안드로이드 AVN 시스템에서 머신러닝을 활용한 파일 파편의 유형 분류 기법 비교분석

김다은, 강해인, 조성제, 김홍근, 황영섭

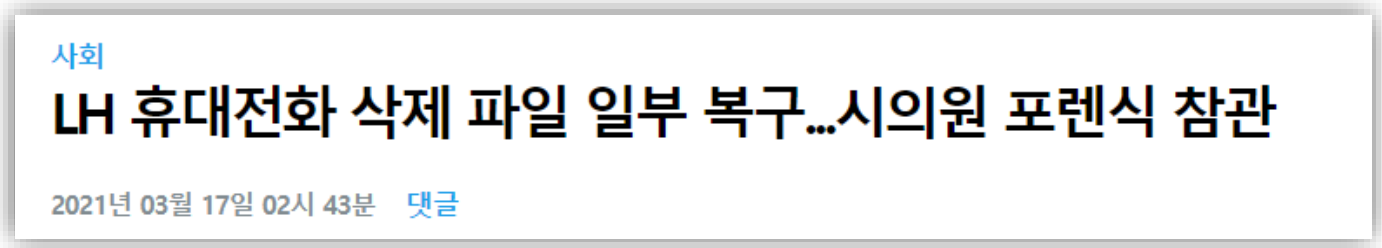
WDSC 2023

2023.08.17

INDEX

01 서론 & 배경지식

서론


사회

LH 휴대전화 삭제 파일 일부 복구…시의원 포렌식 참관

2021년 03월 17일 02시 43분 댓글

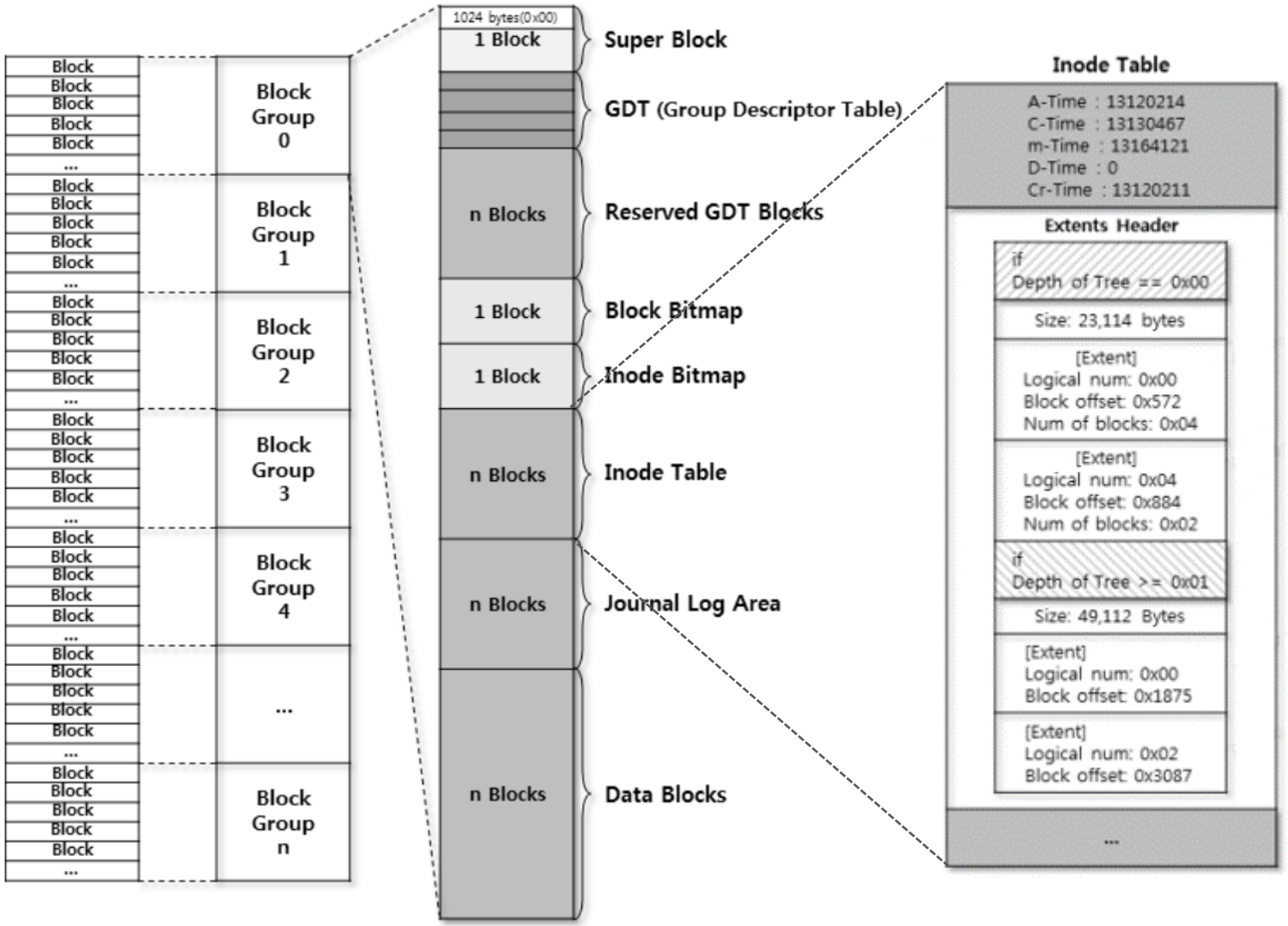
[가] 파일 카빙 사례

❖ 파일 카빙(File Carving)

- 손상·삭제되었으나 저장매체에 남아있는 내용들을, 메타데이터 없이 복구해 내는 기법
- 메타데이터를 이용하면 파편화된 파일들에 쉽게 접근 & 식별 가능하지만
- 메타데이터가 삭제되거나, 이미 손상되었을 가능성 O

- 따라서,
  - ✓ 데이터만을 이용해 파편화된 파일을 합쳐 복구하는 것: 파일 카빙
  - ✓ 우선 파편화된 파일을 유형 별로 분류& 식별하는 과정이 필요 -> 파일 파편 유형 분류(File Fragment Classification)

서론_리눅스 Ext4 파일 시스템 구조



Ext4 파일 시스템의 구조 [나]

***Inode Table 영역**

해당 블록 그룹 안의 모든 데이터에 대한
inode table들이 inode 번호에 따라 순서대로 존재

- 타임스탬프: 데이터 생성, 접근, 변경, 삭제, …
- 데이터 크기
- 실제 데이터가 존재하는 블록 포인터
- etc

배경지식_파일 시스템 삭제 연산

❖ 삭제된 파일을 복구할 수 있는 이유

- 사용자가 파일을 삭제해도

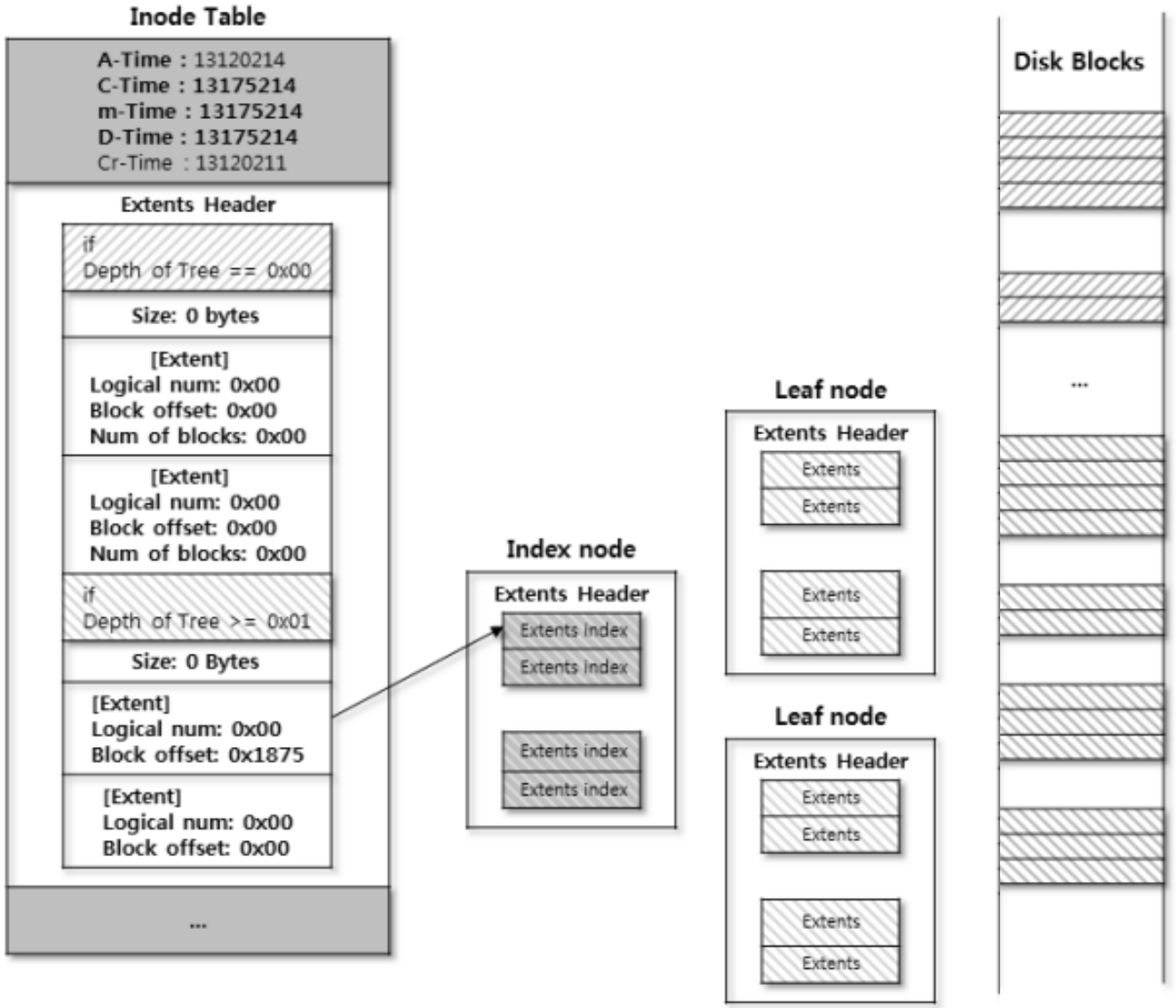
  디스크에서 물리적으로 바로 지워지지 않음.

- 파일을 가리키는 포인터가 해제되어

  파일이 접근할 수 없는 비할당 영역으로 남게 되는 것.

Ext4 파일 시스템의 파일 삭제 후 메타데이터 변화

Ext4 시스템의 삭제 연산 후		extents tree = 0	extents tree >= 1
inode table extent 정보	데이터 블록 offset	삭제	유지(첫 위치만)
	파일 크기 정보	삭제	삭제
타임스탬프	inode 변경 시간	변경(파일 삭제 시간)	삭제
	파일 내용 수정 시간	변경(파일 삭제 시간)	삭제
	파일 생성/삭제 시간	유지	삭제



삭제 연산 후 Ext4 파일 시스템의 파일 inode table extents 구조 [나]

배경지식_파일 복구 방법

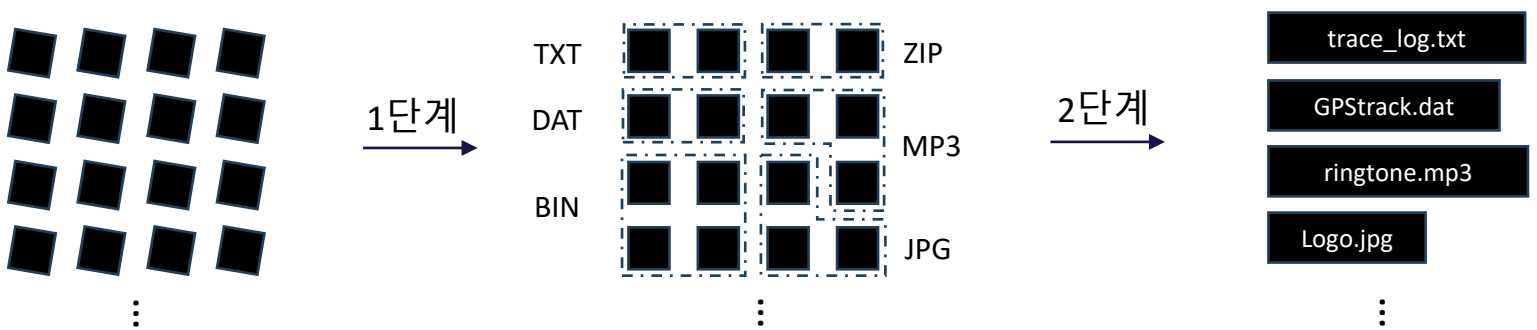
❖ 일반적인 파일 복구 방법: 파일 시스템 정보에 의존

- 파일 시스템 정보가 손상되었을 때는 파일 복구가 어려움

❖ 파일 카빙: 파일 시스템의 메타데이터 없이 파일의 구조와 내용에 따라 파일을 복구하는 방법

- 1단계: 파일 파편 유형 분류
- 2단계: 파일 파편 연결 및 복원

Steps in File Carving



배경지식_파일 카빙 및 파일 파편 유형 분류

❖ 파일 파편 유형 분류: 디스크에 연속적으로 저장되지 않은 파일 파편이 어떤 유형에 해당하는지 분류, 식별하는 것

- 파일 유형 별 특성을 통한 파일 파편 유형 식별의 중요성
  - ✓ 중요도가 높은 파일과 아닌 파일을 분류 가능(데이터와 아티팩트의 차이)
  - ✓ 파일 파편 재구성 및 복원
  - ✓ 복원 후 후속 분석 용이

→ **파일 카빙의 파편 복원 단계 계산량은 파편 유형의 정확한 분류 여부에 달림**

서론

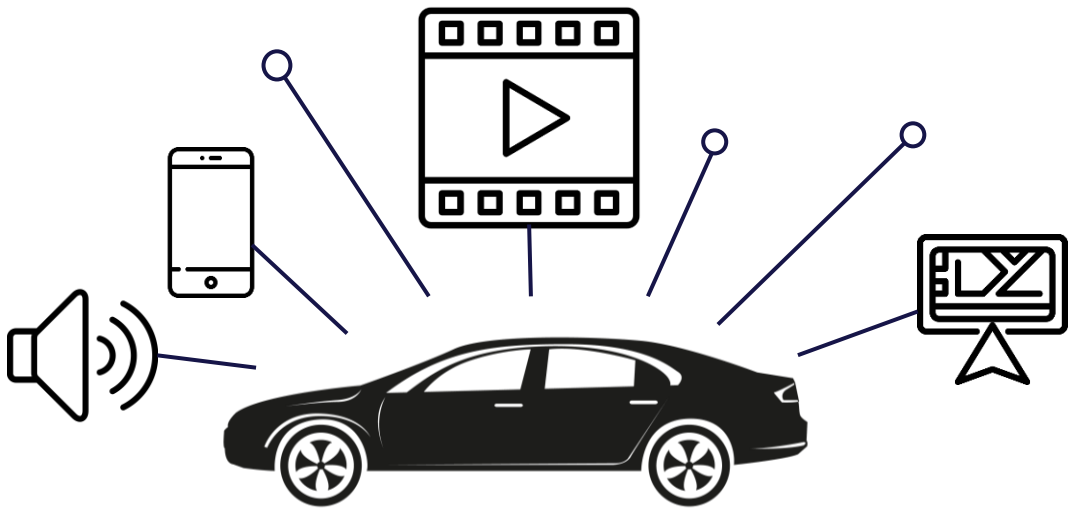
❖ 차량 디지털 포렌식 분야에 파일 카빙 적용 필요성

- 운전자의 여러 이벤트가 저장되는 자동차 AVN(Audio Video Navigation) 시스템에 적용 시
  - ✓ 자동차 사고나 내부 시스템 고장 또는 고의적인 삭제로 인해 손상된 파일을 분류·복원 가능
  - ✓ 운전자 행위 이력, 차량 이동 경로 및 동승자 탑승 여부 등의 이벤트 파악

→ 차량 관련 교통사고의 원인 규명과 범죄 추적에 도움

***AVN system**

- 오디오, 멀티미디어, 내비게이션 등 차량 내·외부의 다양한 장치 등이 통합된 시스템
- 차량 내 인포테인먼트(information + entertainment)를 제공함
- 차량 시스템들이 작동하며 생성하는 수많은 이벤트와 활동들이 로그 등의 형태로 AVN에 기록됨



AVN systems

서론

1. 디지털 포렌식(Digital Forensics)과 파일 카빙(File Carving)의 중요성 증가

2. 머신러닝을 사용하는 파일 파편 유형 분류(File Fragment Classification)의 필요성

3. 차량 디지털 포렌식에도 파일 카빙 적용 시 얻을 수 있는 기대효과가 클 것으로 예상

4. 자동차 AVN 시스템에 파일 파편 유형 분류 기법 적용을 위한 효과적 방법과 한계점 조사, 제시

02 파일 파편 유형 분류에 관한 기존 연구 분석

파일 파편 유형 분류에 관한 기존 연구 분석

[1] S. Fitzgerald, G. Mathews, C. Morris, and O. Zhulyn, "Using NLP techniques for file fragment classification," Digital Investigation 9 : S44-S49, (2012).

[2] T. Xu, M. Xu, Y. Ren, J. Xu, H. Zhang, and N. Zheng, "A File Fragment Classification Method Based on Grayscale Image," J. Comput. 9.8 : 1863-1870, (2014).

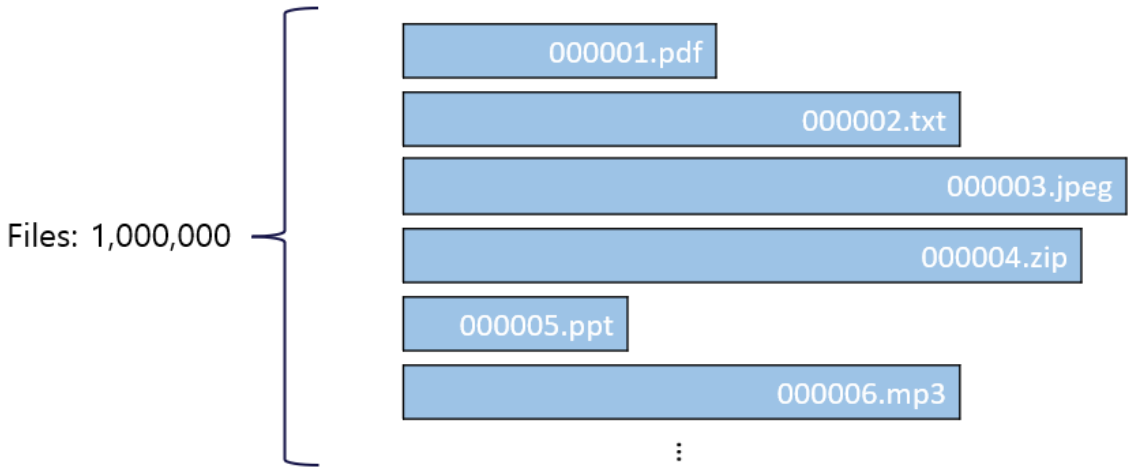
[3] N. Ameen, "Convolutional Neural Networks for Deflate Data Encoding Classification of High Entropy File Fragments," (2021).

[4] W. Qiu, R. Zhu, J. Guo, X. Tang, B. Liu, and Z. Huang, "A new approach to multimedia files carving,"2014 IEEE International Conference on Bioinformatics and Bioengineering, IEEE, (2014).

파일 파편 유형 분류에 관한 기존 연구 분석

❖ govdocs1 데이터 셋

- 2009년 Simson Garfinkel 등의 후원으로 구축
- PDF, PPT, TXT, JPEG, ZIP 등 총 63개의 유형 존재
- 파일 개수: 100만개 (1000개의 디렉토리, 각 1000개의 파일)
- 데이터 셋 불균형(파일 유형 개수 불균형)



File types: 63

File type	Nums
txt	100,000
pdf	20,000
jpeg	10,000

파일 파편 유형 분류에 관한 기존 연구 분석

[1] S. Fitzgerald, G. Mathews, C. Morris, and O. Zhulyn, "Using NLP techniques for file fragment classification," Digital Investigation 9 : S44-S49, (2012).

❖ 학습 데이터 구성- N-gram 방식: NLP에서 사용되는 횟수 기반의 벡터 표현 방식

- 무작위로 선정한 각 파일 파편의 바이트 값을 unigram과 bigram으로 나타냄
- 특징정보: 엔트로피, 해밍 가중치, 평균 바이트 값, 압축된 파일 길이

❖ SVM(linear kernel) 이용: 24개 유형 분류

- 최고 99.7%(.csv) 최저 2.3%(.pptx), 평균 49.1%의 정확도
- 파일 유형별 파편의 수가 많을 수록, 낮은 엔트로피를 가지는 파일 파편일 수록 정확도 높음.

❖ 한계점

- 데이터 스케일링 진행하지 않음
- Bigram 방식을 추가적으로 적용했을 때 정확도 증가량을 밝히지 않음

파일 파편 유형 분류에 관한 기존 연구 분석

[2] T. Xu, M. Xu, Y. Ren, J. Xu, H. Zhang, and N. Zheng, "A File Fragment Classification Method Based on Grayscale Image," J. Comput. 9.8 : 1863-1870, (2014).

❖ 학습 데이터 선정 (각 선정 방식마다 실험 진행)

- 파일 비편향: govdocs1의 모든 디렉토리에서 동일한 수의 파편을 랜덤하게 선택
- 유형 비편향: govdocs1의 모든 유형(29개)에서 동일한 수의 파편을 랜덤하게 선택

❖ 파일 파편들을 Grayscale로 변환 (파일 파편 단위는 1,024바이트로 설정)

- GIST Descriptor를 이용해 여러 개의 grayscale 이미지로 변환
- PCA(Principal Components Analysis)로 이미지의 차원 축소

***GIST Descriptor**

- CV 영역에서 사용되는 이미지 표현법
- 텍스처, 모양 및 색상 분포 등의 속성을 캡처하기에 용이
- 이미지를 여러 개로 쪼개기 때문에 차원이 높아짐

❖ KNN, 그리드 서치 이용: 29개 유형 분류

- 파일 비편향 방식: 평균 39.7% 정확도
- 유형 비편향 방식: 평균 54.7%의 정확도, 최고 100%(.ttf), 최저 11.2%(.pps),

❖ 한계점

- 컴퓨터 비전 분야에서 쓰이는 GIST Descriptor를 실험 목적에 맞게 최적화하는 과정 없이 그대로 사용

파일 파편 유형 분류에 관한 기존 연구 분석

[3] N. Ameen, "Convolutional Neural Networks for Deflate Data Encoding Classification of High Entropy File Fragments," (2021).

❖ 학습 데이터 선정

- 데이터 셋의 불균형을 해소하기 위해 파일을 압축해서 ZIP와 GZIP 파일을 직접 생성
- 8개의 유형에 대해 파일 파편 단위(크기)를 각각 256, 512, 1,024, 4,096 바이트로 grayscale 이미지 생성

❖ CNN + word-embedding layer: 8개의 파일 유형들을 각각 이진 분류

- GZIP에 대한 파일 유형들의 평균 정확도: 69.9% (gzip-jpeg: 99.60%, gzip-gzip9: 50.01%)
- 높은 엔트로피를 가지는 파일들에 대해서도 비교적 높은 정확도를 보임.
- 파일 파편 단위가 클 수록 분류 정확도 높음.

❖ 한계점

- 비슷한 압축 형식을 가질 수록, 압축률이 높을 수록 분류 정확도 낮음
- 워드 임베딩 레이어는 더 높은 정확도를 위해 차원을 높일 수록 더 많은 학습 데이터를 필요로 함

파일 파편 유형 분류에 관한 기존 연구 분석

[4] W. Qiu, R. Zhu, J. Guo, X. Tang, B. Liu, and Z. Huang, "A new approach to multimedia files carving,"2014 IEEE International Conference on Bioinformatics and Bioengineering, IEEE, (2014).

❖ 학습 데이터 선정

- Govdocs1 포함, 총 4가지 데이터 셋에 대해 각각 실험 진행
- 시그니처 정보(헤더, 푸터) 사용

❖ SVM(rbf kernel) 이용 + PUP(복원 규칙): 10개의 유형 중 JPEG 파일만을 식별

- JPEG 파일만을 식별: 각 데이터 셋에 대해 평균 75.2%의 정확도
- 복원까지 보임

❖ 한계점

- JPEG 파일에 한정된 결과
- 시그니처를 이용한 1차 분류는 현실적으로 적용하기 어려움

File Type	Header Signature(Hex)	Footer Signature(Hex)
JPEG	FF D8 FF E0 FF D8 FF E8	FF D9
GIF	47 49 46 38 37 61 47 49 46 38 39 61	00 3B
PNG	89 50 4E 47 0D 0A 1A 0A	49 45 4E 44 AE 42 60 82
PDF	25 50 44 46 2D 31 2E	25 25 45 4F 46
ZIP	50 4B 03 04	50 4B 05 06

파일 시그니처 [다]

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	0123456789ABCDEF
00000000	FF	D8	FF	E0	00	10	4A	46	49	46	00	01	02	01	01	2C	....JFIF.....,
00000010	01	2C	00	00	FF	E1	01	42	45	78	69	66	00	00	4D	4D	.,....BExif..MM
00000020	00	2A	00	00	00	08	00	07	01	12	00	03	00	00	00	01	.*..............
00000030	00	01	00	00	01	1A	00	05	00	00	00	01	00	00	00	62	...............b
00000040	01	1B	00	05	00	00	00	01	00	00	00	6A	01	28	00	03	...........j.(..
00000050	00	00	00	01	00	02	00	00	01	31	00	02	00	00	00	27	.........1.....'

JPEG 파일의 헤더 시그니처 [라]

파일 파편 유형 분류에 관한 기존 연구 분석

표 1. 기존 파일 파편 분류 기법들의 비교 분석

	S. Fitzgerald et al. [1]	T. Xu et al. [2]	N. Ameen [3]	W. Qiu et al. [4]
목적	전체 분류	전체 분류	압축 파일 분류	JPEG 파일만을 식별
파편 크기(byte)	512	1,024	256, 512, 1,024, 4,096	512
시그니처 사용	X	X	X	O
학습 데이터 형태	n-gram	회색조 이미지	회색조 이미지	-
모델	SVM (Linear)	KNN 외 5	CNN (워드 임베딩)	SVM (RBF)
유형 종류	TXT, GZIP, ZIP 외 21	TXT, GZIP 외 27	GZIP, ZIP 외 6	TXT, GZIP, ZIP 외 7
기타 (데이터 선정, 전처리 등)	무작위 추출	파일 비편향, 유형 비편향으로 추출	ZIP, GZIP 생성해 균형 추출	govdocs1 외 추가 데이터 셋

03 자동차 AVN 파일들의 파편 유형 분류방안

AVN 시스템에서 파일들의 유형 분석

❖ Autopsy를 이용한 AVN 시스템 파일 유형 확인

- Kia K5 (2015)
- Kia NIRO EV (2018)
- Hyundai Sonata DN8 (2019)
- Kia All New Morning (2020)

표 2. AVN 시스템에서 확인한 파일 유형

항목	파일 유형	개수
문서	bak, bsmg, cal, cfg, cls, conf, cps, dat, db, dict, fdt, h, html, idc, info, ini, java, kcm, kl, list, log, ok, pak, par, pdc, poi, prop, sdc, txt, udt, xml	31
데이터베이스	db, sqlite	2
아카이브	ar, gz, jar, tar, zip	5
음악	mp3, ogg, wav	3
비디오	swf	1
이미지	bak, env, gif, jfif, jpg, png, so, tif, tmp, webp	10
응용 프로그램	0, apk, ar, ares, bak, bc, bin, bmd, bmsg, bsd, cfg, clm_blob, cnd, conf, dat, data, db, db-journal, db-shm, db-wal, dcs, def, dex, dict, dtb, elf, expkg, ext, fast, filemanager-journal, g2g, gz, h, ida, img, index, info, java, jil, journal, json, ko, layout_version, lcf, lease, localevar, log, mco, metadata, mpd, muf, mufs, nomedia, o, odex, pak, parcel, pcm, pem, pid, pjn, pjo, ppf, preferences, scr, sgt, sh, shaders_cache, skn, so, sqlite, swf, swiarb, swimdl, tar, temp, tmp, ttf, tuxera, txt, ufsd, vtdb, xml, xmp, zip	85
합계	(중복제외)	114

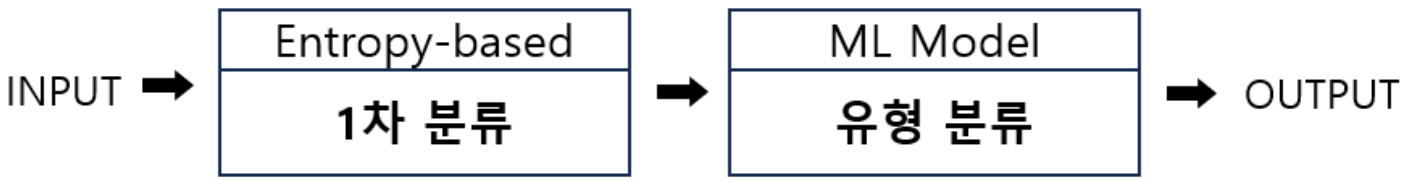
AVN에서 파일들의 파편 유형 분류 방안

❖ 학습 데이터 셋 구성

- 각 유형 별 충분한 파편 수 확보 & 데이터 수가 균등해야 할 필요성
- 파일 파편 크기가 1,024 바이트 이상일 때 효과적인 학습 가능

❖ 분류 모델 구성

- 1차 분류: 엔트로피가 높은 파편과 낮은 파편으로 분류
- 2차 분류: 엔트로피 기반으로 분류된 그룹 각각에 머신러닝 모델을 적용해 유형 분류 진행
  - ✓ CNN 사용: 엔트로피가 높은 유형을 비교적 잘 분류해낸 딥러닝 모델



AVN 파일들의 파편 유형 분류 방안

AVN에서 파일들의 파편 유형 분류 적용 시 예상되는 한계점

❖ 낮은 분류 정확도

- 메타데이터나 시그니처를 활용할 수 없음
- 엔트로피가 높은 파일 유형

❖ AVN 시스템의 파일 파편

- 효과적인 학습을 위해 1,024 바이트 이상의 파편 크기 확보가 가능한 것인지 실험과 분석을 통해 결정
- AVN 시스템의 파일 카빙 연구에 활용할 수 있는 검증된 공개 데이터가 존재하지 않음

04 결론 및 향후 연구

결론

1. 4개의 AVN 시스템에서 유형 목록을 확인

2. 머신러닝을 사용하는 파편 유형 분류 연구 4가지를 비교·분석
   - AVN 시스템에 파일 카빙 기법의 적용 가능성과 방법론 제시
   - 디지털 포렌식에서 차량 포렌식으로 계승될 수 있는 파일 파편 유형 분류 관련 한계점 정리

향후 연구

❖ AVN에 카빙을 적용하기 위한 데이터 셋 확보

- AVN 시스템의 파일 카빙과 파편 유형 분류에 적용할 수 있는 공개 데이터 셋의 부재
- 현대/기아 차량 대상 이미징 파일
- Govdocs1 데이터 셋

❖ AVN 시스템에 머신러닝 적용

- 확보한 데이터 셋에서 특징정보 탐색
- 앞선 연구에서 사용한 SVM, KNN, CNN을 적용

❖ 기타

- 분류한 파편 유형을 이용해 삭제된 파일 복원

Acknowledgement

이 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의
지원을 받아 수행된 기초연구사업임(no. 2021R1A2C2012574),
또한
2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의
지원을 받아 수행된 연구임(No.2022-0-01022, 이벤트 기반 실험시스템
구축을 통한 자동차 내·외부 아티팩트 수집 및 통합 분석 기술 개발).

이미지 자료 출처

[가] "YTN." [단독] LH 휴대전화 삭제 파일 일부 복구...시의원 포렌식 참관, n.d., www.ytn.co.kr/_ln/0103_202103161552488063. 2023년 8월 5일 접속.

[나] Kim, Dohyun, Jungheum Park, and Sangjin Lee. "File carving for Ext4 file system on android OS." Journal of the Korea Institute of Information Security & Cryptology 23.3 (2013): 417-429.

[다] https://sh1r0hacker.tistory.com/90

[라] http://forensic-proof.com/archives/300

Thank you

Q&A