CHUQUR NEYRON TARMOQLARIGA RAQOBATLI HUJUMLAR VA ULARNING BIOTIBBIY TASVIRLAR TAHLILIDAGI SAMARADORLIGI

Boltibayev Shuxratjon Komiljanovich
Namangan davlat universiteti dotsenti
Email: sh.boltibayev@gmail.com
Axmedov Doniyor Akmaljon o‘g‘li
Toshkent Kimyo Xalqaro Universiteti Namangan filiali magistranti
Email: axmedovdoniyor23@gmail.com

Annotatsiya: *Ushbu maqolada zamonaviy tibbiy diagnostika tizimlarining asosi bo‘lgan chuqur neyron tarmoqlarining (CNN) raqobatli hujumlarga (adversarial attacks) nisbatan zaifligi tadqiq etilgan. Tadqiqot jarayonida turli arxitekturalardagi modellar (InceptionV3, ResNet50, DenseNet121 va boshqalar) biotibbiy ma’lumotlar to‘plamlari yordamida testdan o‘tkazilgan. Ishda PGD, DeepFool va CW kabi asosiy hujum algoritmlarining samaradorligi qiyosiy tahlil qilinib, hatto yuqori aniqlikdagi modellar ham inson ko‘zi ilg‘amaydigan minimal o‘zgarishlar (perturbatsiyalar) natijasida noto‘g‘ri xulosa berishi isbotlangan. Maqola yakunida tibbiy sun’iy intellekt tizimlarining kiber-bardoshliligini oshirish bo‘yicha raqobatli o‘qitish va kirish filtrlarini joriy etish kabi amaliy tavsiyalar berilgan.*

Kalit so‘zlar: *Chuqur o‘rganish, raqobatli hujumlar, biotibbiy tasvirlar, neyron tarmoqlar turg‘unligi, PGD, DeepFool, CW algoritmi, kiber-bardoshlilik.*

КОНКУРЕНТНЫЕ АТАКИ НА ГЛУБОКИЕ НЕЙРОННЫЕ СЕТИ И ИХ ЭФФЕКТИВНОСТЬ В АНАЛИЗЕ БИОМЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ.

Болтибаев Шухратжон Комилжанович
Доцент, Наманганский государственный университет
Ахмедов Дониёр Акмалжон угли
Магистрант Наманганского филиала Международного университета Кимё в Ташкенте

Аннотация: *В данной статье исследуется уязвимость глубоких нейронных сетей (CNN), лежащих в основе современных медицинских диагностических систем, к атакам с использованием состязательных методов. В ходе исследования были протестированы модели с различными архитектурами (InceptionV3, ResNet50, DenseNet121 и др.) с использованием биомедицинских наборов данных. В работе представлен сравнительный анализ эффективности основных алгоритмов атак, таких как PGD, DeepFool и CW, и доказано, что даже высокоточные модели могут давать неверные выводы в результате минимальных изменений (возмущений), незаметных для человеческого глаза. В конце статьи даны практические рекомендации по повышению киберустойчивости медицинских систем*

искусственного интеллекта, такие как внедрение состязательного обучения и входных фильтров.

Ключевые слова: *глубокое обучение, атаки с использованием состязательных методов, биомедицинские изображения, стабильность нейронных сетей, PGD, DeepFool, алгоритм CW, киберустойчивость.*

COMPETITIVE ATTACKS ON DEEP NEURAL NETWORKS AND THEIR EFFICIENCY IN BIOMEDICAL IMAGE ANALYSIS

Boltibayev Shuhratjon Komiljanovich
Associate Professor, Namangan State University
Akhmedov Doniyor Akmaljon ogli
Master’s student, Namangan branch of Kimyo international university in Tashkent

Abstract: *This paper examines the vulnerability of deep neural networks (CNNs), which underlie modern medical diagnostic systems, to adversarial attacks. Models with various architectures (InceptionV3, ResNet50, DenseNet121, etc.) were tested using biomedical datasets. The paper presents a comparative analysis of the effectiveness of key attack algorithms, such as PGD, DeepFool, and CW, and demonstrates that even highly accurate models can generate incorrect conclusions due to minimal changes (perturbations) that are imperceptible to the human eye. The paper concludes with practical recommendations for improving the cyber resilience of medical artificial intelligence systems, such as the implementation of adversarial learning and input filters.*

Keywords*: deep learning, adversarial attacks, biomedical images, neural network stability, PGD, DeepFool, CW algorithm, cyber resilience.*

Bugungi kunda sun’iy intellekt (AI) tizimlari, xususan chuqur neyron tarmoqlari, tibbiy diagnostika, rentgen va gistologik tasvirlarni tahlil qilishda yuqori aniqlik ko‘rsatmoqda. Biroq, ushbu tizimlar "raqobatli hujumlar" (adversarial attacks) deb ataluvchi kutilmagan zaifliklarga ega. Tasvirga kiritilgan, inson ko‘zi ilg‘amaydigan minimal o‘zgarishlar (perturbatsiyalar) diagnostika aniqligini keskin pasaytirishi va noto‘g‘ri klinik xulosalarga sabab bo‘lishi mumkin. Tibbiyot kabi inson hayoti bilan bevosita bog‘liq sohada AI tizimlarining kiber-bardoshliligini (robustness) ta’minlash o‘ta muhim va dolzarb vazifadir.

Raqobatli hujumlar (adversarial attacks) — bu sun’iy intellekt tizimlarini, xususan chuqur neyron tarmoqlarini ataylab chalg‘itish san’atidir. Ushbu soha nisbatan yosh bo‘lishiga qaramay, qisqa vaqt ichida fundamental o‘zgarishlarni boshidan kechirdi.

Raqobatli hujumlar tushunchasini fanga olib kirgan ilk fundamental ish Kristian Szegedy va uning jamoasi tomonidan amalga oshirilgan. Ular tasvirga inson ko‘zi ilg‘amaydigan darajadagi kichik o‘zgarish kiritish orqali eng kuchli neyron tarmoqlarini ham adashtirish mumkinligini isbotladilar. Tadqiqotchilar ushbu xatolar tarmoq arxitekturasiga bog‘liq bo‘lmagan holda "umumiy" ekanligini, ya'ni bir model uchun

yaratilgan hujum boshqasini ham aldashi mumkinligini ko‘rsatib berishdi. Szegedy C. va boshqalar, "Neyron tarmoqlarning qiziqarli xususiyatlari" [1].

Ian Goodfellow tomonidan taklif etilgan usul sohadagi eng inqilobiy ishlardan biri bo‘ldi. Mualliflar FGSM (Fast Gradient Sign Method) algoritmini taklif qildilar. Ushbu metod hujumlarni tezkor generatsiya qilish imkonini beradi va modellarni bunday hujumlarga qarshi o‘qitish (adversarial training) uchun asos bo‘lib xizmat qiladi. Goodfellow neyron tarmoqlarining zaifligini ularning "o‘ta nochiziqliligi" bilan emas, balki yuqori o‘lchamli fazodagi chiziqli xususiyatlari bilan tushuntirdi. Goodfellow I. va boshqalar, "Adversarial misollarni tushuntirish va ulardan foydalanish" [2].

Seyed-Mohsen Moosavi-Dezfooli tomonidan ishlab chiqilgan usul optimallashtirishga asoslangan. Bu neyron tarmoqlarini chalg‘itishning oddiy va aniq usuli hisoblanadi. Algoritm tasvirga qo‘llaniladigan o‘zgarishlarni imkon qadar kichik (L_2 normasi bo‘yicha boshlang‘ich tasvirga juda yaqin) saqlashga intiladi. Moosavi-Dezfooli S.M. va boshqalar, "DeepFool: chuqur neyron tarmoqlarini chalg‘itishning oddiy va aniq usuli" [3].

Aleksandr Madry va uning jamoasi adversarial himoyani minimaks optimallashtirish masalasi sifatida ko‘rib chiqishni taklif qilishdi. Ular tomonidan ishlab chiqilgan PGD (Projected Gradient Descent) metodi hozirgi kunda eng kuchli "birinchi tartibli" hujum hisoblanadi. Agar model PGD hujumiga bardosh bera olsa, u boshqa ko‘plab kichikroq hujumlarga ham chidamli bo‘ladi. Madry A. va boshqalar, "Adversarial hujumlarga chidamli chuqur o‘rganish modellari sari" [4].

Nicholas Carlini va David Wagner ko‘plab mavjud "himoya mexanizmlari" aslida mustahkam emasligini isbotladilar.Ular taklif qilgan CW hujumi shunchalik kuchliki, u hatto maxsus himoyalangan tarmoqlarni ham chetlab o‘ta oladi. Bu tadqiqot model barqarorligini oddiy testlar bilan emas, balki murakkab va optimallashtirishga asoslangan hujumlar bilan tekshirish shartligini ko‘rsatdi. Carlini N. va Wagner D., "Neyron tarmoqlarning barqarorligini baholash sari" [5].

Tadqiqot jarayonida 5 ta asosiy tibbiy tasvirlar to‘plami va 8 ta klassifikatsiya masalasi ko‘rib chiqildi:

- Ma’lumotlar: Gistologik tasvirlar (metastazlar, tuxumdon/qalqonsimon bez o‘smalari), ko‘krak qafasi rentgen suratlari (X-ray) va o‘pka kompyuter tomografiyasi (CT).
- Arxitekturalar: InceptionV3, ResNet50, DenseNet121, MobileNet va Xception.
- Dasturiy vositalar: Python, PyTorch, TensorFlow va Adversarial Robustness Toolbox (ART).

Hujumchi tasvir x^* asl tasvir x ga maksimal darajada yaqin bo‘lishi ($\| x^* - x\| \le \tau$) va modelni noto‘g‘ri qaror chiqarishga majbur qilishi kerak.

PGD (Projected Gradient Descent) algoritmi. Ushbu usul iterativ ravishda gradient tushish texnikasini qo‘llaydi:

$$x_{k+1} = Clip_{x,\tau}(x_k + \alpha \cdot sign(\nabla_x L(x_k, y)))$$

Bu yerda τ — buzilish magnitudasi, α — o‘rganish koeffitsienti.

DeepFool algoritmi. Neyron tarmoq chiqishini chiziqlashtirish orqali klassifikatsiya chegarasiga (gipertekislikka) eng yaqin proyeksiyani topadi:

$$x_P = x_0 - \frac{f(x_0)}{\| w\|^2} w$$

CW (Carlini & Wagner) algoritmi. Bu usul nolinear dasturlash masalasini yechishga asoslangan bo‘lib, minimal buzilish bilan maksimal ishonchlilikka intiladi:

$$\min\| \frac{1}{2}(\tanh(w)+1) - x\| + [cite_s tart]c \cdot f_l(\frac{1}{2}(\tanh(w)+1))$$

1-jadval.

Ma’lumotlar to‘plami va modellarning aniqlik darajasi:

Ma’lumotlar to‘plami	Tasnif vazifasi	Rasm soni	Aniqlik
H-MT (Gistologiya)	Norma / Metastaz	100,000	0.97
X-NR3 (Rentgen)	3 ta yosh guruhi	550,080	0.83
CT (O‘pka KT)	Norma / Sil	149,248	0.96
H-ST (Gistologiya II)	6 xil reagent	267,984	0.95

Tajribalar shuni ko‘rsatdiki, CW algoritmi eng yuqori samaradorlikka ega bo‘lib, piksellarning minimal o‘zgarishida ham modelni 100% gacha xato qildira oladi.

- O‘tuvchanlik (Transferability): "Qora quti" sharoitida, bir model (masalan, DenseNet121) uchun yaratilgan hujum boshqa modelni (ResNet50) 36% gacha holatda chalg‘ita oldi.
- Ishonch va turg‘unlik: Model o‘z qaroriga qanchalik ishonchi baland bo‘lsa (confidence > 95%), uni chalg‘itish uchun shunchalik ko‘p shovqin talab etiladi.

Xulosa qilib aytganda tadqiqot shuni ko‘rsatdiki, zamonaviy tibbiy AI tizimlari raqobatli hujumlar oldida mo‘rt hisoblanadi. Tizimlarning xavfsizligini oshirish uchun quyidagilar tavsiya etiladi:

- Raqobatli o‘qitish (Adversarial Training): Modelni o‘qitish jarayonida ataylab buzilgan tasvirlardan foydalanish.
- Ko‘p bosqichli verifikatsiya: Diagnostikada bir nechta model ansamblidan foydalanish.
- Kirish filtrlarini joriy etish: Tasvirni neyron tarmog‘iga kirishidan oldin shovqinlardan tozalash (denoising).

FOYDALANILGAN ADABIYOTLAR RO‘YXATI:

1. Szegedy, C. Intriguing properties of neural networks [Matn] / C. Szegedy, W. Zaremba, I. Sutskever [va boshq.] // arXiv preprint arXiv:1312.6199. – 2013. – P. 1-10.

2. Goodfellow, I. Explaining and harnessing adversarial examples [Matn] / I. Goodfellow, J. Shlens, C. Szegedy // arXiv preprint arXiv:1412.6572v3. – 2015. – P. 1-11.

3. Moosavi-Dezfooli, S. M. DeepFool: a simple and accurate method to fool deep neural networks [Matn] / S. M. Moosavi-Dezfooli, A. Fawzi, P. Frossard // arXiv preprint arXiv:1511.04599v3. – 2015. – P. 1-9.

4. Madry, A. Towards deep learning models resistant to adversarial attacks [Matn] / A. Madry, A. Makelov, L. Schmidt [va boshq.] // arXiv preprint arXiv:1706.06083v3. – 2017. – P. 1-25.

5. Carlini, N. Towards evaluating the robustness of neural networks [Matn] / N. Carlini, D. Wagner // IEEE Symposium on Security and Privacy (SP). – 2017. – P. 39–57.