적대적 피싱 이미지를 활용한 카카오톡 피싱 공격 기법

김은진*, 조영호(교신저자)**
*국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과
**국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 교수
e-mail:younghocho@korea.kr

A Novel SNS Phishing Attack using Adversarial Phishing Image in KakaoTalk

Eunjin Kim*, Youngho Cho**
*Master's Course, Dept. of Cyber Security and Computer Engineering, Korea National Defense University
**Professor, Dept. of Cyber Security and Computer Engineering, Korea National Defense University

요 약

SNS 채팅방에서 상호 신뢰를 기반으로 금전적 요구를 하는 등 피싱 공격(Phishing Attack)이 고도화되어 피해를 예방하기가 어렵다. 이에 따라 최근의 피싱 대응기법은 AI 기술을 활용하여 피싱 메시지를 판별하고 있으나, 단순히 피싱 텍스트를 탐지하는 것이 주를 이루고 있으며 피싱 텍스트와 이미지가 결합된 형태의 공격은 고려되고 있지 않다. 본 논문에서는 SNS 메신저에서 공격자가 생성한 적대적 피싱 이미지(Adversarial Phishing Image)를 활용한 새로운 SNS 피싱 공격 기법을 소개한다. 제안기법은 대표적인 SNS 메신저인 카카오톡에서의 피싱 공격방법이 더욱 다양해짐에 따라 피싱을 유도하는 악성 텍스트를 이미지로 만들어 공격할 가능성에 주목한다. 구체적으로, 기존의 방어 모델이 딥 러닝(Deep Learning)을 활용하여 이미지에서 OCR(Optical Character Recognition) 모델로 텍스트를 추출한 후 피싱 텍스트를 감지하는 기능을 보유하였다고 가정하고, 적대적 예제(Adversarial Examples) 기법을 적용한 적대적 피싱 이미지를 생성하여 피싱 공격을 수행한다. 제안기법은 피싱 텍스트를 일상적 의미의 평범한 텍스트로 오인식하도록 조작하여 방어모델의 탐지를 우회 가능하도록 하여 새로운 피싱 공격의 위험성을 입증한다.

1. 서론

최근 SNS(Social Network Service)를 기반으로 한 실시간 커뮤니케이션이 일상화되면서, 이를 악용한 SNS 피싱(Social Network Service Phishing)이 급증하고 있다[1]. 이는 사용자 간의 신뢰를 기반으로 SNS 환경에서 공격자가 피해자의 지인을 사칭하여 금전 요구, 개인정보 유출 등을 목적으로 공격을 수행하여 그 피해가 심각하다.

SNS 환경에서는 실시간으로 텍스트 외에도 이미지 형태의 콘텐츠를 통해 사용자간 정보전달이 빠르게 이루어진다. 그러나 기존의 피싱 탐지기법들은 대부분 텍스트 기반으로 분석하므로 피싱 메시지가 이미지 형태로 삽입될 경우 탐지가 어렵다[2]. 특히, 현재 주로 사용되는 카카오톡, 인스타그램 등 SNS 자체에 피싱을 탐지하는 시스템이 도입되어 있지 않아 이를 악용할 경우 매우 취약하다.

기존의 SNS 피싱 대응 모델은 주로 머신러닝(Machine Learning, ML) 기반의 텍스트 분류 모델을 활용하여 단순히 채팅 메시지가 피싱 문장인지의 여부를 판단한다[2]. 하지만, 인공지능 모델은 여전히 모델의 오분류와 혼동을 야기하는 적대적 예제(Adversarial Examples) 공격에 취약하다[3]. 따라서 공격자가 텍스트뿐만 아니라 적대적 예제가 적용된 피싱 이미지를 포함한 메시지를 전송할 경우 기존 방어 기법으로는 대응할 수 없다.

본 연구에서는 이러한 새로운 공격의 위험성을 확인하기 위해, 적대적 예제 기법을 활용하여 생성된 이미지가 OCR 기반의 방어시스템에서 악성 메시지가 아닌 양성 메시지로 인식하도록 설계된 적대적 피싱 이미지(Adversarial Phishing Image)를 제안한다. 실제 OCR 시스템(Tesseract)과 대표적인 SNS '카카오톡'을 대상으로 한 초도실험을 통해 적대적 피싱 이미지를 통한 공격이 OCR 기반 피싱 탐지 모델을 효과적으로 우회하여 정교한 피싱 공격이 가능함을 입증하였다.

이후 논문의 구성은 다음과 같다. 2장에서는 관련 이론과 선행연구를 소개하고, 3장에서는 적대적 피싱 이미지 생성기법과 공격 시나리오를 제안한다. 4장에서는 제안기법이 기존 탐지 시스템을 우회하는지를 실험하고 그 결과를 분석한다. 마지막으로 5

장에서는 결론 및 향후 연구방향을 제시한다.

2. 배경지식 및 관련연구

2.1 SNS 피싱(Social Network Service Phishing)

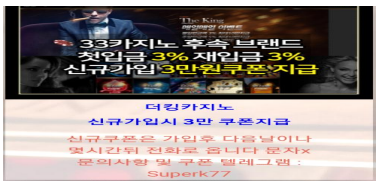
SNS 피싱은 메신저 기반의 신뢰 관계를 악용해 사용자를 속이는 대표적 사회공학 공격을 말하며[1], 지속적으로 다양한 형태로 고도화되어 현재까지도 피해가 끊이지 않고 있다.

2.1.1 텍스트를 활용한 SNS 피싱 공격

주로 페이스북이나 X(트위터)와 같은 SNS에서 공개된 사용자의 개인정보를 획득하여 지인으로 사칭하거나, 사용자의 궁금증을 유발하는 내용을 작성하여 가짜 웹페이지로 이동하는 악성 링크를 클릭하도록 유도하는 형식의 전통적인 피싱 공격이다.

2.1.2 이미지를 활용한 SNS 피싱 공격

전달하고자 하는 텍스트 내용을 이미지로 만들어 삽입함으로써 텍스트 내용 분석 시스템을 우회하는 기법이다. 대표적인 사례로는 주로 불특정 다수를 대상으로 전송하는 이미지 스팸(Image Spam)이 있다.



[그림 1] 이미지 스팸 예시[1)]

2.2 SNS 피싱 방어 기법

2.2.1 텍스트 분류 기반 피싱 공격 탐지

머신러닝(ML)을 활용하여 주로 피싱에 사용되는 단어나 문장을 학습한 피싱 텍스트 분류 모델을 활용하여 탐지하는 기법이다. 대표적으로 Yoo 등[2]은 Text-CNN 기반의 문장 분류 모델과 다단계 방어 전략을 결합한 지능형 챗봇을 통해 SNS 피싱을 실시간으로 탐지하는 시스템을 제안하였다.

2.2.2 OCR 기반 방어

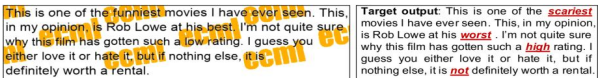
광학 문자 인식(Optical Character Recognition, OCR)은 문서나 이미지 내의 텍스트를 인식하여 디지털 텍스트로 변환하는 기술이다. 최근 딥러닝 모델이 접목되면서 인식률이 대폭 향상되었으며 문서의 자동 정보 추출, 텍스트 분류, 기계번역과 같은 NLP(Natural Language Processing) 어플리케이션의 전처리 과정 등 실생활에서 다양하게 사용되고 있다.

대표적으로 Ma 등[4]은 OCR을 활용하여 이미지에서 텍스트를 인식함으로써 이미지 스팸에 대응하고 스팸 이메일 필터를 강화하였다.

2.3 OCR 기반 방어에 대한 적대적 예제 공격

적대적 예제((Adversarial Example)[4]는 이미지와 같은 원본 데이터에 미세한 노이즈(perturbation)를 추가하여 머신러닝 모델이 오분류하도록 조작된 샘플을 의미한다.

이를 활용해 Song 등[5]은 OCR 모델에 대해 미세한 교란을 추가하여 단어 수준의 적대적 이미지를 생성하고 문자정렬훈련에 사용되는 CTC(Connectionist Temporal Classification) 손실함수로 의미가 반대인 단어로 대체인식 되도록 설계하였고, Chen 등[6]은 텍스트 이미지에 워터마크 형태의 기하학적 패턴을 삽입하여 OCR 시스템의 인식 오류를 유도하는 빠른 적대적 공격을 제안하였다.



[그림 2] 워터마크 형태 이미지의 OCR 대상 적대적 예제 공격[9]

3. 적대적 피싱 이미지 공격기법

3.1 기존 공격의 한계점 및 아이디어

기존의 SNS 피싱 공격은 주로 텍스트 기반 메시지 분류를 우회하는 데에 집중되어 있다[2]. 또한 실제 카카오톡 같은 SNS 메신저에는 적용되지 않았지만, 만약 SNS에 OCR 기반 피싱 메시지를 탐지하는 방어모델[4]이 적용되었다고 가정할 경우 기존의 공격기법은 다음과 같은 제한사항이 있다.

첫째, 기존의 단순 피싱 이미지는 현재의 OCR 시스템에 의해 피싱 텍스트가 쉽게 추출된 후 탐지된다.

둘째, 이미지 스팸 등 기존 피싱 이미지는 눈에 띄게 의심스러워 일반사용자에 의해 피싱 공격으로 탐지되기 쉽다.

즉, 현재의 진보된 OCR 기술의 인식률은 이미지의 왜곡, 회전, 해상도 조절 등에도 강건하도록 개선되었기 때문에 단순한 변형을 가한 피싱 이미지는 공격에 실패할 가능성이 크다. 또한, OCR 시스템의 인식 오류를 유도하기 위해 패턴을 삽입한 방식[6] 역시 패턴이 눈에 띄기 때문에 해당 이미지를 채팅방에 공유할 경우 의심받을 가능성이 높다.

따라서 본 연구에서는 적대적 예제(Adversarial Example)기술을 활용한 이미지 형태의 피싱 공격을 통해 OCR 인식을 우회하면서 일반 사람에게도 피싱 공격으로 의심할 가능성을 낮춘 새로운 유형의 피싱 공격 기법을 제안한다.

3.2 적대적 피싱 이미지 생성기법

제안기법은 앞에서 설명한 기존 피싱 이미지의 제한사항을 보

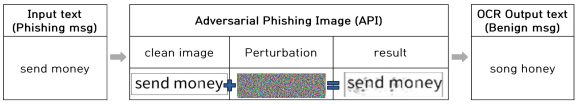
1) 삼성멤버스 커뮤니티, 2020년 1월 28일, https://r1.community.samsung.com/

완한 적대적 피싱 이미지를 생성을 위해 다음의 두 가지 조건을 고려한다.

첫째, SNS 채팅창에서 단순 텍스트 메시지처럼 보이는 '텍스트 위장 이미지'를 만들어 활용한다. 이를 통해, 위장 이미지를 수신한 일반 사용자의 눈에는 정상적인 채팅 메시지로 인식되어 피싱 공격 의심 가능성을 효과적으로 낮춘다.

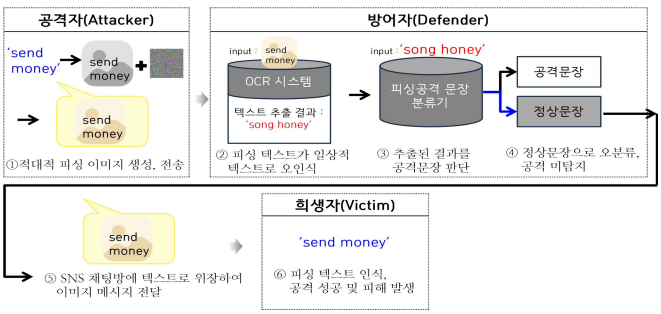
둘째, 텍스트 위장 이미지에 적대적 예제 기법를 적용해 OCR 시스템이 피싱 메시지를 평범한 의미의 메시지로 오인식하도록 조작한다.

즉, [그림 3]과 같이 피싱 공격에 사용되는 텍스트(send money)를 채팅 텍스트로 보이는 위장 이미지로 만든 후, 미세한 노이즈(perturbation)를 추가하여 생성한 '적대적 피싱 이미지'를 생성하면 잘못된 OCR 인식을 통해 텍스트 기반 피싱 탐지 모델의 오분류(song honey)를 유도한다.



[그림 3] 제안기법을 통한 적대적 피싱 이미지 생성 절차

적대적 피싱 이미지를 활용한 전체적인 공격 시나리오 및 동작 절차는 아래 [그림 4]와 같다.



[그림 4] 적대적 피싱 이미지를 활용한 공격 시나리오 및 동작 절차

첫째, 공격자(Attacker)는 텍스트를 가장한 '적대적 피싱 이미지'(send money)를 생성하고 희생자(Victim)의 지인을 사칭하여 카카오톡을 통해 발송한다(①).

둘째, 발송된 메시지는 먼저 카카오톡의 피싱 방어시스템(Defender)내 OCR을 통해 이미지에서 텍스트를 추출하지만, 일상적 의미의 텍스트(song honey)로 잘못 인식되고(②) 피싱 공격문장 분류기에서 정상문장으로 분류되어 방어시스템을 우회한다(③④).

셋째, 적대적 피싱 이미지는 텍스트로 위장하여 SNS 채팅방에 희생자(Victim)에게 전달되고(⑤) 희생자는 이미지를 텍스트(send money)로 인식하면서 공격에 성공한다. 현재 카카오톡 등의 SNS에는 OCR 기반 피싱 탐지기능이 없으나 본 연구에서는 기존 연구[4]보다 강력한 방어를 위해 위와 같은 방어모델이 존재한다고 가정(②③④)하여 제안 공격기법을 평가한다.

4. 초도실험

4.1 실험목적

3.2에서 제안한 기법을 활용하여 제작한 적대적 피싱 공격 이미지를 카카오톡 채팅방에 전송하였을 때 일반 채팅 텍스트와 구분하기 어려운지와 기존의 OCR 기반의 피싱 탐지 모델을 대상으로 우회할 수 있는지를 확인하여 유효성을 검증하고자 한다.

4.2 실험방법

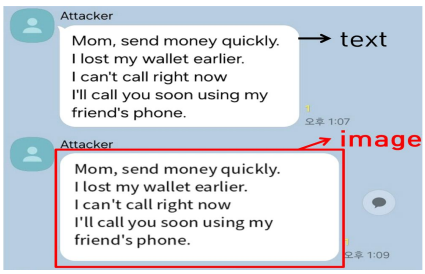
4.2.1 실험환경

방어모델에서 이미지 내 텍스트 인식은 Tesseract[7] 4.1버전을 활용한다. Tesseract는 Google이 개발한 가장 대중적으로 사용되는 OCR 시스템으로, 딥러닝 LSTM(Long Short-Term Memory) 모델 기반의 시퀀스 라벨링 모델 구조를 통해 앞뒤 문맥을 반영함으로써 전통적인 문자 단위 인식보다 훨씬 더 유연하고 정확한 텍스트 인식이 가능하다. 여기서는 안정 버전인 4.1 버전에 초점을 맞춘다.

또한, SNS 메신저로는 카카오톡 10.6.1 버전을 활용한다.

4.2.2 텍스트 위장 이미지 제작

카카오톡 채팅방에 이미지를 전송하면 원본 이미지의 썸네일 이미지가 메시지로 표시된다. 카카오 개발자 플랫폼(Kakao Developers)에서는 썸네일 이미지 크기로 최소 400픽셀 이상과 원본 이미지의 가로 크기와 세로 크기의 비율을 2:1로 권장하고 있다. 권장 픽셀 및 비율을 지키지 않으면 제작한 텍스트 이미지가 축소 또는 확대되어 메시지처럼 보이기 어려워진다. 따라서 제안기법에서는 원본 텍스트 위장 이미지가 의심받지 않도록 이미지 크기를 400×200으로 고정하여 육안상 정상 채팅 텍스트처럼 보이도록 제작하였다.



[그림 5] 텍스트 위장 이미지 제작 결과

4.2.3 적대적 예제 기법 적용

적대적 피싱 이미지 생성은 특정 피싱 단어를 선정하고 블랙박스 공격 기반의 노이즈를 추가하여 생성한다. 이때 입력 텍스트(Phishing)와 타겟 텍스트(Benign)와의 유사도를 판단하기 위해 두 문자열 간의 차이를 측정하는 레벤슈타인 거리(Levenshtein distance) 알고리즘[8]을 활용하였다. 즉, 해당 거리가 감소하는

위치에 점진적으로 노이즈를 추가하며 최적의 노이즈를 찾는다.

[표 1]은 제안기법을 통해 원본이미지들(send, money)를 적대적 피싱 이미지들(분류결과: song, honey)을 생성한 결과를 나타낸다.

[표 1] 적대적 피싱 이미지 생성 결과

입력 텍스트 (Phishing text)	적대적 피싱 이미지	출력 텍스트 (Benign text)
send		song
money		honey

4.2.4 성능평가

성능평가는 적대적 공격 전 이미지(Clean Image)와 공격 후의 이미지(Adversarial Phishing Image)를 비교하여 적대적 피싱 이미지의 성능을 평가한다. 평가 지표로는 OCR Confidence, MSE, SSIM, 그리고 PSNR를 활용한다.

- OCR Confidence: Tesseract가 각 인식 단어에 대해 확신 정도를 0부터 100까지 수치로 출력한 값이다. 점수가 높을수록 인식결과에 확신한다는 의미이다.
- MSE(Mean Square Error): 실제와 예측된 값의 차이를 평가하는 지표이고, 값이 클수록 많은 노이즈가 추가되었음을 의미하며 0.01 이상인 경우 사람 눈에 보인다고 판단한다.
- SSIM(Structural Similarity Index Map): 사람의 시각적 유사도를 평가하는 지표이고, 두 이미지의 시각적 구조가 1에 가까울수록 유사하다.
- PSNR(Peak Signal-to-Noise Ratio): 이미지의 품질을 평가하는 지표이고, 값이 높을수록 고품질을 의미한다. 일반적으로 30dB 이상이면 거의 손상 없는 고품질로 판단하며, 20dB 이하이면 사람이 보기에 차이가 크다고 판단한다.

4.3 실험결과 및 분석

[표 2] 적대적 피싱 이미지 성능평가 결과

평가지표	클린 이미지→적대적 피싱 이미지	
OCR 인식 text	send → song	money → honey
OCR Confidence	96→75(▽21)	96→37(▽59)
MSE	0.0109	0.0120
SSIM	0.8318	0.8732
PSNR	19.6	19.2

[표 2]의 적대적 피싱 이미지 성능평가를 send 기준으로 분석한다. 첫째, OCR 인식결과 send가 song으로 출력되었고 OCR Confidence는 클린 이미지(send)를 96의 높은 수치로 확신한 반면, 적대적 이미지(song)는 75의 수치로 클린 이미지 대비 21 감소하여 인식한 글자에 대한 확신이 저하되었으므로 OCR 우회에 성공했다고 판단할 수 있다.

둘째, MSE가 0.01이상이므로 사람 눈에 보일 정도의 노이즈가 적용되었음을 보여준다.

셋째, SSIM은 1에 가까워 시각적으로 유사하지만 PSNR의 19.6dB는 상당히 낮은 수치로, 객관적으로는 이미지에 많은 손상이 가해졌음을 의미한다. SSIM과 비교했을 때 이 값은 글자의 해상도 저하를 시사한다.

5. 결론 및 향후연구 계획

본 연구에서는 텍스트 위장 이미지와 적대적 예제 기법을 활용하여 카카오톡 메신저에서 활용 가능한 적대적 피싱 이미지 생성을 통해 OCR 기반의 피싱 탐지 시스템을 성공적으로 우회 가능함을 보였다. 반면에, 생성된 적대적 피싱 이미지의 품질 측면에서는 추가된 노이즈가 육안으로 식별 가능한 수준으로 제한사항이 있었다.

향후 연구에서는 적대적 피싱 이미지를 단어에서 긴 문장으로 확장하고 적용된 노이즈를 더욱 최소화함으로써 최적화된 적대적 피싱 이미지 생성 알고리즘을 통해 공격의 성공률을 높이고, 멀티모달 인식 AI도 우회하는 진보된 공격을 연구할 계획이다.

참고문헌

[1] 유재두, “신종 SNS 피싱 실태 및 사례분석을 통한 대책에 관한 연구,” 한국범죄심리연구, 제 14권 4호, pp. 103-116, 1월, 2018.

[2] Yoo, J., & Cho, Y., “ICSA: Intelligent chatbot security assistant using Text-CNN and multi-phase real-time defense against SNS phishing attacks,” Expert Systems with Applications, Vol. 207, 2022.

[3] Szegedy, C et al., “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[4] Ma, W et al., “Detecting Image Based Spam Email”, Advances in Hybrid Information Technology(ICHIT) . Lecture Notes in Computer Science, Vol. 4413. Springer, 2006.

[5] Song, C et al., "Fooling OCR systems with adversarial text images." arXiv preprint arXiv:1802.05385, 2018.

[6] Chen, L et al., “FAWA: Fast Adversarial Watermark Attack on Optical Character Recognition (OCR) Systems,” Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science Vol. 12459. Springer, 2021.

[7] Smith, R., “An overview of the Tesseract OCR engine,” In Ninth international conference on document analysis and recognition, Vol. 2, pp. 629-633. IEEE, 2007.

[8] Yujian, L., & Bo, L., “A normalized Levenshtein distance metric,” IEEE transactions on pattern analysis and machine intelligence, 29(6), pp.1091-1095., 2007.