ADVERSARIAL HUJUM USULLARI YORDAMIDA OBYEKTLARNI ANIQLASh MODELLARINI ChALG‘ITISh VA ULARNING BARQARORLIGINI BAHOLASh

Samatov Boburmirzo Xamidillo o‘g‘li – *o‘qituvchi*
boburmirzosamatov@gmail.com
Yunusov Odiljon Pozilovich – *dotsent*
Sayidova Nigora Komildjonovna – *katta o‘qituvchi*
Andijon davlat universiteti

ИССЛЕДОВАНИЕ ВОЗДЕЙСТВИЯ ADVERSARIAL АТАК НА МОДЕЛИ ОБНАРУЖЕНИЯ ОБЪЕКТОВ И ОЦЕНКА ИХ УСТОЙЧИВОСТИ

Саматов Бобурмирзо Хамидуллаевич – *преподаватель*
Юнусов Одилжон Позилович – *доцент*
Саидова Нигора Комилджоновна – *старший преподаватель*
Андижанский государственный университет

EVALUATION OF THE ROBUSTNESS OF OBJECT DETECTION MODELS AGAIUST ADVERSARIAL ATTACKS

Samatov Boburmirzo Khamidullaevich – *teacher*
Yunusov Odiljon Pozilovich – *Associate Professor*
Sayidova Nigora Komildjonovna – *senior lecturer*
Andijan State University

Tayanch so‘zlar: FGSM, YOLOv5, PGD, Faster R-CNN, ResNet18, CNN.

Rezyume. Ushbu tadqiqotda konvolyutsion neyron tarmoqlarga asoslangan obyekt aniqlash tizimlari, xususan, YOLOv5 modeliga qarshi adversarial hujumlar samaradorligi o‘rganildi. Hujum usullari sifatida Fast Gradient Sign Method (FGSM) hamda uning takomillashtirilgan variantlari (Iterative FGSM va PGD) qo‘llanildi. ResNet18 modeli yordamida hosil qilingan adversarial perturbatsiyalar asosida original rasm va hujumga uchragan rasmning obyekt aniqlash natijalari solishtirildi. Natijalar shuni ko‘rsatdiki, hatto kichik miqdordagi shovqin ham obyekt aniqlash modelini sezilarli darajada chalg‘itishi mumkin. Adversarial hujumlar sun’iy intellekt modellarining, ayniqsa chuqur o‘rganishga asoslangan tasvirni tanish tizimlarining asosiy zaifliklaridan biri hisoblanadi. FGSM va PGD kabi hujumlar model kirishini kichik perturbatsiyalar orqali o‘zgartirib, modelni noto‘g‘ri klassifikatsiya qilishga majbur qiladi.

Ключевые слова: FGSM, YOLOv5, PGD, Faster R-CNN, ResNet18, CNN.

Резюме. В данном исследовании были изучены системы обнаружения объектов, основанные на сверточных нейронных сетях, в частности эффективность адверсариальных атак против модели YOLOv5. В качестве методов атаки были применены Fast Gradient Sign Method (FGSM) и его усовершенствованные варианты (Iterative FGSM и PGD). Адверсариальные возмущения, сгенерированные с использованием модели ResNet18, применялись для сравнения результатов обнаружения объектов на исходном изображении и на изображении, подвергшемся атаке. Результаты показали, что даже небольшое количество шума может существенно ввести в заблуждение модель обнаружения объектов. Адверсариальные атаки являются одной из основных уязвимостей моделей искусственного интеллекта, особенно систем распознавания изображений, основанных на глубоком обучении. Такие атаки, как FGSM и PGD, изменяют входные данные модели с помощью небольших возмущений, заставляя модель выполнять некорректную классификацию.

Key words: FGSM, YOLOv5, PGD, Faster R-CNN, ResNet18, CNN.

Summary. In this study, object detection systems based on convolutional neural networks were investigated, particularly the effectiveness of adversarial attacks against the YOLOv5 model. As attack methods, the Fast Gradient Sign Method (FGSM) and its improved variants (Iterative FGSM and PGD) were applied. Adversarial perturbations generated using the ResNet18 model were used to compare the object detection results of the original image and the attacked image. The results demonstrated that even a small amount of noise can significantly mislead an object detection model. Adversarial attacks represent one of the major vulnerabilities of artificial intelligence models, especially deep learning–based image recognition systems. Attacks such as FGSM and PGD modify the model input with small perturbations, forcing the model to produce incorrect classifications.

Kirish. So‘nggi yillarda chuqur o‘rganish (Deep Learning) asosidagi obyekt aniqlash tizimlari kompyuter ko‘rish sohasida yuqori samaradorlikka erishdi. Ayniqsa, konvolyutsion neyron tarmoqlarga (CNN) asoslangan arxitekturalar, jumladan YOLOv5 real vaqt rejimida yuqori aniqlik va tezkorlikni ta’minlay olishi bilan ajralib turadi. Bunday modellar video kuzatuv, sanoat nazorati, biometrik autentifikatsiya hamda avtonom transport tizimlarida keng qo‘llanilmoqda. Ushbu sohalarda modelning ishonchliligi va barqarorligi muhim ahamiyat kasb etadi.

Shu bilan birga, chuqur o‘rganish modellarining adversarial hujumlarga nisbatan zaifligi dolzarb muammolardan biri hisoblanadi. Adversarial hujum – bu model kirish ma’lumotlariga juda kichik, inson ko‘zi bilan deyarli sezilmaydigan perturbatsiya qo‘shish orqali modelni noto‘g‘ri qaror qabul qilishga majbur qilish usulidir. Amaliy jihatdan bu holat obyektni noto‘g‘ri aniqlash, klassifikatsiya xatolari yoki aniqlash ishonchliligining keskin pasayishiga olib kelishi mumkin. Xususan, gradient asosidagi hujum usullari, masalan, Fast Gradient Sign Method (FGSM) va Projected Gradient Descent (PGD)

model parametrlarining sezgirligidan foydalanib, samarali chalg‘ituvchi shovqinlar hosil qiladi.

Obyekt aniqlash tizimlarida adversarial tahdidlarning mavjudligi real tizimlar xavfsizligiga bevosita ta’sir ko‘rsatadi. Masalan, kuzatuv tizimida inson obyektining aniqlanmasligi yoki noto‘g‘ri aniqlanishi jiddiy xavfsizlik muammolariga sabab bo‘lishi mumkin. Shuning uchun obyekt aniqlash modellarining adversarial sharoitdagi barqarorligini eksperimental baholash va ularning zaif tomonlarini aniqlash muhim ilmiy va amaliy vazifa hisoblanadi.

Mazkur tadqiqotning maqsadi – konvolyutsion neyron tarmoqlarga asoslangan obyekt aniqlash modeli, xususan YOLOv5 ning adversarial hujumlarga nisbatan barqarorligini baholashdir. Tadqiqotda adversarial perturbatsiyalar ResNet-18 modeli yordamida generatsiya qilinib, tasvirlarga FGSM va PGD usullari orqali qo‘llaniladi hamda original va hujumga uchragan tasvirlar natijalari o‘zaro solishtiriladi.

Adabiyotlar tahlili va metodologiya. Mazkur tadqiqotda obyekt aniqlash modeli sifatida YOLOv5 tanlandi. Ushbu model real vaqt rejimida yuqori aniqlik va tezlik ko‘rsatkichlariga ega bo‘lgan bir bosqichli (one-stage) detektor hisoblanadi. Adversarial perturbatsiyalarni generatsiya qilish uchun esa tasvirlarni klassifikatsiya qiluvchi chuqur konvolyutsion neyron tarmoq ResNet-18 modelidan foydalanildi. Model ImageNet ma’lumotlar bazasida oldindan o‘qitilgan (pre-trained) vaziyatda ishlatildi.

Tadqiqot arxitekturasi ikki bosqichdan iborat:

1. Tasvirga adversarial shovqin generatsiya qilish.
2. Olingan adversarial tasvirni obyekt aniqlash modelida sinovdan o‘tkazish va natijalarni solishtirish.

Ushbu hujumni amalga oshirish quyidagi natijalar orqali tekshirildi va amalga oshirildi.

ResNet18 modelidan foydalanib FGSM yordamida tasvirga adversarial shovqin qo‘shiladi. Natijada modelning oldingi to‘g‘ri prediksiyasi o‘zgaradi. Maqsad – hujumning samaradorligini baholash va himoya mexanizmlarini taklif qilish.

Tadqiqotda ResNet18 (ImageNet) – 1000 sinfli pre-trained CNN modeli ishlatiladi. Modelning birinchi prediksiyasi real tasvir bo‘yicha olinadi.

FGSM quyidagi formula asosida ishlaydi:

$$x_{adv} = x + \epsilon * sign(\nabla_x J(\theta, x, y))$$

Bu yerda:

x – original rasm,

ϵ – shovqin miqdori,

$\nabla_x J$ – yo‘qotish funksiyasi gradienti,

$sign(\,)$ – belgini qaytaruvchi funksiya.

Dastur ishlash jarayoni quyidagi tartibda amalga oshirildi:

1. ResNet18 model yuklanadi (1-rasm).

```
model = models.resnet18(weights=models.ResNet18_Weights.IMAGENET1K_V1)
model.eval()
```

1-rasm. ResNet18 model yuklanishi

2. Rasm o‘lchami 224x224 ga keltiriladi.
3. Modelga berilib normal prediksiya olinadi.
4. Loss funksiyasi orqali gradient hisoblanadi.
5. FGSM qo‘llanib shovqinli rasm (adv_image) yaratiladi (2-rasm).

```
epsilon = 0.06
adv_image = img_tensor + epsilon * img_tensor.grad.sign()
adv_image = torch.clamp(adv_image, min: 0, max: 1)
```

2-rasm. FGSM qo’llanilishi

6. Adversarial rasm qayta modelga beriladi(3-rasm).

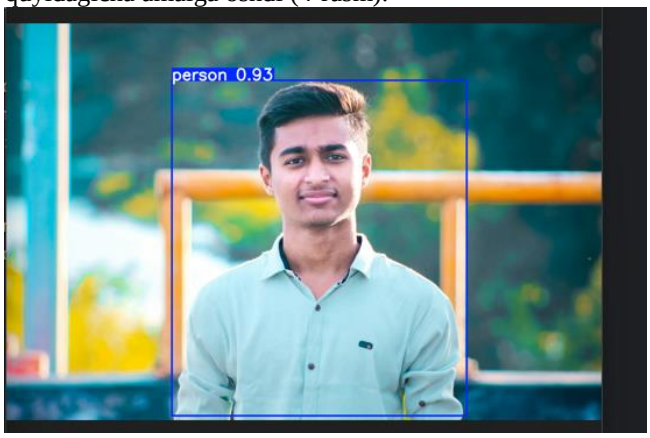
```
yolo_model = torch.hub.load( repo_or_dir: 'ultralytics/yolov5', model: 'yolov5s', pretrained=True, skip_validation=True)
results_orig = yolo_model(img)
results_orig.show()
results_adv = yolo_model(adv_img_pil)
results_adv.show()
```

3-rasm. Yolo model va undan natija olinishi

7. Normal va hujumdan 39eying natijalar solishtiriladi

Muhokama va natijalar.

Natijalarni vizual ko‘rinishda dastur natijalari quyidagicha amalga oshdi (4-rasm).



4-rasm. Asl tasvirning YOLO natijasi

Tasvirda ko‘rinib turibdiki bu yerda YOLO insonni aniqlayapti va 93% aniqlikda. Adversarial rasmni huddi shu modelda aniqlaydigan bo’lsak 30% inson ekanligini korsatyapti (5-rasm).

5-rasm. Adversarial o‘zgartirilgan rasm

Adversarial hujumlar real tizimlar uchun xavf tug‘diradi. Modellarni himoyalash uchun ko‘p bosqichli yondashuv — eng samarali usul hisoblanadi.

Ushbu ishda tasvirlarni klassifikatsiya qiluvchi ResNet18 modeli ustida FGSM (Fast Gradient Sign Method) asosida adversarial hujum amalga oshirildi. Rasmga inson ko‘zi tomonidan sezilmaydigan, ammo model uchun ahamiyatli bo‘lgan gradient yo‘nalishidagi

kichik shovqin qo‘shish orqali modelning prediksiyasi o‘zgartirib yuborilishi kuzatildi. Bu usul orqali yaratilgan adversarial rasm model tomonidan noto‘g‘ri klass tanlanishiga olib keldi, bu esa chuqur o‘rganish modellarining adversarial ta’sirlarga juda sezgirligini ko‘rsatdi.

Natijalar shuni ko‘rsatdiki, hatto juda kichik o‘zgarishlar (epsilon = 0.06) ham modelni chalg‘itib, to‘g‘ri klassifikatsiya qila olmaydigan holatga olib keladi. Bu esa real hayotda video kuzatuv, biometrik xavfsizlik, avtopilot tizimlari va tibbiy diagnostika kabi ko‘plab sohalarda chuqur o‘rganish modellarining ishonchliligi va xavfsizligi bilan bog‘liq jiddiy muammolarni keltirib chiqaradi.

Adversarial hujumlarning asosiy sababi – neyron tarmoqlarning kirish ma’lumotlaridagi past darajadagi matematik o‘zgarishlarga sezgirligi va ularni «inson ko‘rmaydigan» shaklda o‘zgartirish mumkinligidir. Ushbu tajribada modelni alday olgan shovqinlar tasvirning vizual sifatini sezilarli darajada o‘zgartirmagan bo‘lsa-da, u modelning qaror qabul qilish jarayoniga kuchli ta’sir ko‘rsatgan.

Adversarial hujumlarning oldini olish va modelni mustahkamlash uchun quyidagi himoya choralaridan foydalanish tavsiya etiladi:

1. Adversarial Training (eng samarali usul).

Modelni o‘qitishda adversarial rasm va original rasmni birgalikda qo‘llash.

Model hujumga barqaror bo‘lib qoladi.

2. Input Transformation Defense.

Kirish tasvirini o‘zgartirib, shovqinni yo‘qotish:

- Gaussian blur
- Median filtering
- JPEG compression
- Bit-depth reduction (feature squeezing)
- Random resize + padding

Bu usullar adversarial signallarni zaiflashtiradi.

3. Gradient Masking / Obfuscation.

Modelning gradientini yashirish yoki buzish orqali hujumchi to‘g‘ri yo‘nalishni topolmaydi.

Bir nechta modelni birgalikda ishlatish – bitta modelga qarshi bajarilgan hujum boshqasiga ta’sir qilmay qolishi mumkin.

Xulosa. Ushbu tadqiqot shuni ko‘rsatdiki, chuqur o‘rganish modellarida adversarial tahdidlar mavjud bo‘lib, ular tizim ishonchliligini sezilarli darajada kamaytiradi. Shu sababli har qanday real tizimda AI modellarini qo‘llashda adversarial hujumlardan himoya qilish bo‘yicha chora-tadbirlarni joriy etish juda muhimdir. Himoya choralarini qo‘llash modelning ishonchliligini oshiradi va uni turli xavfsizlik xatarlaridan himoya qiladi.

Adabiyotlar

1. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. // arXiv preprint arXiv:1412.6572. 2014. https://arxiv.org/abs/1412.6572

2. Szegedy C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks. // arXiv preprint arXiv:1312.6199. 2013. https://arxiv.org/abs/1312.6199

3. Kurakin A., Goodfellow I., Bengio, S. Adversarial machine learning at scale. // arXiv preprint arXiv:1611.01236. 2016. https://arxiv.org/abs/1611.01236

4. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. Proceedings of the 6th International Conference on Learning Representations (ICLR). 2018.

5. Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z.B., Swami A. Practical black-box attacks against deep learning systems using adversarial examples. // IEEE Symposium on Security and Privacy (SP), 2017. – P. 1-18.