatório).
- Isso expõe uma vulnerabilidade fundamental das soluções baseadas em DL.