CHUQUR NEYRON TARMOQLARNING BIOTIBBIYOT TASVIRLARINI TAHLIL QILISHDAGI RAQOBATBARDOSH HUJUMLARGA CHIDAMLILIGINI O‘RGANISH

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 asosini tashkil etuvchi chuqur konvolyutsion neyron tarmoqlarining (CNN) raqobatli hujumlarga nisbatan zaif tomonlari o‘rganilgan. Tadqiqot davomida turli tuzilishga ega bo‘lgan modellar (jumladan, InceptionV3, ResNet50, DenseNet121) biotibbiyot ma’lumotlar to‘plamlari asosida sinovdan o‘tkazilgan. Maqolada PGD, DeepFool va CW kabi asosiy hujum usullarining samaradorligi qiyosiy tahlil qilingan hamda yuqori aniqlikdagi modellarning o‘zi ham inson ko‘zi sezmaydigan juda kichik o‘zgartirishlar (perturbatsiyalar) tufayli xato xulosalar chiqarishi isbotlangan. Xulosa qismida tibbiy sun’iy intellekt tizimlarining kiberxavfsizlikka chidamliligini oshirish uchun raqobatli o‘qitish va kirish filtrlari 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.*

ИЗУЧЕНИЕ УСТОЙЧИВОСТИ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ К КОНКУРЕНТНЫМ АТАКАМ 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 работе исследуются уязвимости глубоких сверточных нейронных сетей (CNN), лежащих в основе современных медицинских диагностических систем, к конкурентным атакам. В ходе исследования модели с различными структурами (включая InceptionV3, ResNet50, DenseNet121) тестировались на биомедицинских наборах данных. В работе сравнивается эффективность основных методов атак, таких как PGD, DeepFool и CW, и доказывается, что даже высокоточные модели могут делать ошибочные выводы из-за очень малых изменений (возмущений), незаметных для человеческого глаза. В*

заключении приводятся практические рекомендации, такие как внедрение конкурентного обучения и входных фильтров, для повышения кибербезопасности медицинских систем искусственного интеллекта.

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

STUDYING THE RESILIENCE OF DEEP NEURAL NETWORKS TO COMPETITIVE ATTACKS 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 studies the vulnerabilities of deep convolutional neural networks (CNNs), which are the basis of modern medical diagnostic systems, to competitive attacks. During the study, models with different structures (including InceptionV3, ResNet50, DenseNet121) were tested on biomedical datasets. The paper compares the effectiveness of the main attack methods such as PGD, DeepFool, and CW, and proves that even high-fidelity models can make erroneous conclusions due to very small changes (perturbations) that are not noticeable to the human eye. The conclusion provides practical recommendations, such as the introduction of competitive training and input filters, to increase the cybersecurity resilience of medical artificial intelligence systems.*

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

Chuqur neyron tarmoqlarning biotibbiyot tasvirlarini tahlil qilishdagi raqobatbardosh hujumlarga chidamliligini o‘rganish

Chuqur neyron tarmoqlari (ChNT) tibbiy tasvirlarni tahlil qilishda, jumladan, saraton, sil va boshqa kasalliklarni avtomatik diagnostika qilishda yuqori natijalarga erishmoqda [3, 4]. Biroq, so‘nggi tadqiqotlar shuni ko‘rsatdiki, bunday tarmoqlar raqobatbardosh hujumlar (adversarial attacks) ga nisbatan juda zaifdir: original tasvirga inson ko‘zi sezmaydigan darajada kichik o‘zgartirish kiritish orqali tarmoqni noto‘g‘ri klassifikatsiya qilishga majbur qilish mumkin [2, 6]. Bu muammo, ayniqsa, xatolik inson hayotiga xavf tug‘dirishi mumkin bo‘lgan biotibbiyot sohasida juda dolzarbdir. Shu sababli, raqobatbardosh hujumlarning xususiyatlarini va ularning turli rejimlarda (oq va qora quti) qanday ta’sir qilishini o‘rganish muhim ilmiy-texnik vazifadir.

Raqobatbardosh hujumlar bo‘yicha birinchi ishlardan biri Szegedy va boshq. [6, 13] tomonidan amalga oshirilgan bo‘lib, ular neyron tarmoqlarning kichik buzilishlarga nisbatan beqarorligini aniqlagan. Goodfellow va boshq. [7] “Fast Gradient Sign Method” (FGSM) algoritmini taklif qilib, hujumlarni samarali tushuntirgan. Madry va boshq. [8] PGD

(Projected Gradient Descent) algoritmini ishlab chiqib, uni eng kuchli hujumlardan biri sifatida taqdim etgan. Carlini va Wagner [12] (CW algoritmi) L-BFGS ga asoslangan yanada samarali, ammo hisoblash jihatidan og‘irroq usulni taklif qilgan. Moosavi-Dezfooli va boshq. [14] DeepFool algoritmi orqali minimal L_2 buzilish bilan hujum qilish mumkinligini ko‘rsatgan.

1-jadval

Ishlatilgan ma’lumotlar to‘plamlari va ular asosida qurilgan klassifikatsiya masalalari

Ma’lumotlar to‘plami	Qisqartma	Klassifikatsiya masalasi	Tasvirlar soni	Sinflar bo‘yicha taqsimot	Aniqlik
Gistologiya (metastazlar)	H-MT	Normal / Metastaz	100000	50000 / 50000	0.97
Gistologiya (tuxumdon/qalqonsimon bez)	H-OV	Tuxumdon: normal / o‘sma	96000	48000 / 48000	0.92
-	H-TH	Qalqonsimon bez: normal / o‘sma	96000	48000 / 48000	0.94
-	H-OV-TH	4 sinf (tuxumdon normal/o‘sma, qalqonsimon normal/o‘sma)	192000	48000×4	0.91
Rentgen (o‘pka)	X-NR2	2 yosh guruhi: 20-35 / 50-70 yosh	200000	100000 / 100000	0.98
-	X-NR3	3 yosh guruhi: 17-24 / 25-41 / 42-80 yosh	550080	183360×3	0.83
Kompyuter tomografiyasi (sil)	CT	Normal / Sil	149248	111990 / 37258	0.96
Gistologiya (6 kimyoviy agent)	H-ST	6 sinf (turli agentlar)	267984	59568, 37488, 55296, 35280, 24192, 56160	0.95

Adabiyotlarda asosan umumiy maqsadli tasvirlar (CIFAR-10, ImageNet) bo‘yicha tadqiqotlar olib borilgan. Biroq, biotibbiyot tasvirlari sohasida bunday hujumlarning ta’siri kam o‘rganilgan. Ushbu ishda aynan shu bo‘shliqni to‘ldirish maqsad qilingan.

Jami 5 xil biotibbiyot tasvirlari to‘plami ishlatilgan. Ularning qisqacha tavsifi 1-jadvalda keltirilgan.

Oq quti rejimida tajribalar uchun InceptionV3, DenseNet121, ResNet, MobileNet, Xception arxitekturalari ishlatilgan. Qora quti rejimida ham xuddi shu arxitekturalar “maqsadli tarmoq” va “imitatsion tarmoq” sifatida qo‘llanilgan.

Quyidagi uchta raqobatbardosh hujum algoritmi qo‘llanilgan:

- PGD (Projected Gradient Descent) [8]: Yo‘naltirilmagan hujum uchun iteratsiya qoidasi:

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

bu yerda m – original sinf indeksi, α – o‘rganish koeffitsienti, ϵ – buzilish chegarasi ($L\infty$ normasi bo‘yicha).

- DeepFool [14]: Minimal L_2 buzilishni topishga mo‘ljallangan iterativ usul. Har bir iteratsiyada funksiyani chiziqlilashtirish orqali proyeksiya hisoblanadi:

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

- Carlini & Wagner (CW) [12]: Cheklangan optimizatsiya masalasi:

$$\| x^* - x\| \to \min, \quad F(y(x^*)) = l, \quad x \in [0,1]^n$$

Oq quti rejimida hujumlar har bir klassifikatsiya masalasi uchun:

1. Neyron tarmoq o‘qitiladi.
2. Test to‘plamidagi har bir tasvir uchun PGD (ε = 0.02 dan 0.2 gacha, qadam 0.02), DeepFool va CW algoritmlari yordamida hujum tasvirlari yaratiladi.
3. Muvaffaqiyatli hujumlar ulushi (norma chegarasiga nisbatan) hisoblanadi.
4. PGD uchun iteratsiyalar soni (1 dan 100 gacha) va original tasvirning ehtimollik darajasi (0.5–1 oralig‘ida 10 bo‘lak) bo‘yicha bog‘liqlik o‘rganiladi.

Qora quti rejimida quyidagi metodika qo‘llanilgan:

1. Maqsadli tarmoqning o‘quv to‘plamidan foydalanib, imitatsion tarmoq o‘qitiladi.
2. Imitatsion tarmoqqa qarshi oq quti hujumi (PGD, ε=0.1) o‘tkazilib, hujum tasvirlari yaratiladi.
3. Yaratilgan hujum tasvirlari maqsadli tarmoqqa beriladi va muvaffaqiyat ulushi hisoblanadi.

Jami 4 ta klassifikatsiya masalasi (X-NR3, CT, H-OV-TH, H-ST) va 5 ta arxitektura uchun 25 ta juftlik (20 xil tarmoq + 5 bir xil tarmoq) tekshirilgan.

Oq quti rejimi natijalari:

- $L\infty$ normasi bo‘yicha (6.1-rasm): PGD algoritmi $\varepsilon \approx 0.1–0.15$ da muvaffaqiyatli hujumlar ulushi 80–95% ga yetadi. DeepFool va CW esa kichik buzilishlarda ham yuqori samaradorlik ko‘rsatadi.
- L_2 normasi bo‘yicha (6.2-rasm): DeepFool va CW o‘rtasida sezilarli farq yo‘q; ikkalasi ham o‘rtacha $L_2 \approx 1.5–2.0$ oralig‘ida muvaffaqiyatli hujumlar yaratadi.
- PGD iteratsiyalari soni (6.3-rasm): 8 ta masaladan 5 tasida 20–30 iteratsiyadan keyin muvaffaqiyat ulushi platosiga chiqadi; qolgan 3 tasida asta-sekin o‘sish kuzatilgan.
- Original tasvirning ehtimollik darajasi (6.4-rasm): Original tasvirga berilgan ehtimollik qancha yuqori bo‘lsa (masalan, >0.95), hujum muvaffaqiyati shuncha past bo‘lgan. Bu tasvir sinf ichida “chuqur” joylashganligidan dalolat beradi.

- L_2 norma va ehtimollik bog‘liqligi (6.5-rasm): Yuqori ehtimolli tasvirlar uchun DeepFool va CW ham kattaroq L_2 buzilish talab qiladi.

Qora quti rejimi natijalari

- Bir xil tarmoq uchun (oq quti) muvaffaqiyat ulushi 85–99% ni tashkil etadi.
- Turli tarmoqlar orasida hujum ko‘chirilganda (imitatsion → maqsadli) muvaffaqiyat keskin pasayadi: aksariyat hollarda 10–30% atrofida.
- Eng yaxshi ko‘chiruvchanlik X-NR3 (rentgen) va H-ST (gistologiya) to‘plamlarida kuzatilgan, ammo bu ham 50% dan oshmagan.
- CT (kompyuter tomografiyasi) va H-OV-TH to‘plamlarida ko‘chiruvchanlik deyarli nolga yaqin.

Bu xulosa shuni ko‘rsatadiki, oddiy PGD yordamida qora quti hujumlari biotibbiyot tasvirlari uchun unchalik samarali emas. Bunga sabab – turli arxitekturalarning xususiyat fazolaridagi farq va ma’lumotlarning murakkab tuzilishi.

Ushbu ishda biotibbiyot tasvirlarini tahlil qiluvchi chuqur neyron tarmoqlarga qarshi raqobatbardosh hujumlarning samaradorligi keng miqyosda o‘rganildi. Asosiy ilmiy va amaliy xulosalar:

- Muammoning dolzarbligi tasdiqlandi: barcha sinovdan o‘tkazilgan hujum algoritmlari neyron tarmoqlarning aniqligini 15% dan pastga tushirishga qodir. Demak, biotibbiyotda ChNTlarni qo‘llashda xavfsizlik choralari majburiy.
- Algoritmlarni taqqoslash: PGD algoritmi bir xil L_∞ buzilish sharoitida DeepFool va CW dan sezilarli darajada past samaradorlikka ega. L_2 normasiga ko‘ra DeepFool va CW deyarli bir xil sifatda hujum tasvirlarini yaratadi.
- Ehtimollik va mustahkamlik: Agar neyron tarmoq original tasvirga 0.95 dan yuqori ehtimollik bersa, bu tasvirga qarshi muvaffaqiyatli hujum yaratish qiyinroq yoki kattaroq buzilish talab etadi. Bu xususiyat kelajakda “hujumga chidamli” tarmoqlarni yaratishda qo‘llanishi mumkin.
- Qora quti hujumlari cheklangan: 4 ta klassifikatsiya masalasidan 3 tasida PGD asosidagi qora quti hujumlari deyarli samarasiz (muvaffaqiyat ulushi <30%). Bu shuni anglatadiki, real dasturlarda faqat oq quti hujumlaridan himoyalanish yetarli emas; qora quti hujumlariga qarshi ham maxsus usullar (masalan, tasvirlarni siqish, kirishni tekshirish [10]) ishlab chiqilishi kerak.
- Amaliy tavsiya: Ishlab chiqilgan metodika va natijalar avtomatik va yarim avtomatik diagnostika tizimlarida xavfsizlikni oshirish uchun qo‘llanilishi mumkin. Xususan, yuqori ishonchli (ehtimollik >0.95) prognozlarni qayta tekshirish yoki ularga qo‘shimcha filtrlar qo‘llash maqsadga muvofiq.

ADABIYOTLAR RO‘YXATI:

1. Recht B., Roelofs R., Schmidt L., Shankar V. Do CIFAR-10 classifiers generalize to CIFAR-10? // arXiv preprint. – 2018. – arXiv:1806.00451.

2. Akhtar N., Mian A.S. Threat of adversarial attacks on deep learning in computer vision // IEEE Access. – 2018. – Vol. 6. – P. 14410–14430.

3. Litjens G., Kooi T., Bejnordi B., Setio A., Ciompi F., Ghafoorian M. A survey on deep learning in medical image analysis // Medical Image Analysis. – 2017. – Vol. 42. – P. 60–88.

4. Ker J., Wang L., Rao J., Lim T. Deep learning applications in medical image analysis // IEEE Access. – 2018. – Vol. 6. – P. 9375–9389.

5. Wu Z., Lim S.-N., Davis L., Goldstein T. Making an invisibility cloak: real world adversarial attacks on object detectors // arXiv preprint. – 2019. – arXiv:1910.14667.

6. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // International Conference on Learning Representations (ICLR) 2014. – Banff: Springer, 2014. – P. 1–10.

7. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv preprint. – 2015. – arXiv:1412.6572v3.

8. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // arXiv preprint. – 2017. – arXiv:1706.06083v3.

9. Ozdag M. Adversarial attacks and defenses against deep neural networks: a survey // Procedia Computer Science. – 2018. – Vol. 140. – P. 152–161.

10. Xu W., Evans D., Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks // arXiv preprint. – 2017. – arXiv:1704.01155v2.

11. Wang H., Yu C.-N. A direct approach to robust deep learning using adversarial networks // arXiv preprint. – 2019. – arXiv:1905.09591v1.

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

13. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // arXiv preprint. – 2013. – arXiv:1312.6199.

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