Przegląd ataków adversarialnych na sieci konwolucyjne (CNN)

Wojciech Bartoszek
Łukasz Checiak

Opiekun: prof. dr hab. inż. Rafał Scherer

Abstrakt

Sieci konwolucyjne (CNN) zrewolucjonizowały przetwarzanie obrazów, jednak pozostają podatne na ataki adversarialne — celowo wprowadzone subtelne zakłócenia w danych wejściowych, które wprowadzają błędy w przewidywaniach modeli. W niniejszej pracy eksperymentalnej przeanalizowano metody ataków adversarialnych na CNN, ich skutki oraz strategie obronne. Scharakteryzowano podstawowe techniki takie jak Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), ataki Carliniego-Wagnera oraz DeepFool, ze szczególnym uwzględnieniem ich podstaw matematycznych i skuteczności praktycznej. Przeprowadzono eksperymenty na podzbiorze danych walidacyjnych ImageNet oraz obrazach hiperspektralnych z wykorzystaniem pięciu architektur sieci konwolucyjnych (ResNet50, VGG16, DenseNet121, MobileNetV2, HybridSN). Dodatkowo, wykonano systematyczną analizę wpływu kluczowych parametrów każdej metody ataku na ich skuteczność, ujawniając nieliniowe zależności i optymalne konfiguracje parametrów. Omówiono mechanizmy obronne, w tym przetwarzanie wstępne danych poprzez kompresję JPEG, wskazując na ich skuteczność w mitigacji ataków adversarialnych. Wykazano uniwersalną podatność architektur CNN na ataki, przy czym ResNet50 wykazuje najwyższą odporność, a ataki Carlini & Wagner okazują się najskuteczniejsze. Przeanalizowano także efekt kompresji JPEG jako naturalnego mechanizmu defensywnego. Wyniki wskazują na fundamentalne wyzwanie związane z równoważeniem dokładności i odporności modeli oraz konieczność rozwoju wielowarstwowych strategii ochronnych. Zaproponowano kierunki przyszłych badań obejmujące projektowanie inherentnie odpornych architektur oraz zaawansowane metody preprocessing'u defensywnego. Praca integruje perspektywę teoretyczną i praktyczną, oferując kompleksowy obraz wyzwań bezpieczeństwa w głębokich sieciach neuronowych.

1 Wprowadzenie

Współczesne sieci konwolucyjne (CNN) osiągnęły niezwykłe sukcesy w zadaniach rozpoznawania obrazów, jednak ich podatność na ataki adversarialne stanowi poważne wyzwanie dla bezpieczeństwa systemów opartych na uczeniu maszynowym[11]. Ataki adversarialne polegają na wprowadzeniu celowych, często niedostrzegalnych dla człowieka, modyfikacji do danych wejściowych, które prowadzą do błędnych przewidywań modelu. Zjawisko

to budzi szczególne obawy w kontekście zastosowań krytycznych, takich jak autonomiczne pojazdy, systemy medyczne czy bezpieczeństwo publiczne.

Celem niniejszej pracy jest przeprowadzenie kompleksowej analizy wpływu różnych metod ataków adversarialnych na wydajność wybranych architektur CNN. Badanie koncentruje się na czterech fundamentalnych technikach ataku:

- **FGSM (Fast Gradient Sign Method)**[2] — jednokrokowa metoda gradientowa charakteryzująca się wysoką efektywnością obliczeniową
- **PGD (Projected Gradient Descent)**[5] — iteracyjne rozszerzenie FGSM zapewniające większą skuteczność
- **Carlini & Wagner (C&W)**[1] — zaawansowany atak optymalizacyjny minimalizujący perturbacje
- **DeepFool**[6] — metoda geometryczna znajdująca minimalne zaburzenia prowadzące do zmiany klasyfikacji

Analiza empiryczna obejmuje cztery reprezentatywne architektury CNN, charakteryzujące się różnymi podejściami projektowymi:

- **ResNet50** — architektura wykorzystująca połączenia rezydualne
- **VGG16** — klasyczna głęboka sieć konwolucyjna
- **DenseNet121** — architektura z gęstymi połączeniami między warstwami
- **MobileNetV2** — efektywna obliczeniowo sieć przeznaczona na urządzenia mobilne

Eksperymenty przeprowadzono na reprezentatywnym podzbiorze zbioru walidacyjnego ImageNet, zawierającym 128 obrazów z różnych klas, co umożliwia kontrolowaną ocenę odporności modeli przy zachowaniu obliczeniowej wykonalności eksperymentów. Dodatkowo, w celu rozszerzenia zakresu analizy, zbadano także wpływ ataków na obrazy hiperspektralne z wykorzystaniem modelu HybridSN na zbiorze danych Indian Pines, oraz skuteczność kompresji JPEG jako metody defensywnej.

Istotnym elementem badania jest przeprowadzona systematyczna analiza wpływu kluczowych parametrów każdej z metod ataków na ich skuteczność. Dla ataku FGSM przeanalizowano wpływ parametru epsilon, dla PGD zbadano interakcje między epsilon, alpha oraz liczbą kroków iteracyjnych, dla ataku Carlini & Wagner oceniono wpływ współczynnika regularyzacji, learning rate i liczby kroków optymalizacji, natomiast dla DeepFool przebadano wpływ liczby kroków oraz parametru overshoot. Analiza ta ujawniła nieliniowe zależności między parametrami a skutecznością ataków oraz dostarczyła praktycznych wskazówek dotyczących optymalnej konfiguracji każdej metody.

Kod źródłowy implementacji oraz szczegółowe wyniki eksperymentów zostały udostępnione pod adresem: `https://github.com/f4rys/Cross-Domain-Adversarial-Analysis`

2 Metodologia ataków adversarialnych

W niniejszym badaniu zaimplementowano cztery fundamentalne metody ataków adversarialnych w frameworku PyTorch, wykorzystując specyficzne warianty i optymalizacje dostosowane do charakterystyki badanych modeli. Wszystkie implementacje bazują na oryginalnych algorytmach z literatury, lecz zostały zoptymalizowane pod kątem wydajności obliczeniowej i stabilności numerycznej.

2.1 Fast Gradient Sign Method (FGSM)

Fast Gradient Sign Method[2] stanowi fundament jednokrokowych ataków adversarialnych. W eksperymentach wykorzystano wariant *untargeted*, który maksymalizuje funkcję straty dla prawdziwej etykiety bez określania konkretnej klasy docelowej. Implementacja wykorzystuje CrossEntropyLoss i wykonuje gradient ascent względem danych wejściowych.

Matematycznie, adversarialny przykład x' generowany jest zgodnie z równaniem:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

gdzie x oznacza oryginalny obraz, ϵ parametr kontrolujący intensywność ataku. $J(\theta, x, y)$ funkcję straty modelu z parametrami θ dla danej pary (obraz, etykieta), a $\nabla_x J$ gradient tej funkcji względem danych wejściowych.

Główną zaletą FGSM jest jego efektywność obliczeniowa — wymaga jedynie jednokrotnego obliczenia gradientu.

2.2 Projected Gradient Descent (PGD)

Projected Gradient Descent[5] stanowi iteracyjne rozszerzenie metody FGSM, zaimplementowane jako wariant *untargeted* z normą ℓ_∞. Algorytm wykonuje wielokrotną aktualizację perturbacji z projekcją na kulę ℓ_∞ o promieniu ϵ.

Algorytm PGD można sformalizować jako:

$$x^{t+1} = \Pi_S(x^t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^t, y))) \tag{2}$$

gdzie Π_S oznacza operację projekcji na dopuszczalny zbiór perturbacji S (kulę ℓ_∞ o promieniu ϵ), α rozmiar kroku, a t numer iteracji.

Implementacja zawiera mechanizm wczesnego zatrzymywania — jeśli gradient względem obrazu adversarialnego wynosi zero (co może oznaczać maksymalne błędne przewidywanie), algorytm przerywa iteracje dla danej próbki.

2.3 Carlini & Wagner (C&W)

Atak Carlini & Wagner[1] reprezentuje zaawansowane podejście optymalizacyjne zaimplementowane jako wariant *untargeted* z normą ℓ_2. W przeciwieństwie do metod gradientowych, C&W formułuje problem jako zadanie optymalizacji z ograniczeniami wykorzystując AdamW optimizer.

Implementacja wykorzystuje transformację atanh dla zapewnienia, że wartości pikseli pozostają w dopuszczalnym zakresie:

$$x' = \frac{1}{2}(\tanh(w) + 1), \quad w = \operatorname{atanh}(2x - 1) \tag{3}$$

Funkcja celu łączy stratę klasyfikacyjną z regularyzacją ℓ_2:

$$\min \|x' - x\|_2 + c \cdot f(x') \tag{4}$$

gdzie $f(x') = \max(Z(x')_y - \max_{j \neq y} Z(x')_j + \kappa, 0)$ z $Z(x')$ oznaczającym logity.

Implementacja zawiera mechanizm $best - candidatetracking$ — przechowuje najlepsze adversarialne przykłady (o najmniejszej normie ℓ_2) spośród tych, które skutecznie wprowadzają błędną klasyfikację.

2.4 DeepFool

DeepFool[6] wykorzystuje geometryczne podejście implementowane jako wariant *untargeted* minimalizujący normę ℓ_2. Algorytm iteracyjnie aproksymuje granice decyzyjne klasyfikatora poprzez znajdowanie najbliższej hiperpłaszczyzny separującej klasy.

Implementacja wykorzystuje obliczenia macierzy Jacobian dla efektywnego wyliczenia gradientów względem wszystkich klas jednocześnie. W każdej iteracji t, DeepFool oblicza minimalną perturbację r_t potrzebną do przekroczenia lokalnej granicy decyzyjnej:

$$r_t = \frac{|f_k(x_t) - f_{j^*}(x_t)|}{\|\nabla f_k(x_t) - \nabla f_{j^*}(x_t)\|_2^2} (\nabla f_k(x_t) - \nabla f_{j^*}(x_t)) \tag{5}$$

gdzie j^* oznacza klasę znajdującą się najbliżej granicy decyzyjnej, f_k różnicę logitów między klasą aktualnie przewidywaną a klasą k.

3 Architektury sieci neuronowych

W niniejszym badaniu wykorzystano cztery reprezentatywne architektury CNN oraz jedną architekturę dedykowaną obrazom hiperspektralnym. Każda z nich charakteryzuje się odmiennym podejściem projektowym i różnymi mechanizmami uczenia. Wybór tych modeli pozwala na kompleksową ocenę wpływu różnych strategii architekturalnych na odporność względem ataków adversarialnych.

- **ResNet50**[3] — architektura wprowadzająca koncepcję połączeń rezydualnych (skip connections), które umożliwiają trenowanie bardzo głębokich sieci poprzez rozwiązanie problemu zaniku gradientu. Połączenia te pozwalają na bezpośredni przepływ informacji między odległymi warstwami, co teoretycznie może wpływać na odporność modelu na perturbacje adversarialne.

- **VGG16**[10] — klasyczna architektura charakteryzująca się prostą, sekwencyjną strukturą złożoną z małych filtrów konwolucyjnych (3×3). Mimo swojej prostoty, VGG16 pozostaje ważnym punktem odniesienia w badaniach nad sieciami konwolucyjnymi oraz stanowi reprezentację tradycyjnych podejść architekturalnych.

- **DenseNet121**[4] — sieć wykorzystująca gęste połączenia, w której każda warstwa otrzymuje informacje ze wszystkich poprzedzających warstw. Taka architektura promuje ponowne wykorzystanie cech (feature reuse) i może prowadzić do lepszej regularyzacji, co potencjalnie wpływa na odporność adversarialną.
- **MobileNetV2**[9] — efektywna obliczeniowo architektura zaprojektowana z myślą o urządzeniach o ograniczonych zasobach. Wykorzystuje separowalne konwolucje depthwise oraz połączenia rezydualne w blokach z wąskim gardłem (inverted residuals), co znacząco redukuje liczbę parametrów.
- **HybridSN**[8] — specjalistyczna architektura przeznaczona do klasyfikacji obrazów hiperspektralnych. Model ten wykorzystuje hybrydowe podejście, łącząc konwolucje 3D (do ekstrakcji cech spektralno-przestrzennych) z konwolucjami 2D (do dalszej ekstrakcji cech przestrzennych). Taka hierarchiczna struktura pozwala na efektywne przetwarzanie danych hiperspektralnych o wysokiej wymiarowości.

Wszystkie modele CNN zostały wczytane jako przedtrenowane wersje z biblioteki torchvision, przeszkolone na pełnym zbiorze danych ImageNet. W niniejszych eksperymentach wykorzystano implementację modelu HybridSN opartą na PyTorch oraz przedtrenowane wagi dostępne w repozytorium GitHub[7]. W eksperymentach wykorzystano reprezentatywny podzbiór 128 obrazów z różnych klas ze zbioru walidacyjnego ImageNet dla modeli CNN, oraz 8192 próbek z 16 klas dla modelu HybridSN na zbiorze Indian Pines, co umożliwia kontrolowaną ocenę odporności modeli przy zachowaniu obliczeniowej wykonalności eksperymentów. Przed przystąpieniem do ataków adversarialnych dokonano oceny baseline'owej dokładności na wybranych podzbiorach.

4 Metodologia eksperymentalna i wyniki

4.1 Konfiguracja eksperymentów

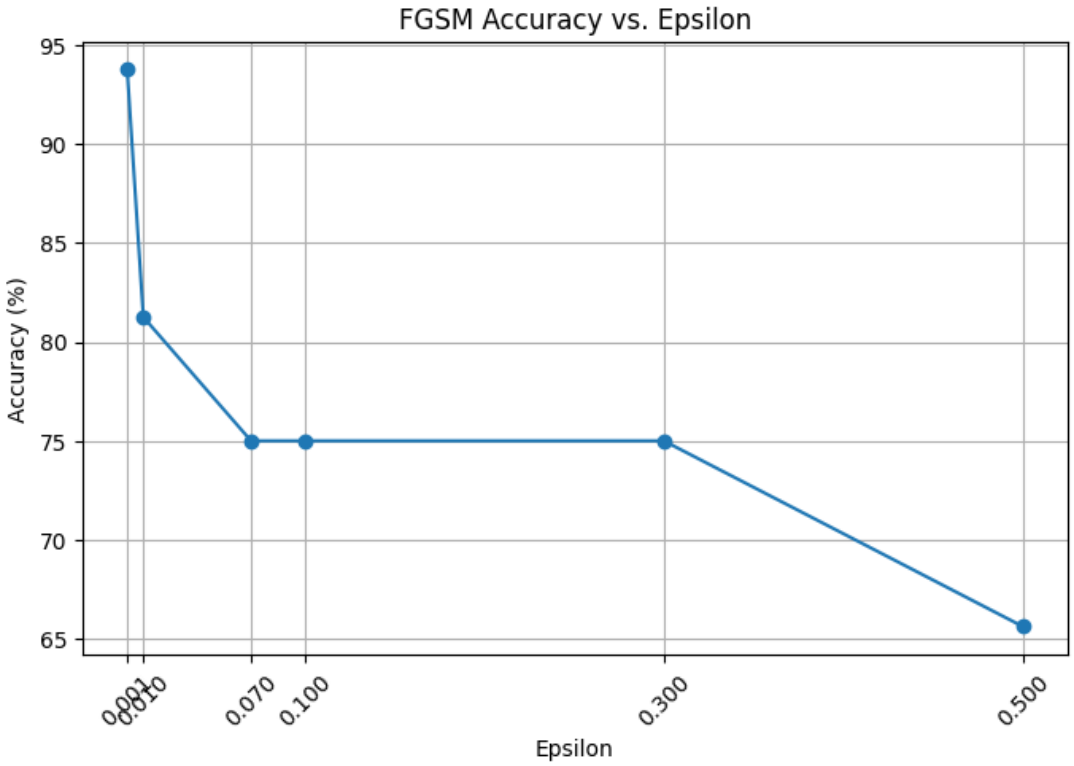
Wszystkie eksperymenty zostały przeprowadzone w środowisku PyTorch z wykorzystaniem GPU NVIDIA dla przyspieszenia obliczeń. Dla zapewnienia powtarzalności wyników, ziarno losowości zostało ustalone na stałą wartość. Każdy model był poddawany atakom w odrębnych sesjach w celu uniknięcia interferencji między eksperymentami.

4.2 Analiza wpływu parametrów ataków adversarialnych

W celu głębszego zrozumienia mechanizmów działania ataków adversarialnych przeprowadzono systematyczną analizę wpływu kluczowych parametrów na skuteczność każdej z badanych metod. Eksperymenty zostały wykonane na architekturze ResNet50 z wykorzystaniem 128 obrazów z podzbioru ImageNet, co umożliwiło kontrolowaną ocenę zachowania ataków przy różnych konfiguracjach parametrów.

4.2.1 Analiza parametru epsilon w ataku FGSM

Parametr ϵ w metodzie FGSM kontroluje magnitudo perturbacji adversarialnej i bezpośrednio wpływa na siłę ataku. Przeprowadzono eksperymenty dla wartości w zakresie $\epsilon \in \{0.001, 0.01, 0.07, 0.1, 0.3, 0.5\}$.

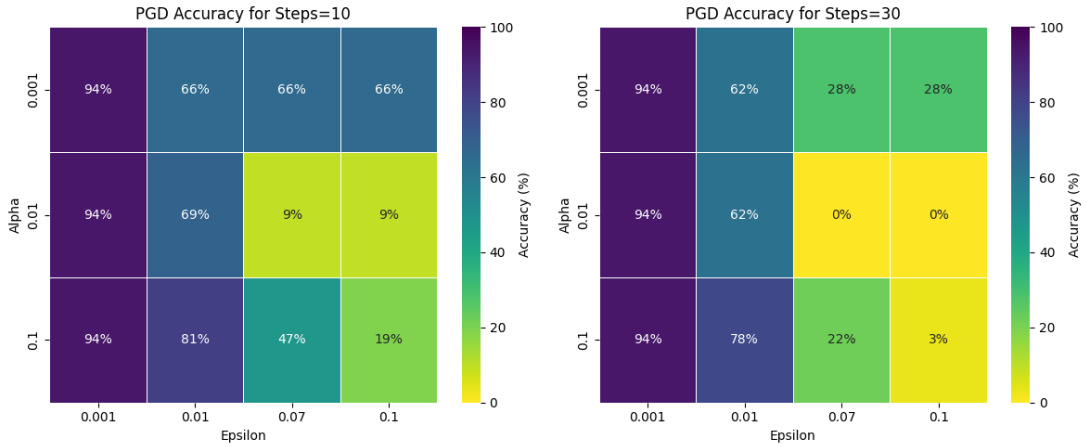


Rysunek 1: Wpływ parametru epsilon na skuteczność ataku FGSM względem architektury ResNet50

Analiza wyników (Rysunek 1) ujawnia nieliniową zależność między wartością ϵ a skutecznością ataku. Dla małych wartości ($\epsilon = 0.001$) dokładność modelu pozostaje na wysokim poziomie (93.75%), co wskazuje na niedostateczną siłę perturbacji. Istotna degradacja wydajności występuje dla $\epsilon \geq 0.01$, gdzie dokładność spada do około 80%. Szczególnie interesujący jest plateau obserwowany dla wartości $\epsilon \in \{0.07, 0.1, 0.3\}$, gdzie dokładność ustabilizowała się na poziomie 75%. Dalsze zwiększenie do $\epsilon = 0.5$ prowadzi do dodatkowej degradacji do 65.62%.

4.2.2 Analiza parametrów ataku PGD

Metoda PGD charakteryzuje się trzema kluczowymi parametrami: epsilon (ϵ), alpha (α) oraz liczbą kroków iteracyjnych. Przeprowadzono systematyczną analizę dla kombinacji: $\epsilon \in \{0.001, 0.01, 0.07, 0.1\}$, $\alpha \in \{0.001, 0.01, 0.1\}$, oraz liczby kroków $\in \{10, 30\}$.

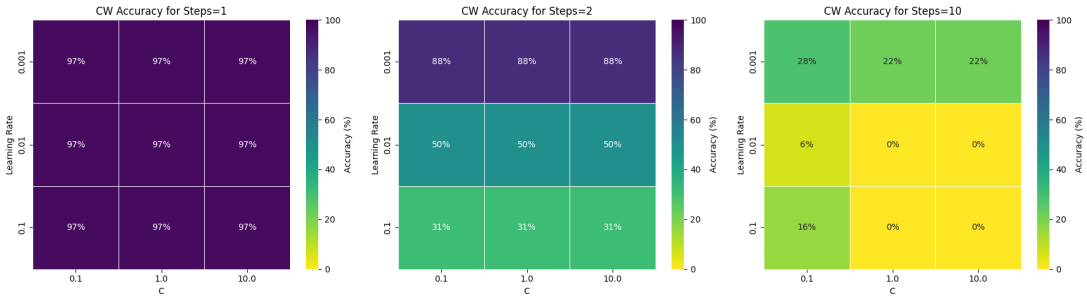


Rysunek 2: Analiza heatmap skuteczności ataku PGD w funkcji parametrów epsilon i alpha dla różnych liczb kroków

Analiza heatmap (Rysunek 2) dostarcza kompleksowego obrazu interakcji między parametrami PGD. Dla 10 kroków obserwuje się, że kombinacja wysokiego ϵ (0.07-0.1) z umiarkowanym α (0.01) prowadzi do najskuteczniejszych ataków. Zwiększenie liczby kroków do 30 wzmacnia tę tendencję, umożliwiając osiągnięcie zerowej dokładności dla szerszego zakresu kombinacji parametrów. Szczególnie widoczne jest, że zbyt wysokie wartości α (0.1) mogą paradoksalnie zmniejszać skuteczność ataku poprzez wprowadzenie niestabilności w procesie optymalizacji.

4.2.3 Analiza parametrów ataku Carlini & Wagner

Atak C&W charakteryzuje się kilkoma kluczowymi parametrami: współczynnikiem regularyzacji c, learning rate, liczbą kroków optymalizacji oraz parametrem confidence κ. Analiza została przeprowadzona dla $c \in \{0.1, 1, 10\}$, learning rate $\in \{0.001, 0.01, 0.1\}$, oraz liczby kroków $\in \{1, 2, 10\}$ przy stałym $\kappa = 0$.



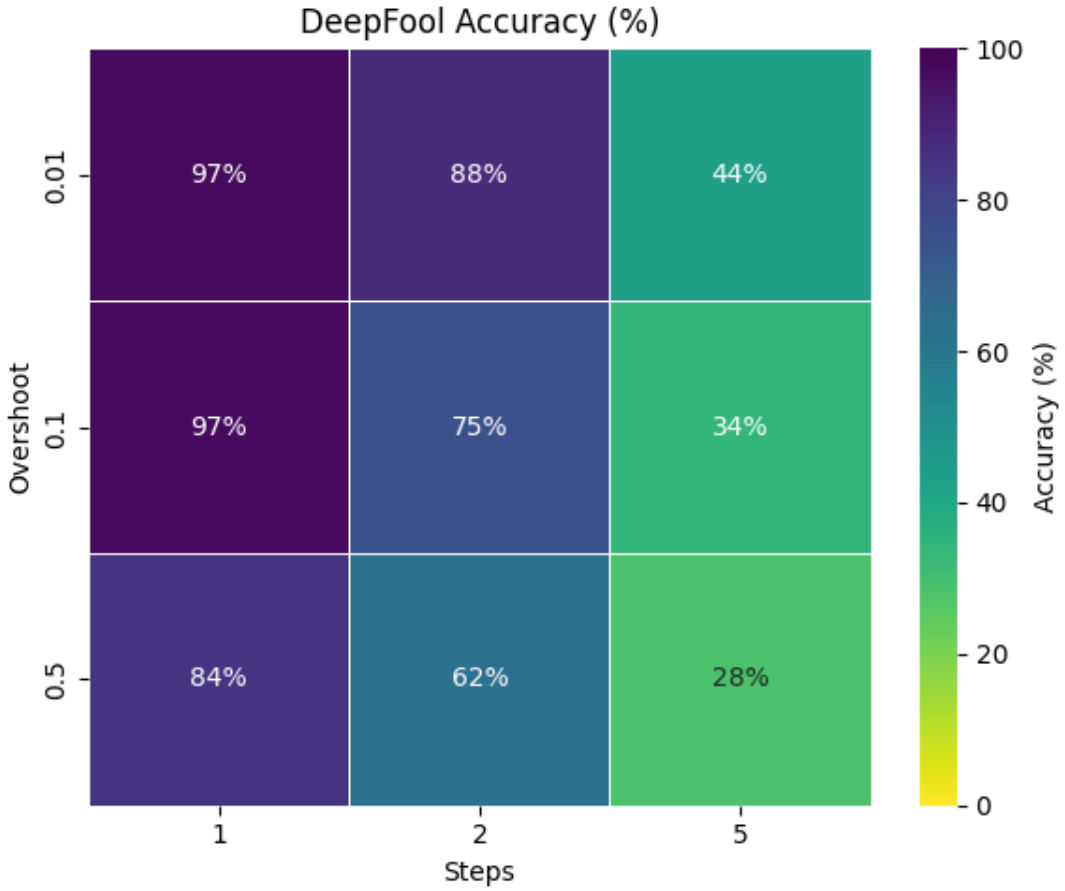
Rysunek 3: Analiza heatmap skuteczności ataku C&W w funkcji parametrów C i learning rate dla różnych liczb kroków

Heatmap C&W (Rysunek 3) ujawnia złożone interakcje między parametrami optymalizacji. Dla 1 kroku żadna z kombinacji parametrów nie jest skuteczna, co potwierdza potrzebę iteracyjnej optymalizacji w tej metodzie. Zwiększenie do 2 kroków znacząco poprawia skuteczność, szczególnie dla kombinacji średnich wartości c i learning rate. Dalsze

zwiększenie do 10 kroków prowadzi do osiągnięcia zerowej dokładności dla większości kombinacji parametrów, demonstrując potęgę tej metody przy odpowiedniej konfiguracji.

4.2.4 Analiza parametrów ataku DeepFool

Metoda DeepFool charakteryzuje się dwoma głównymi parametrami: liczbą kroków iteracyjnych oraz parametrem overshoot kontrolującym margines przekroczenia granicy decyzyjnej. Analiza została przeprowadzona dla liczby kroków $\in \{1, 2, 5, 10\}$ oraz overshoot $\in \{0.01, 0.1, 0.5\}$.



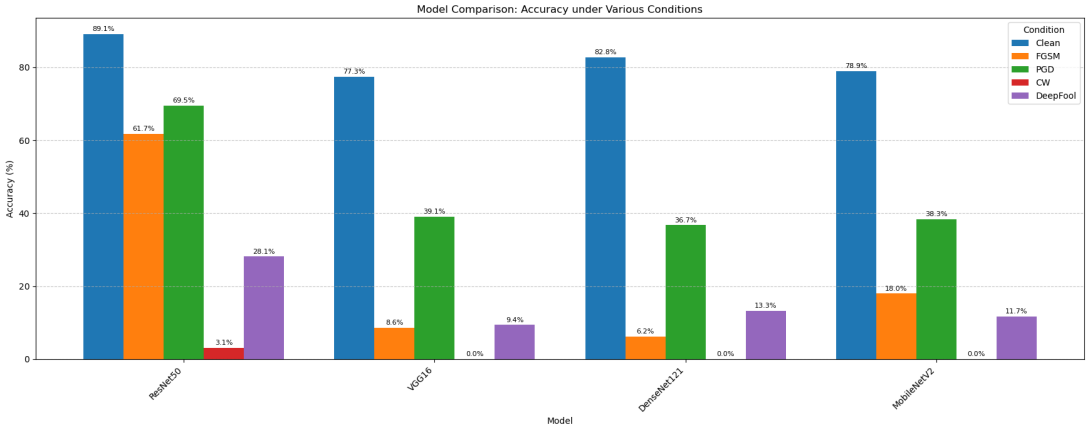
Rysunek 4: Analiza heatmap skuteczności ataku DeepFool w funkcji liczby kroków i parametru overshoot

Heatmap DeepFool (Rysunek 4) ujawnia , że liczba kroków ma kluczowe znaczenie dla skuteczności ataku. Dla 1 kroku atak jest praktycznie nieskuteczny, co potwierdza potrzebę iteracyjnej natury tej metody. Zwiększenie liczby kroków do 2 lub 5 znacząco poprawia skuteczność, osiągając spadek dokładności. Większy overshoot (0.5) prowadzi do jeszcze większej degradacji wydajności, co sugeruje, że atak DeepFool jest wrażliwy na ten parametr.

4.3 Wyniki ataków na klasyfikację obrazów

Parametry ataków zostały ustalone następująco: $\epsilon = 0.03$ dla ataków FGSM i PGD, liczba iteracji PGD $= 1$, rozmiar kroku $\alpha = 0.005$. Dla ataku C&W wykorzystano parametr confidence $\kappa = 0$ oraz maksymalnie 5 iteracji optymalizacji oraz parametrem $c = 5$. DeepFool był uruchamiany z maksymalnie 5 iteracjami oraz parametrem *overshoot* $= 0.05$.

Ocena wpływu ataków adversarialnych została przeprowadzona poprzez pomiar dokładności klasyfikacji przed i po zastosowaniu perturbacji. Wyniki przedstawione na Rysunku 5 oraz w Tabeli 1 ujawniają znaczące różnice w odporności poszczególnych architektur.



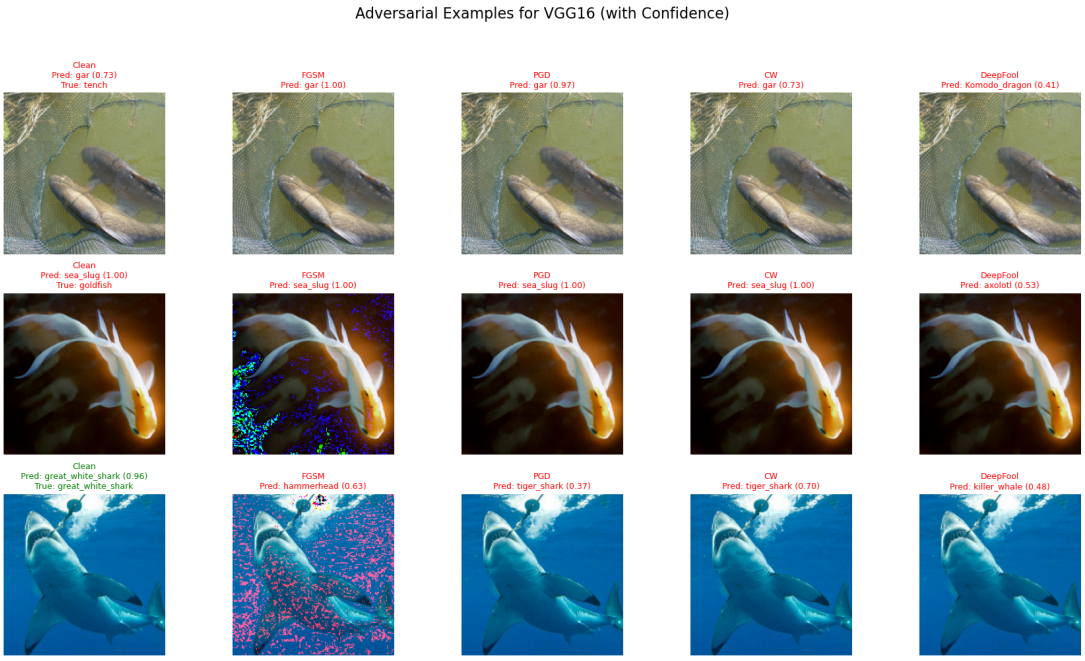
Rysunek 5: Wpływ różnych ataków adversarialnych na dokładność klasyfikacji poszczególnych architektur CNN

Model	Clean	FGSM	PGD	C&W	DeepFool
ResNet50	89.06%	61.72%	69.53%	3.12%	28.12%
VGG16	77.34%	8.59%	39.06%	0.00%	9.38%
DenseNet121	82.81%	6.25%	36.72%	0.00%	13.28%
MobileNetV2	78.91%	17.97%	38.28%	0.00%	11.72%

Tabela 1: Dokładność klasyfikacji poszczególnych modeli przed i po zastosowaniu ataków adversarialnych

Analiza przedstawionych rezultatów ujawnia znaczące różnice w podatności badanych architektur. ResNet50 wykazuje najwyższą odporność na większość typów ataków, co może być związane z mechanizmem połączeń rezydualnych ułatwiającym gradientowy przepływ informacji. W przeciwieństwie do tego, architektura VGG16 okazuje się być najbardziej podatna, szczególnie w przypadku ataku FGSM, gdzie dokładność spada do zaledwie 8.59%.

Szczególnie istotnym obserwacjom podlega uniwersalna skuteczność ataku C&W, który powoduje niemal całkowitą degradację wydajności wszystkich testowanych modeli. Wynik ten potwierdza przewagę metod optymalizacyjnych nad jednokrokowymi atakami gradientowymi w kontekście znajdowania minimalnych perturbacji adversarialnych.



Rysunek 6: Porównanie wizualnej percepcji oryginalnych obrazów i ich adversarialnych odpowiedników

Analiza wizualna przedstawionych przykładów adversarialnych (Rysunek 6) ujawnia fundamentalną właściwość tego typu ataków — ich subtelność względem ludzkiej percepcji. Podczas gdy perturbacje wprowadzone przez atak FGSM są wizualnie dostrzegalne jako zwiększony szum w obrazie, modyfikacje wynikające z zastosowania PGD, C&W oraz DeepFool pozostają praktycznie niewidoczne dla ludzkiego oka.

Ta obserwacja ma kluczowe znaczenie dla bezpieczeństwa systemów wizyjnych, ponieważ wskazuje na możliwość wprowadzenia złośliwych modyfikacji do obrazów bez wzbudzenia podejrzeń u operatorów lub użytkowników końcowych. Szczególnie problematyczne jest to w kontekście zastosowań krytycznych, gdzie decyzje podejmowane przez systemy CNN mogą mieć bezpośredni wpływ na bezpieczeństwo lub zdrowie.

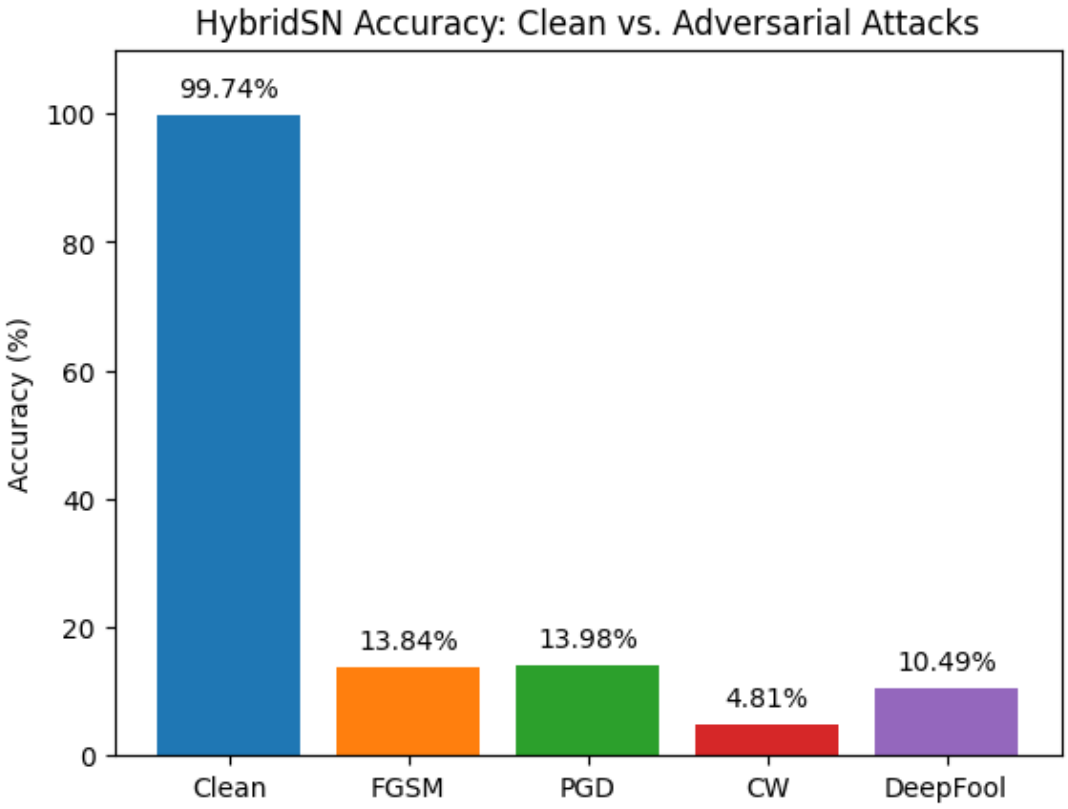
4.4 Analiza ataków na dane hiperspektralne

Obrazy hiperspektralne charakteryzują się znacznie większą złożonością informacyjną w porównaniu do konwencjonalnych obrazów RGB, zawierając setki pasm spektralnych dla każdego piksela. Ta właściwość czyni je szczególnie wartościowymi w zastosowaniach takich jak teledetekcja, monitoring środowiskowy, czy analiza medyczna, ale jednocześnie stwarza nowe wyzwania w kontekście ataków adversarialnych.

W niniejszym badaniu zastosowano architekturę HybridSN[8], specjalnie zaprojektowaną do klasyfikacji obrazów hiperspektralnych poprzez kombinację konwolucji 3D i 2D. Model ten wykorzystuje zarówno informacje przestrzenne jak i spektralne, co teoretycznie może wpływać na jego odporność względem perturbacji adversarialnych.

W celu zapewnienia stabilności eksperymentów na danych hiperspektralnych, parametry ataków zostały skalibrowane na niższe wartości w porównaniu do eksperymentów na obrazach ImageNet: FGSM z $\epsilon = 0{,}01$, PGD z $\epsilon = 0{,}01$, $\alpha = 0{,}003$ i 1 iteracją, C&W

z parametrem $c = 100$, $\kappa = 0$, 20 iteracjami i learning rate 0,05, oraz DeepFool z 1 iteracją i parametrem overshoot 0,01. Mimo zastosowania słabszych parametrów ataków, model HybridSN wykazał znacznie większą podatność niż CNN testowane na ImageNet z silniejszymi parametrami.

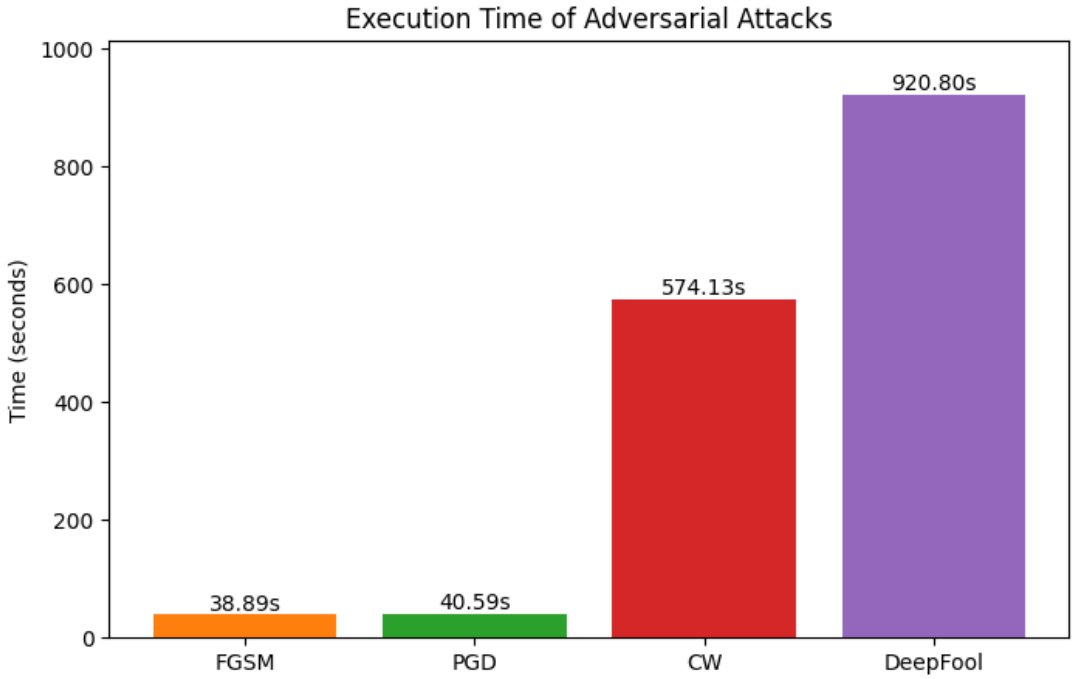


Rysunek 7: Dokładność klasyfikacji obrazów hiperspektralnych po zastosowaniu różnych typów ataków adversarialnych

Wyniki przedstawione na Rysunku 7 ujawniają znaczące różnice w odporności modelu HybridSN na ataki adversarialne w porównaniu do konwencjonalnych architektur CNN. Wszystkie badane ataki degradują skuteczność HybridSN w większym stopniu niż jakąkolwiek z badanych architektur CNN, co stanowi istotną obserwację dla bezpieczeństwa systemów klasyfikacji hiperspektralnej.

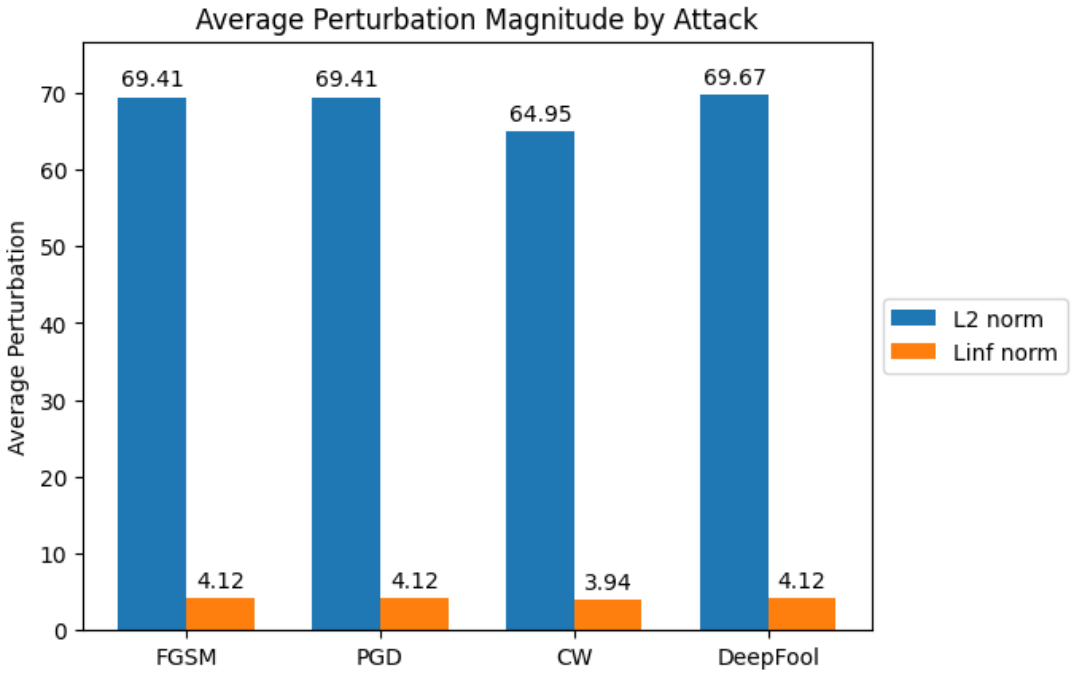
Ta zwiększona podatność HybridSN może wynikać z architekturalnych czynników:

- **Prostota architektury** — HybridSN wykorzystuje stosunkowo prostą kombinację warstw konwolucyjnych 3D i 2D bez zaawansowanych mechanizmów regularyzacyjnych obecnych w nowoczesnych CNN, takich jak połączenia rezydualne w ResNet50 czy normalizacja batch.
- **Mniejsza głębokość sieci** — W porównaniu do głębokich architektur CNN (ResNet50: 50 warstw, DenseNet121: 121 warstw), HybridSN charakteryzuje się znacznie mniejszą liczbą warstw, co może ograniczać jego zdolność do uczenia się odpornych reprezentacji.



Rysunek 8: Analiza czasowa wykonania ataków adversarialnych na dane hiperspektralne

Analiza efektywności obliczeniowej (Rysunek 8) ujawnia wyraźny podział ataków na dwie kategorie. Metody gradientowe (FGSM, PGD) charakteryzują się wysoką efektywnością czasową, wykonując się w czasie około 40 sekund. W przeciwieństwie do tego, ataki optymalizacyjne (C&W, DeepFool) wymagają znacznie większych zasobów obliczeniowych, z czasami wykonania sięgającymi odpowiednio około 10 i 15 minut.



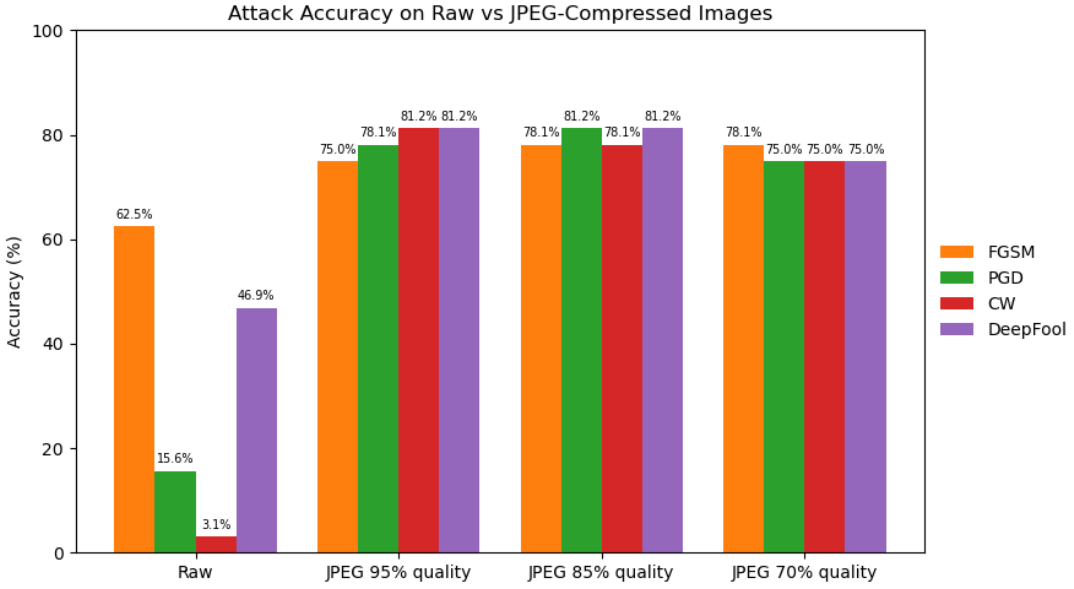
Rysunek 9: Porównanie magnitud perturbacji adversarialnych mierzonych normami ℓ_2 i ℓ_∞

Analiza magnitud perturbacji została przeprowadzona w celu weryfikacji, czy parametry poszczególnych ataków zostały dobrane tak, aby siła generowanych zaburzeń była porównywalna, co jest warunkiem koniecznym dla rzetelnej oceny ich skuteczności. W analizie wykorzystano dwie fundamentalne normy:

- norma ℓ_2 - mierzy całkowitą energię zaburzenia, czyli jego długość w przestrzeni euklidesowej. W praktyce oznacza to, że jeśli zaburzenie jest rozproszone równomiernie po wielu pikselach i pasmach spektralnych, ale o małej sile, to norma ℓ_2 będzie niska. Jest to często stosowane do oceny subtelnych, „gładkich" perturbacji.
- norma ℓ_∞ - mierzy maksymalną zmianę wartości w jednym pikselu. Określa więc, jak bardzo zaburzenie może zmienić dowolną pojedynczą wartość w obrazie. Tę normę stosuje się, gdy celem jest ograniczenie „najgorszego przypadku" zmiany, nawet jeśli cała reszta obrazu pozostaje praktycznie nietknięta.

Wartości uzyskane są bardzo podobne, norma ℓ_2 dla C&W jest trochę mniejsza, co może sugerować, że potrzebna była mniejsza czułość do skutecznego przeprowadzenia ataku. Uzyskane wyniki raczej pokazują odporność samego zbioru danych na ataki, niż skuteczność ich samą w sobie. Ogólnie więc możemy ten wykres interpretować tak, że potrzebne było dość duże zaburzenie globalne (norma ℓ_2) oraz że zmiana lokalna (maksymalna dla pojedynczego piksela) może być widoczna, jednak raczej dla obrazów o niższej rozdzielczości (ℓ_∞).

4.5 Wpływ kompresji JPEG na skuteczność ataków adversarialnych



Rysunek 10: Wpływ kompresji JPEG na skuteczność ataków adversarialnych w architekturze ResNet50

Kompresja JPEG stanowi powszechnie stosowaną metodę redukcji rozmiaru plików obrazowych poprzez eliminację wysokoczęstotliwościowych składowych sygnału. W kontekście ataków adversarialnych, proces ten może nieintencjonalnie działać jako mechanizm defensywny poprzez usunięcie perturbacji zakodowanych w wysokich częstotliwościach przestrzennych.

Analiza ujawnia znaczący wpływ kompresji JPEG na mitigację ataków adversarialnych. Po zastosowaniu kompresji stratnej, dokładność klasyfikacji wzrosła do poziomu około 80% dla wszystkich typów ataków, co w przypadku ataku C&W oznacza spektakularny wzrost o 77 punktów procentowych w porównaniu do 3% dokładności przed kompresją.

Ten fenomen można wytłumaczyć dwoma fundamentalnymi mechanizmami:

- **Eliminacja wysokoczęstotliwościowych artefaktów** — algorytm JPEG usuwa komponenty o wysokich częstotliwościach przestrzennych, które często zawierają perturbacje adversarialne generowane przez ataki takie jak FGSM czy PGD.
- **Efekt wygładzania** — proces kwantyzacji i kompresji wprowadza naturalne wygładzenie obrazu, co może osłabiać lokalne perturbacje adversarialne poprzez ich rozmycie w otaczającym kontekście przestrzennym.

Należy jednak podkreślić, że kompresja JPEG jako metoda defensywna ma istotne ograniczenia. Po pierwsze, jest to transformacja stratna, co może negatywnie wpływać na jakość danych wejściowych. Po drugie, zaawansowane ataki mogą być dostosowane do odporności na tego typu preprocessing, co potencjalnie mogłoby zniwelować obserwowany efekt defensywny.

5 Analiza wyników

Z przedstawionych wyników wynika, że skuteczność ataków adversarialnych znacząco różni się w zależności od użytego modelu i rodzaju ataku. Najważniejsze obserwacje to:

- **ResNet50** wykazuje największą odporność na większość ataków. Dla ataku PGD dokładność spadła do 69,5%, co nadal jest relatywnie wysokim wynikiem w porównaniu do innych modeli. Również przy ataku FGSM ResNet50 zachowuje stosunkowo dobrą skuteczność (61,7%).
- **Atak Carlini & Wagner (C&W)** okazał się najskuteczniejszy — doprowadził do niemal całkowitej dezintegracji wszystkich modeli (dokładność od 0% do 3,1%). Świadczy to o jego precyzyjnej konstrukcji i zdolności do tworzenia zaburzeń minimalnie wpływających na percepcję wizualną, ale bardzo skutecznych wobec klasyfikatorów.
- **Modele VGG16, DenseNet121 oraz MobileNetV2** cechuje znacznie niższa odporność na ataki w porównaniu do ResNet50. VGG16, na przykład, osiąga zaledwie 8,6% dokładności przy ataku FGSM i całkowicie zawodzi przy ataku C&W.
- **DeepFool** osiąga umiarkowaną skuteczność — dokładność po ataku waha się od 9,4% do 28,1%. Ciekawym spostrzeżeniem jest jego wizualna subtelność, mimo że potrafi znacząco zaburzyć klasyfikację.
- **Hiperspektralne ataki** wykazały podobne właściwości — ataki C&W ponownie były najskuteczniejsze, co potwierdza ich uniwersalność. Natomiast czas ich wykonania oraz poziom wymaganych zaburzeń wskazują na kompromis pomiędzy skutecznością a kosztami obliczeniowymi.
- **Kompresja JPEG** okazuje się być interesującą metodą obrony — znacząco redukuje wpływ ataku, szczególnie dla C&W, gdzie dokładność wzrosła z 3,1% do około 80%. Oznacza to, że nawet proste techniki przetwarzania obrazu mogą w pewnych przypadkach niwelować efekty perturbacji.
- **Analiza parametrów** ujawnia nieliniowe zależności między siłą ataków a ich parametrami. Dla FGSM i PGD obserwuje się efekt nasycenia przy wysokich wartościach ϵ, podczas gdy ataki optymalizacyjne (C&W, DeepFool) wykazują większą stabilność względem parametrów. Szczególnie istotne jest odkrycie, że zwiększenie liczby iteracji w PGD i C&W prowadzi do dramatycznej poprawy skuteczności, co wskazuje na kluczową rolę czasu obliczeniowego w projektowaniu skutecznych ataków.

6 Podsumowanie i wnioski

Przeprowadzone badania eksperymentalne dostarczają kompleksowego obrazu podatności współczesnych architektur CNN na różne typy ataków adversarialnych. Analiza obejmująca cztery reprezentatywne architektury (ResNet50, VGG16, DenseNet121, MobileNetV2) oraz cztery fundamentalne metody ataków (FGSM, PGD, C&W, DeepFool) pozwala na sformułowanie następujących kluczowych wniosków:

6.1 Architekturalne determinanty odporności adversarialnej

Wyniki eksperymentów jednoznacznie wskazują na znaczące różnice w odporności poszczególnych architektur CNN na ataki adversarialne. **ResNet50** wykazuje najwyższą odporność spośród badanych modeli, co można przypisać mechanizmowi połączeń rezydualnych ułatwiających stabilny przepływ gradientów i potencjalnie zwiększających regularność przestrzeni reprezentacji. W przeciwieństwie do tego, **VGG16** okazuje się najbardziej podatny na perturbacje, co może wynikać z jego prostej, sekwencyjnej architektury pozbawionej mechanizmów regularyzacyjnych obecnych w nowszych projektach.

6.2 Hierarchia skuteczności ataków adversarialnych

Analiza porównawcza różnych metod ataków ujawnia wyraźną hierarchię ich skuteczności:

1. **Carlini & Wagner (C&W)** — najskuteczniejszy atak we wszystkich scenariuszach, osiągający niemal zerową dokładność klasyfikacji dla wszystkich testowanych architektur. Jego przewaga wynika z zaawansowanego podejścia optymalizacyjnego minimalizującego perturbacje przy maksymalizacji skuteczności.

2. **DeepFool** — druga w kolejności skuteczność, wykorzystująca geometryczne podejście do znajdowania minimalnych perturbacji względem granic decyzyjnych.

3. **PGD** — iteracyjne rozszerzenie FGSM wykazujące umiarkowaną skuteczność, szczególnie przeciwko modelom o większej odporności jak ResNet50.

4. **FGSM** — jednokrokowa metoda o najniższej skuteczności, ale najwyższej efektywności obliczeniowej.

6.3 Implikacje dla bezpieczeństwa systemów wizyjnych

Obserwowana subtelność wizualna perturbacji adversarialnych, szczególnie w przypadku ataków C&W i DeepFool, stwarza poważne zagrożenie dla bezpieczeństwa systemów opartych na CNN. Możliwość wprowadzenia niepostrzegalnych modyfikacji powodujących błędne klasyfikacje ma krytyczne znaczenie w zastosowaniach takich jak:

- Systemy autonomicznych pojazdów
- Diagnostyka medyczna oparta na analizie obrazów
- Systemy bezpieczeństwa i nadzoru
- Aplikacje rozpoznawania biometrycznego

6.4 Podatność danych hiperspektralnych

Rozszerzenie analizy na obrazy hiperspektralne potwierdza uniwersalność problemu ataków adversarialnych niezależnie od modalności danych. Architektura HybridSN, mimo swojej specjalizacji w przetwarzaniu informacji spektralno-przestrzennych, wykazuje podobną podatność na ataki jak konwencjonalne CNN. Ta obserwacja ma szczególne znaczenie dla zastosowań w teledetekcji, monitoringu środowiskowym oraz analizie materiałów.

6.5 Potencjał metod defensywnych

Badanie wpływu kompresji JPEG jako nieintencjonalnej metody defensywnej ujawnia obiecujący kierunek rozwoju mechanizmów ochronnych. Spektakularny wzrost dokładności klasyfikacji po kompresji (z 3% do 80% dla ataku C&W) wskazuje na potencjał prostych metod preprocessing'u w mitigacji ataków adversarialnych. Jednakże, należy uwzględnić ograniczenia tej metody, w tym degradację jakości obrazu oraz możliwość adaptacji ataków do odporności na kompresję.

6.6 Rekomendacje dla przyszłych badań

Na podstawie przeprowadzonej analizy można wskazać następujące kierunki dalszych badań:

1. **Rozwój architektur inherentnie odpornych** — projektowanie sieci neuronowych z wbudowanymi mechanizmami odporności na ataki adversarialne.
2. **Zaawansowane metody preprocessing'u defensywnego** — eksploracja różnych technik transformacji obrazów jako metod ochronnych.
3. **Teoretyczne podstawy transferowalności ataków** — głębsze zrozumienie mechanizmów przenoszenia perturbacji między różnymi architekturami.
4. **Standaryzacja metodologii oceny** — opracowanie ujednoliconych benchmarków dla oceny odporności adversarialnej.
5. **Ataki adaptacyjne** — badanie ataków dostosowanych do konkretnych mechanizmów defensywnych.

Podsumowując, wyniki niniejszego badania podkreślają fundamentalne wyzwanie jakim są ataki adversarialne dla współczesnych systemów uczenia głębokiego. Przeprowadzona analiza parametrów ujawnia kluczową rolę precyzyjnego dostrajania ataków dla maksymalizacji ich skuteczności przy zachowaniu subtelności wizualnej. Skuteczna ochrona przed tego typu zagrożeniami wymaga wielowymiarowego podejścia łączącego zaawansowane architektury, metody treningu adversarialnego oraz preprocessing defensywny. Jednocześnie, ciągła ewolucja technik ataku wymaga stałego rozwoju i adaptacji mechanizmów obronnych, ze szczególnym uwzględnieniem nieliniowych zależności między parametrami ataków a ich skutecznością.

Bibliografia

[1] Nicholas Carlini i David Wagner. „Towards evaluating the robustness of neural networks". W: *IEEE Symposium on Security and Privacy (SP)* (2017).

[2] Ian J Goodfellow, Jonathon Shlens i Christian Szegedy. „Explaining and harnessing adversarial examples". W: *arXiv preprint arXiv:1412.6572* (2014).

[3] Kaiming He i in. „Deep residual learning for image recognition". W: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, s. 770–778.

[4] Gao Huang i in. „Densely connected convolutional networks". W: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, s. 4700–4708.

[5] Aleksander Madry i in. „Towards deep learning models resistant to adversarial attacks". W: *International Conference on Learning Representations*. 2018.

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi i Pascal Frossard. „DeepFool: a simple and accurate method to fool deep neural networks". W: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, s. 2574–2582.

[7] Pancakerr. *HybridSN*. Implementation and pretrained weights, accessed: June 2025. 2022. URL: `https://github.com/Pancakerr/HybridSN`.

[8] Swalpa Kumar Roy i in. „HybridSN: Exploring 3D-2D CNN Feature Hierarchy for Hyperspectral Image Classification". W: *IEEE Geoscience and Remote Sensing Letters* 17 (2020).

[9] Mark Sandler i in. „MobileNetV2: Inverted residuals and linear bottlenecks". W: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, s. 4510–4520.

[10] Karen Simonyan i Andrew Zisserman. „Very deep convolutional networks for large-scale image recognition". W: *arXiv preprint arXiv:1409.1556* (2014).

[11] Christian Szegedy i in. „Intriguing properties of neural networks". W: *arXiv preprint arXiv:1312.6199* (2013).