Contact:
ristar1234@naver.com
+82-10-3115-5452

호기심 많은
Advancer & Implementer
연구원 황혜경입니다.

AI 연구원 (2024)

Contents

SUNGKYUNKWAN UNIVERSITY 1398 성균관대학교

MSL Media System Laboratory

Skills & Tools:

- **Tools:** Python, OpenCV, DL/ML 라이브러리, MATLAB, R, C++
- **수행 가능 딥러닝 Task:**
  Explainable AI, Classification,
  Active Learning, Representation Learning,
  Object Detection, Semantic Segmentation,
  Image Generation, Vision Language Model
- **XAI:**
  개념 기반 설명, 히트맵 기반 설명, Interpretable DNN, 고차원 특징 공간 시각화

STUDY:

Year	Contents
2018	Instance segmentation
2018~2021	Object Detection under Multi-type corruptions
2020	Image Generation
2021~2024	Explainable AI – Visualization – Interpretable structure

Publications:

- 논문: 7편 (국내: 1편, 해외: 5편/SCI급 주저자: 2편 게재, 1편 arxiv)
- 학회발표: 7회 (국내: 4회, 해외: 3회)

Projects:

- 기업 과제 : 2건 [㈜ SK Telecom, ㈜ Crescom]
- 국책 과제 : 5건 [제안서 작성/선발: 2회, 기술문서 작성 1회]

Stats:

연구 주제: The Quantification of Vulnerability and Uncertainties of Deep Learning

Computer Vision

Vulnerability Quantification,
Interpretable DL, Explainable AI,
Classification, Object Detection, Segmentation,
Image Generation

Post-hoc model analysis

- Identify responsible neurons
- Decision-base heatmap visualization
- Decision boundary approximation
- Bias visualization

Test Case Prioritization

- Failure Analysis
- Active Learning
- Adversarial training

Concept-based Explanation

- Unsupervised Concept discovery
- Indicator Parametrization
- Contrastive Learning
- Vision Language Model
- Concept activation vector

POST-HOC MODEL ANALYSIS

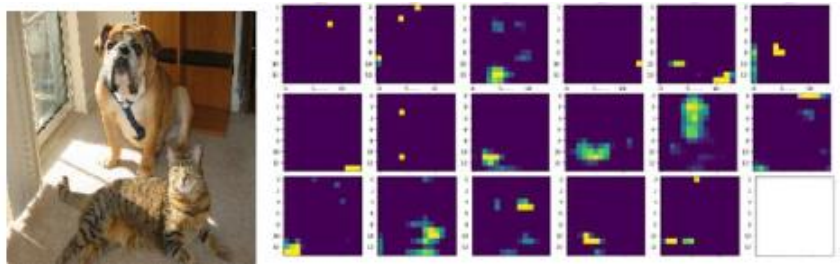
- 딥러닝 모델은 Feature sparsity로 인해, 의사 결정에 주요한 영향을 미치는 뉴런은 소수이다.

- **방법 :**

1. 단일 뉴런 중요도 = 단일 뉴런 제거 시 DNN 출력 변화 빈도 및 민감도
2. N개의 뉴런의 중요도를 모두 계산한 후, 각 레이어마다 출력 변화 빈도가 가장 큰 뉴런들을 추출하여 그래프로 연결

- **결과 및 활용**

1) **성능**: 당시 SOTA (NeuronShapley) 대비 **단위 뉴런 영향력이 1.3배 더 높은 뉴런을 찾음**
2) **Efficiency**:
   - **NeuronShapley**: 분산 컴퓨팅으로 서버 100대가 필요, hyperparameter sensitivity가 큼
   - **제안 방법:** Single GPU machine. Hyper-parameter **free**, 그래프 제작에는 30초 소요
3) **Pruning으로의 확장**: **하위 뉴런 20%를 삭제하더라도 성능이 유지됨**

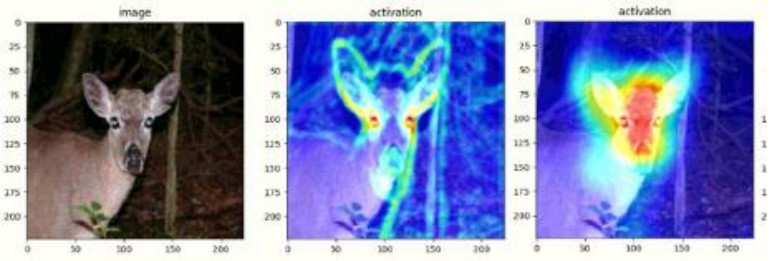


[그림] 딥러닝 특징 sparsity의 예시

(a) Concept patches from ox that is contrast to cow

(b) Concept patches from cow that is contrast to ox

[그림] (좌) Class 소가 물소로 오인되는 요인
(우) Class 물소가 소로 오인되는 요인



[그림] Layer 별 주요 뉴런의 활성 영역

POST-HOC MODEL ANALYSIS

- 딥러닝 모델은 Feature sparsity로 인해, 의사 결정에 주요한 영향을 미치는 뉴런은 소수이다.

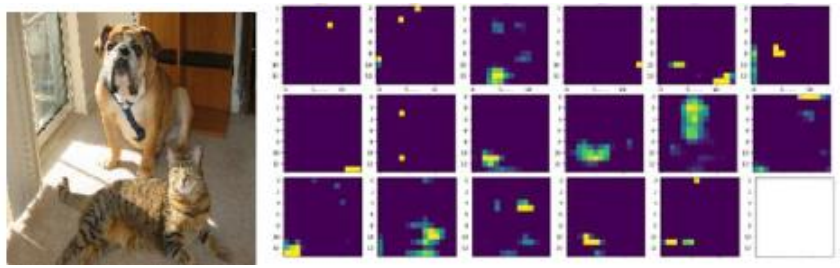
- **방법 :**

1. 단일 뉴런 중요도 = 단일 뉴런 제거 시 DNN 출력 변화 빈도 및 민감도
2. N개의 뉴런의 중요도를 모두 계산한 후, 각 레이어마다 출력 변화 빈도가 가장 큰 뉴런들을 추출하여 그래프로 연결

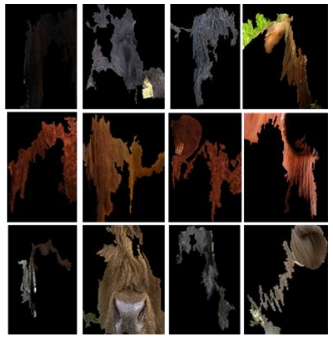
- **결과 및 활용**

1) **성능**: 당시 SOTA (NeuronShapley) 대비 **단위 뉴런 영향력이 1.3배 더 높은 뉴런을 찾음**
2) **Efficiency**:
   - NeuronShapley: 분산 컴퓨팅으로 서버 100를 요구하며, hyperparameter sensitivity가 크
   - **제안 방법:** Single GPU machine. Hyper-parameter **free**, 그래프 제작에는 30초 소요
3) **Pruning으로의 확장 가능성**: **하위 뉴런 20%를 삭제하더라도 성능이 유지됨**

1. Hyper-parameter 없는 고효율 뉴런 추출
2. 분류 편향 시각화
3. 모델 경량화



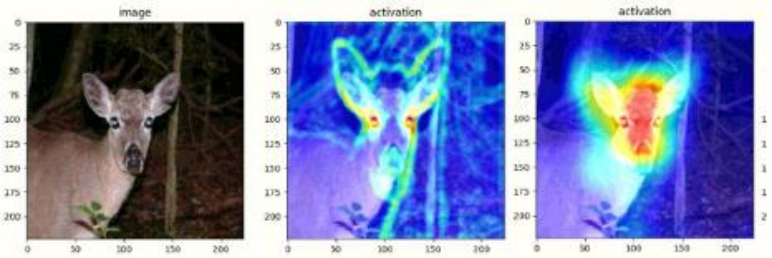
[그림] 딥러닝 특징 sparsity의 예시



(a) Concept patches from ox that is contrast to cow

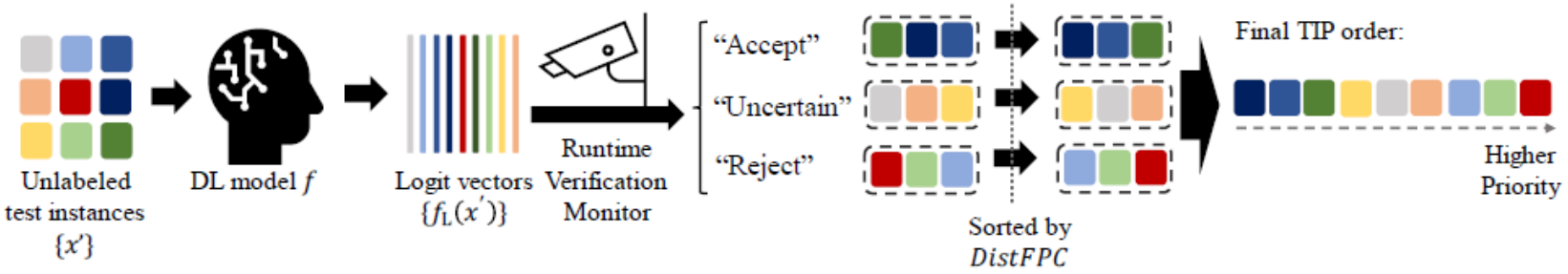
(b) Concept patches from cow that is contrast to ox

[그림] (좌) Class 소가 물소로 오인되는 요인
(우) Class 물소가 소로 오인되는 요인



[그림] Layer 별 주요 뉴런의 활성 영역

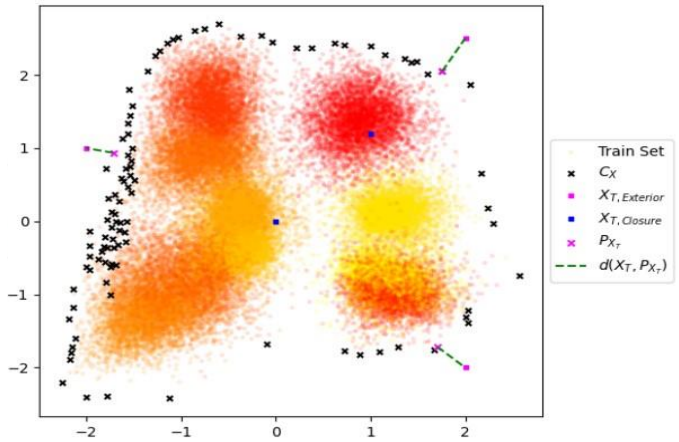
TEST CASE PRIORITIZATION



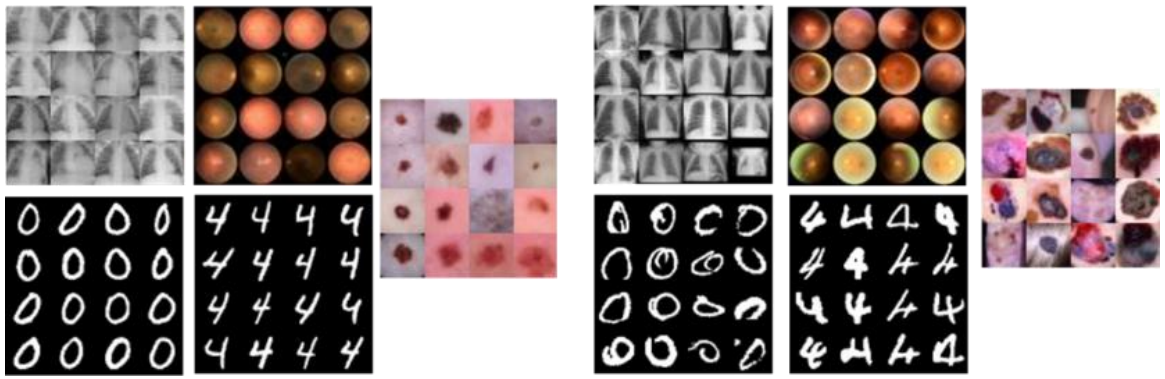
- Unlabeled 상황에서 비관측 데이터의 위험도 측정
- **방법**:
  1. Higher-dimension convex polytope approximation 을 통한 학습 데이터 Polygon 기반의 거리 측정법 제안 (To-hull Distance)
  2. False prediction cluster 추정 및 거리, 각도를 모두 고려한 triangular distance 제안

- **결과 및 활용**:

1. 8가지 데이터셋에서 7개의 기존 metric 와 비교하여 false prediction 검출률 SOTA 달성 (4.24% 개선)
2. Multi-type corruption 데이터에 대한 false prediction 검출률 SOTA 달성 (95.11%)
3. 제안 metric 기반 adversarial sample 검출 정확도 최대 94.54%
4. Active learning image sampling 의 모델 성능 개선도 SOTA 달성

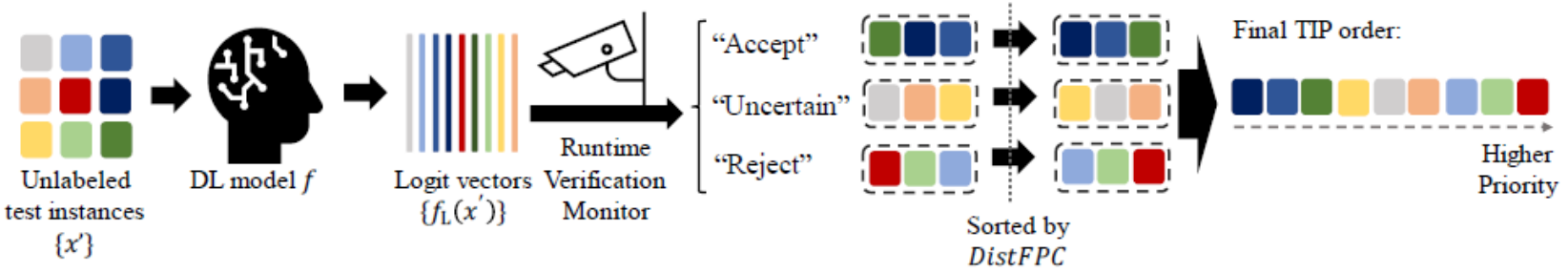


[그림] CIFAR10-ResNet50 의 학습 데이터셋에 특징 공간 polygon 추정 결과



[그림] (좌) 제안 measure에 따른 easy sample
(우) 제안 measure에 따른 Hard sample

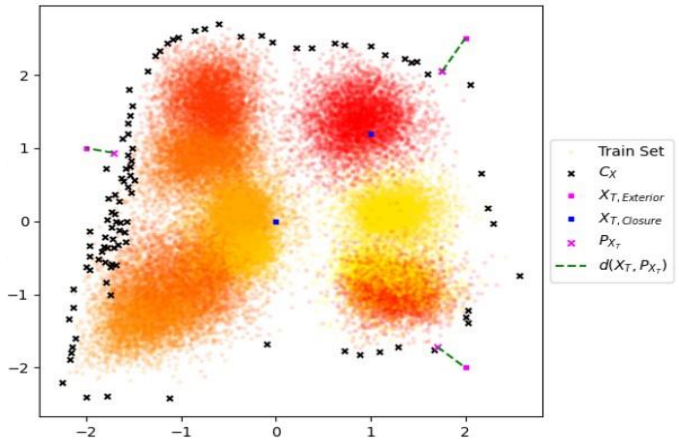
TEST CASE PRIORITIZATION



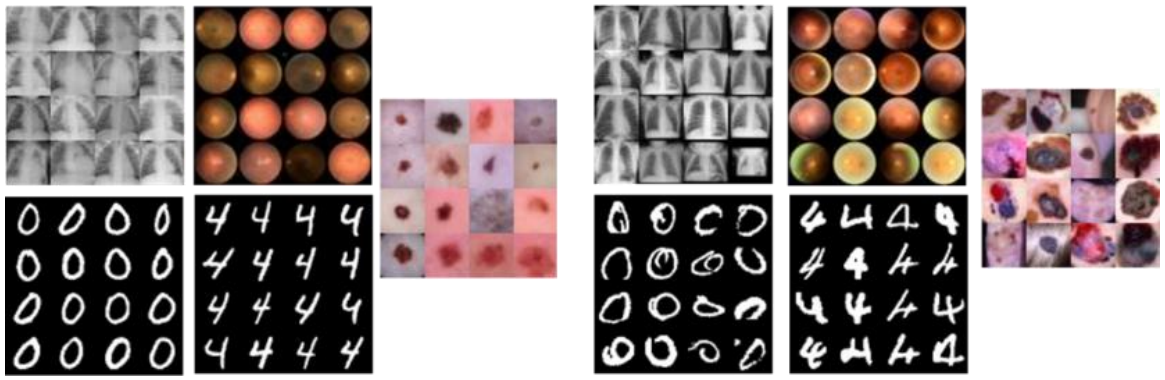
- Unlabeled 상황에서 비관측 데이터의 위험도 측정
- **방법**:
  1. Higher-dimension convex polytope approximation 을 통한 학습 데이터 Polygon 기반의 거리 측정법 제안 (To-hull Distance)
  2. False prediction cluster 추정 및 거리, 각도를 모두 고려한 triangular distance 제안

- **결과 및 활용**:
1. 8가지 데이터셋에서 7개의 기존 metric 와 비교하여 false prediction 검출률 SOTA 달성 (4.24% 개선)
2. Multi-type corruption 데이터에 대한 false prediction 검출률 SOTA 달성 (95.1%)
3. 제안 metric 기반 adversarial sample 검출 성능 최대 94.94%
4. Active learning image sampling 의 모델 성능 개선도 SOTA 달성

1. 난이도 추정을 통한 모델 개선 데이터 수집
2. 수집된 난이도 기반 데이터의 모델 개선도 검증
3. 딥러닝 모델에 대한 입력 난이도 실시간 모니터링

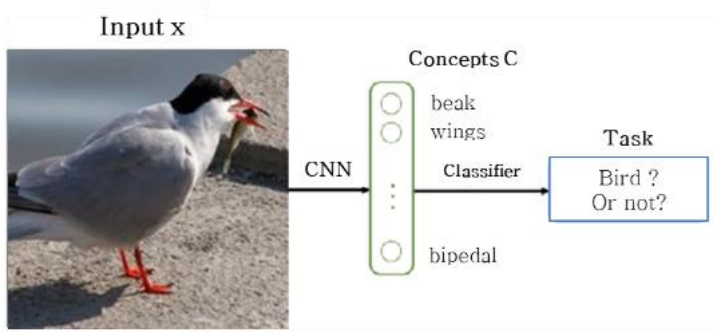


[그림] CIFAR10-ResNet50 의 학습 데이터셋에 특징 공간 polygon 추정 결과

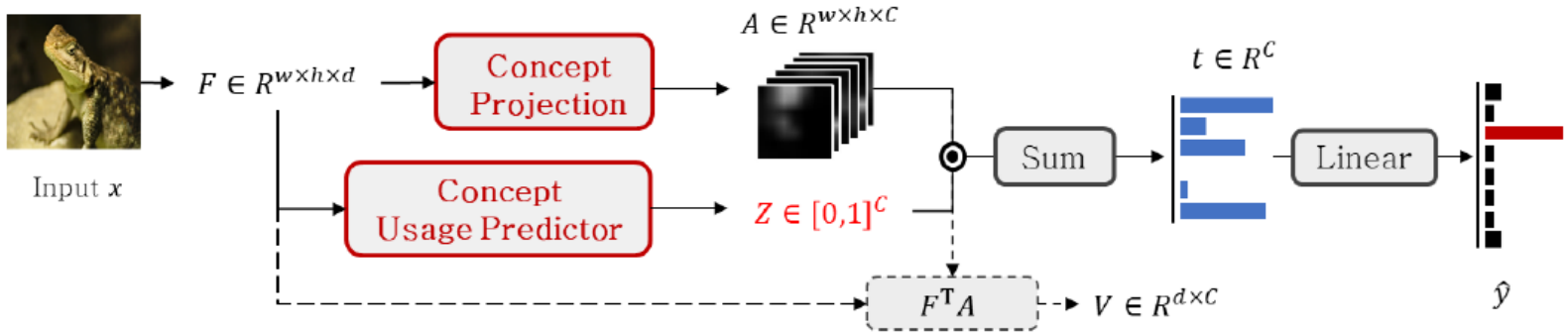


[그림] (좌) 제안 measure에 따른 easy sample
(우) 제안 measure에 따른 Hard sample

CONCEPT-BASED EXPLANATION

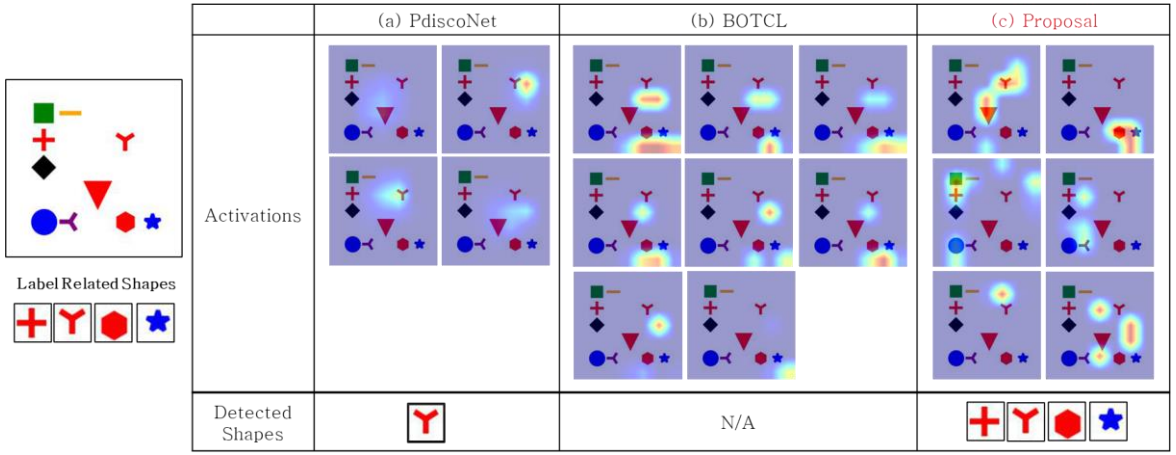


[그림] Concept bottleneck model



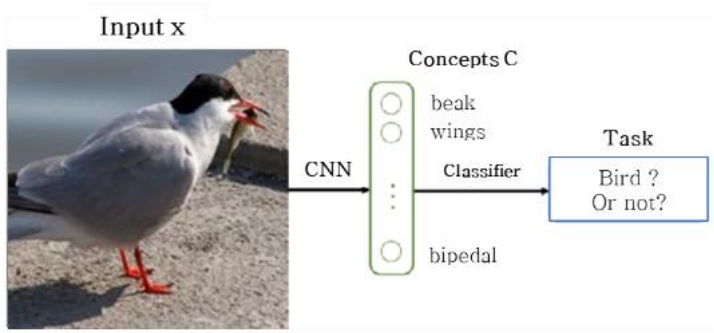
[그림] Proposed Method

- 딥러닝 특징 공간을 사람이 이해할 수 있는 semantics로 분해하고자 함
- **방법**:
  - Image를 Graph representation으로 변환하는 방법 제안
  - Downstream task에 발견된 개념의 사용 여부를 결정하기 위한 선형 레이어 제안
  - 대조학습을 통한 클래스 간 개념 및 클래스 내 개념 특징 관계 조정
- **결과 및 적용**:
  - Interpretable Model 중 fine-grained classification 정확도 SOTA 달성 → (부록 28p 참고)
  - Black box model 과 Compatible 한 분류 정확도 달성
  - 발견한 개념의 Completeness, Purity, Oracle Impurity Score, Niche Impurity Score 모두 SOTA 달성 → (부록 30p 참고)

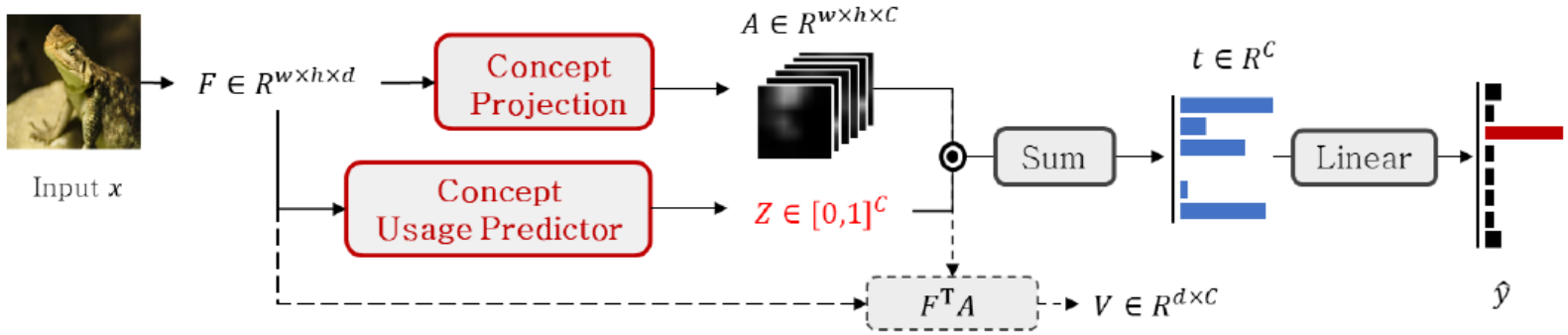


[그림] SOTA 와의 개념 발견 성능 비교

CONCEPT-BASED EXPLANATION



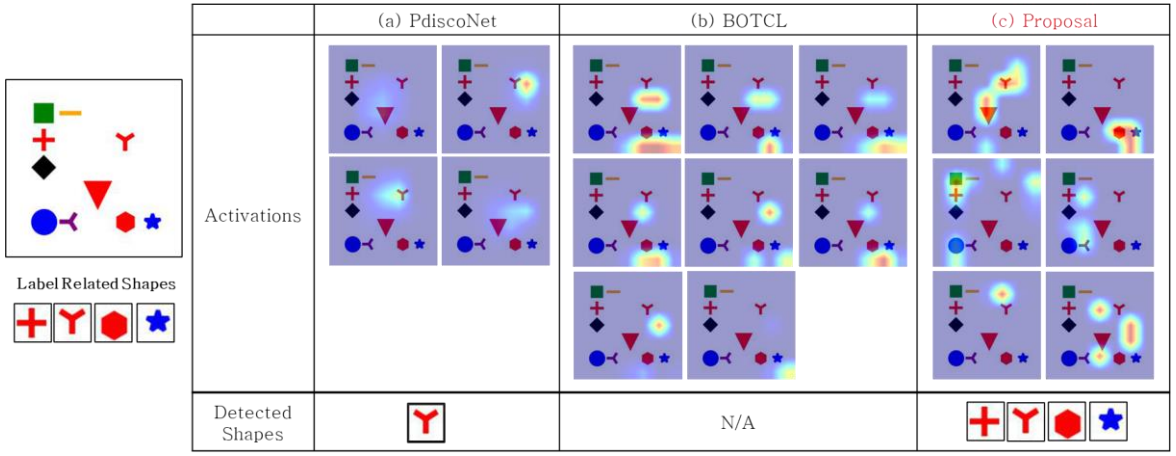
[그림] Concept bottleneck model



[그림] Proposed Method

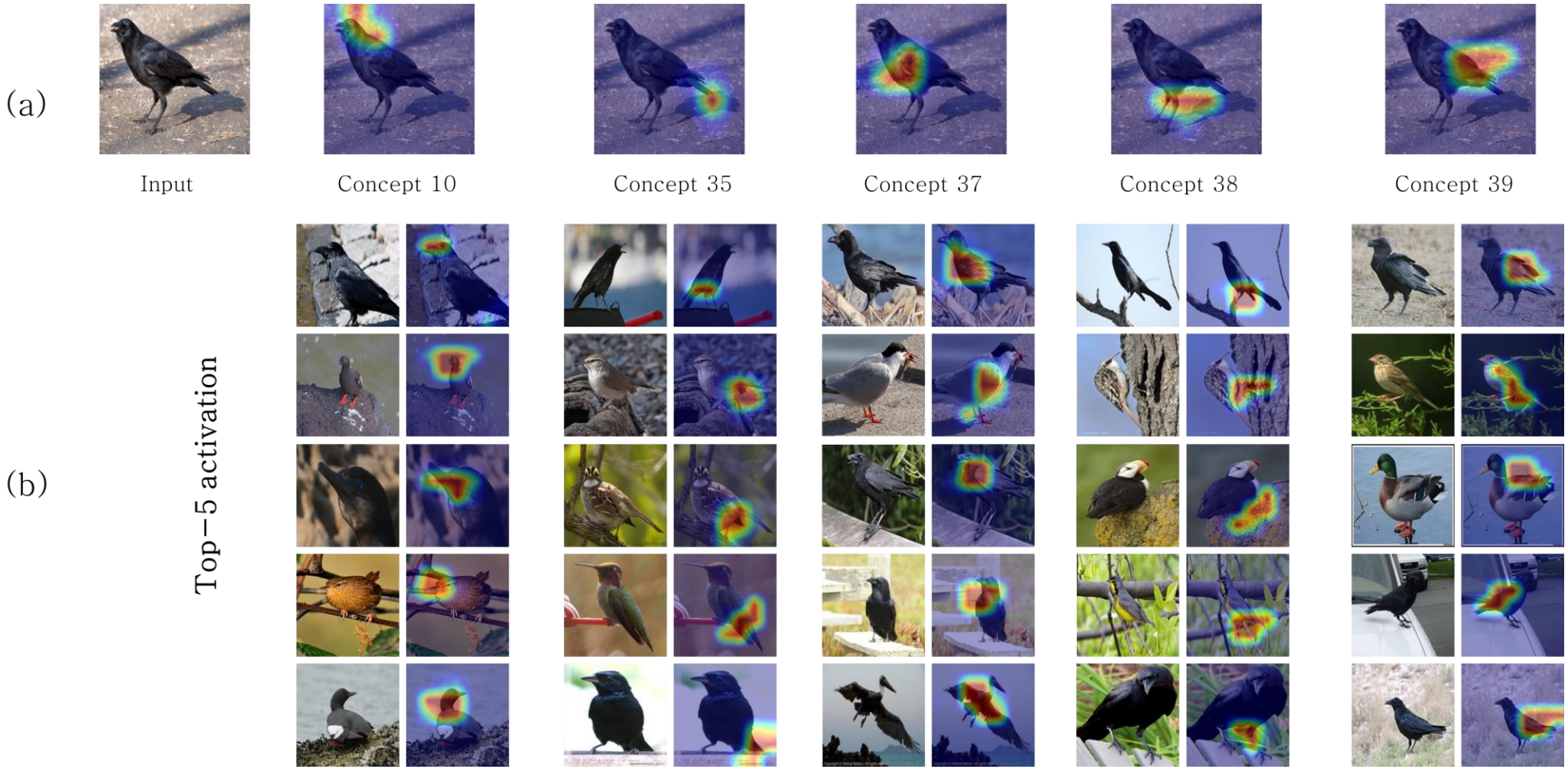
- 딥러닝 특징 공간을 사람이 이해할 수 있는 semantics로 분해하고자 함
- **방법**:
  - Image를 Graph representation으로 변환하는 방법 제안
  - Downstream task에 발견된 개념의 사용 여부를 결정하기 위한 선형 레이어 제안
  - 대조학습을 통한 클래스 간 개념 및 클래스 내 개념 특징 관계 조정
- **결과 및 적용**:
  - 발견한 개념의 Completeness, Purity, Oracle Impurity Score, Niche Impurity Score 등 SOTA 달성
  - Interpretable Model 중 fine-grained classification 정확도 SOTA 달성
  - Black box model 과 Compatible 한 분류 정확도 달성

1. 블랙박스 모델을 해석가능한 모델로 전환
2. 개념 라벨이 없는 상황에서 의미론적 단서 제공
3. Blackbox 와 견줄 수 있는 성능을 가진 해석 가능한 DL 설계

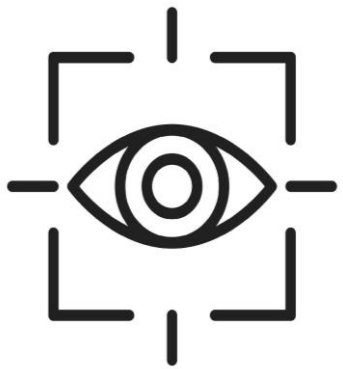


[그림] SOTA 와의 개념 발견 성능 비교

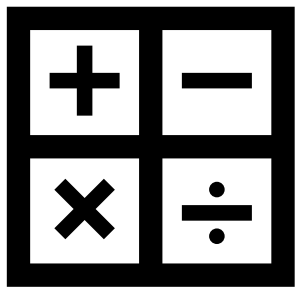
- Fine-grained classification 에서 제안 방법으로 발견 가능한 concepts



업무 강점

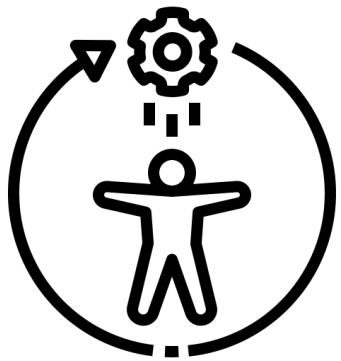


Computer Vision



Mathematics

Explainable AI



Responsibility

CAREER VISION

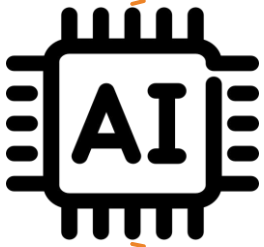
- 지원 동기 및 향후 포부

높은 성능

다양한 기능

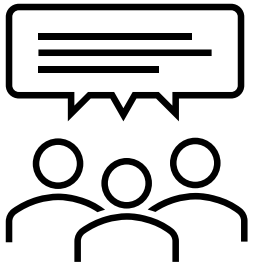
표준 수립

AI 의사 결정에 대한 신뢰도

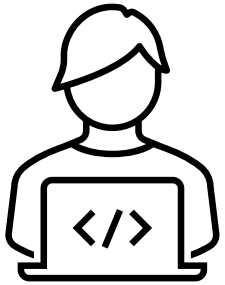


고객의 다양한 니즈에 맞춘 다양한 솔루션

영상 처리 딥러닝 전문가



요구사항 분석



기술적 적용

분석 및 비전 제시

PUBLICATIONS

요약

논문: 7편 (국내: 1편, 해외: 5편/SCI급 주저자: 2편 게재, arxiv 1편)
학회발표: 7회 (국내: 4회, 해외: 3회)

논문명	저자구분	저널명	SCI급 여부	Impact Factor
Improved Test Input Prioritization Using Verification Monitors with False Prediction Cluster Centroids	주저자	Electronics	Y	2.9
Chain Graph Explanation of Neural Network Based on Feature-Level Class Confusion	주저자	Applied Sciences	Y	2.7
Robust object detection under harsh autonomous-driving environments	공저자	IET Image Processing	Y	2.3
Optimization of Object Detection and Inference Time for Autonomous Driving	공저자	한국통신학회 논문지	N	-
Pedestrian detection using multi-scale squeeze-and-excitation module	공저자	Machine Vision and Applications	Y	3.3
Performance of ensemble methods with 2D pre-trained deep learning networks for 3D MRI brain segmentation	공저자	International Journal of Information and Electronics Engineering	N	-
Uncertainty Measurement of Deep Learning System based on the Convex Hull of Training Sets	주저자	Arxiv	N	-

논문명	저자구분	학회명	국내/해외
Test case prioritization with z-Score based neuron coverage	주저자	ICACT2024	해외
Singular Value Threshold Score CAM	주저자	IPIU2021	국내
Language of Gleam: Impressionism Artwork Automatic Caption Generation for People with Visual Impairments	공저자	ICMV2020 (최우수 구두발표상)	해외
Gaussian Clustering을 활용한 Single Shot Object Detector	주저자	한국통신학회 동계종합학술발표회	국내
Receptive Filed Steam Block을 이용한 실시간 객체검출 기법	공저자	한국통신학회 동계종합학술발표회	국내
KOREN 연구망을 활용한 Smart Safety Campus 서비스를 위한 real-time anomaly detection	주저자	한국통신학회 동계종합학술발표회	국내
Skipped-Hierarchical Feature Pyramid Networks for Nuclei Instance Segmentation	주저자	APSIPA 2018	해외

감사합니다 ☺

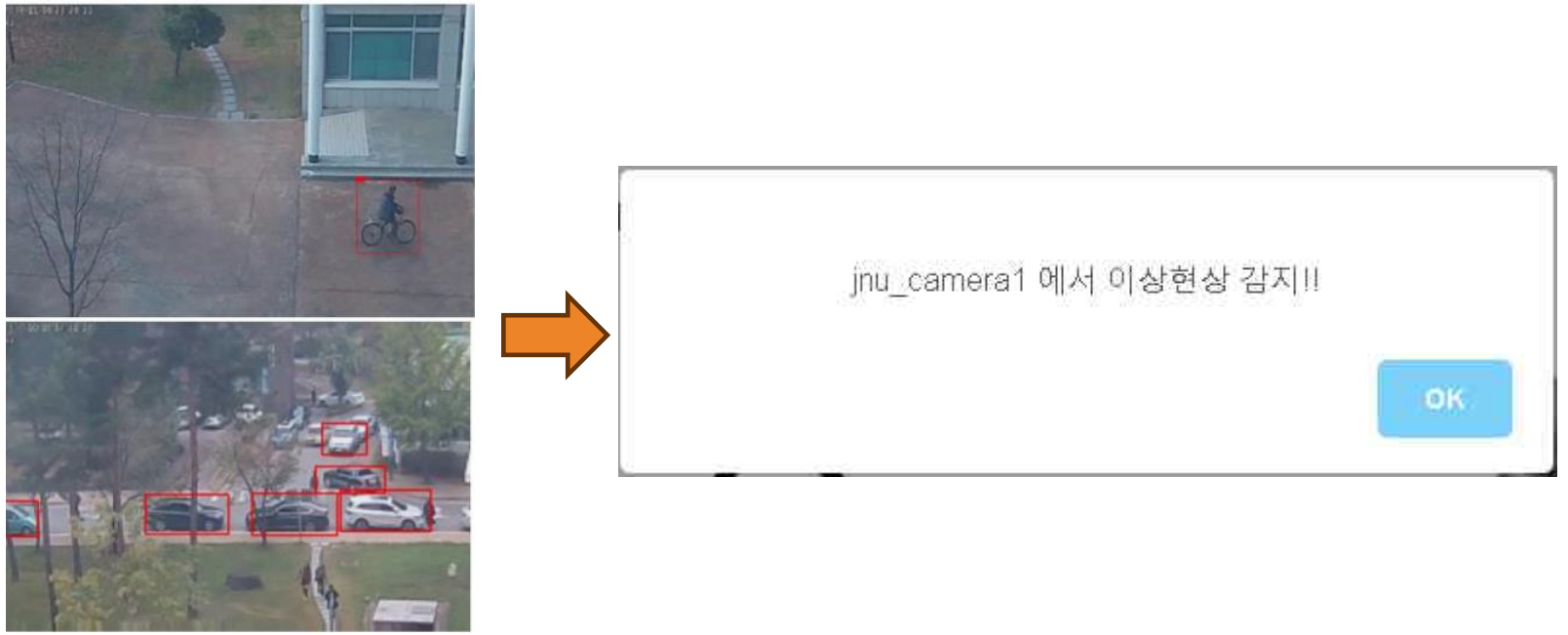
Any sufficiently advanced technology is indistinguishable from magic - Arthur Charles Clarke

부록 - 프로젝트 참여 내역 및 기여

KOREN SDI 기반 오픈 플랫폼 실증 (2018 - 2019)

기여: Object Detection

1. Campus CCTV 영상에 대한 이상 물체 레이블 제작
2. 딥러닝 기반 객체 검출기 (SSD)를 활용한 실시간 CCTV 영상 내 이상 물체 검출
3. Confidence Threshold 0.8 에서 mAP 0.75 확보
4. 『AI Network Lab 인사이트』 - “최신 AI 불확실성 정량화 동향 및 시사점” 보고서 발간



제안한 객체 검출기로 대학 캠퍼스 CCTV 영상 내 이상 물체 검출 프로젝트 실증 화면

자율주행 환경에서 고난이도 상황중심의 딥러닝 기반과 희소성 표현 기반의 영상처리 기술 (2018 – 2020)

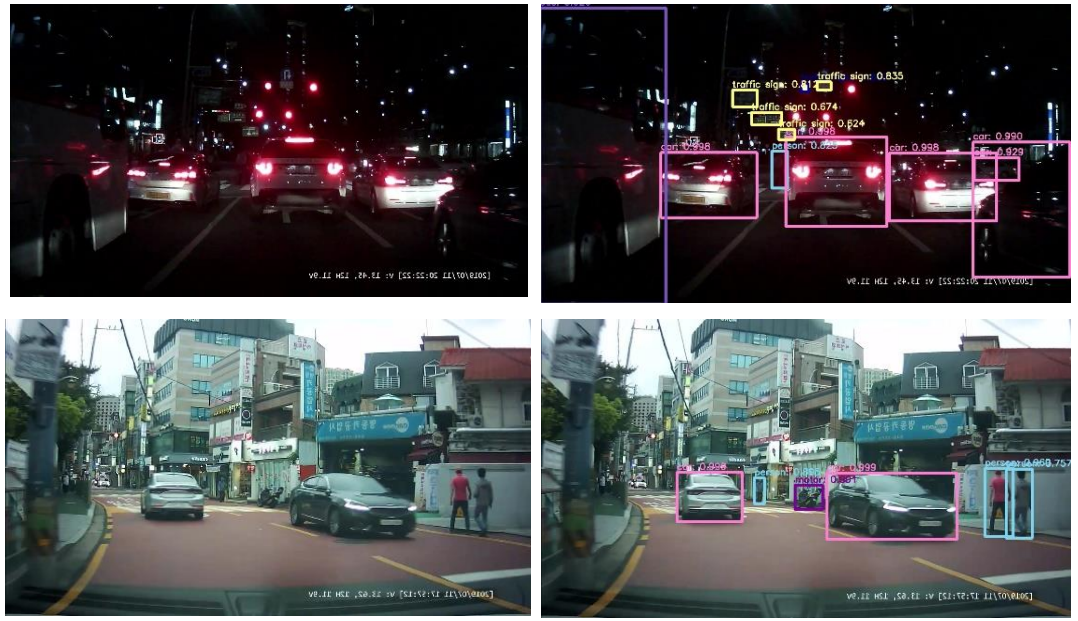
기여: Object Detection

1. 객체 검출 모델의 Bounding box candidate 의 reliability 를 기반으로 한 Bbox candidate filtering 을 통한 당시 SOTA (SSD) 대비 **모델 추론 시간 10% 감축, 1.3% 높은 mAP 달성**
2. VGG Annotator, MMDetection (M2Det) 을 활용한 국내 도로 객체 검출을 위한 데이터 라벨링

Table 1 Experimental Results on VOC07 test dataset

Model	Aero	Bicy	Bird	Boat	Bottle	Bus	Car	Cat	Chair	cow	Dining	Dog	Horse	Motor	Person	Pot	Sheep	Sofa	Train	Tv	mAP
SSD	80.7	84.0	74.7	70.7	52.4	**86.1**	85.7	**87.8**	61.3	82.0	**75.6**	83.9	85.9	**84.1**	76.4	49.4	74.3	78.1	83.3	76.1	76.6
Ours	**82.9**	**84.4**	**77.8**	70.7	**52.8**	85.5	**86.9**	87.1	**62.2**	**82.1**	75.2	**85.3**	**86.1**	**84.1**	**79.4**	**52.2**	**75.6**	**82.1**	**86.6**	**78.3**	**77.9**

제안 객체 검출 방법을 SSD에 적용한 결과에 따른 객체 검출 mAP 성능 표



라벨링 프레임워크 제작한 라벨링 프레임워크를 통한 라벨링 결과
(좌) 원본 블랙박스 영상, (우) 라벨링 결과

자율주행 환경에서 고난이도 상황중심의 딥러닝 기반과 희소성 표현 기반의 영상처리 기술 (2018 – 2020)

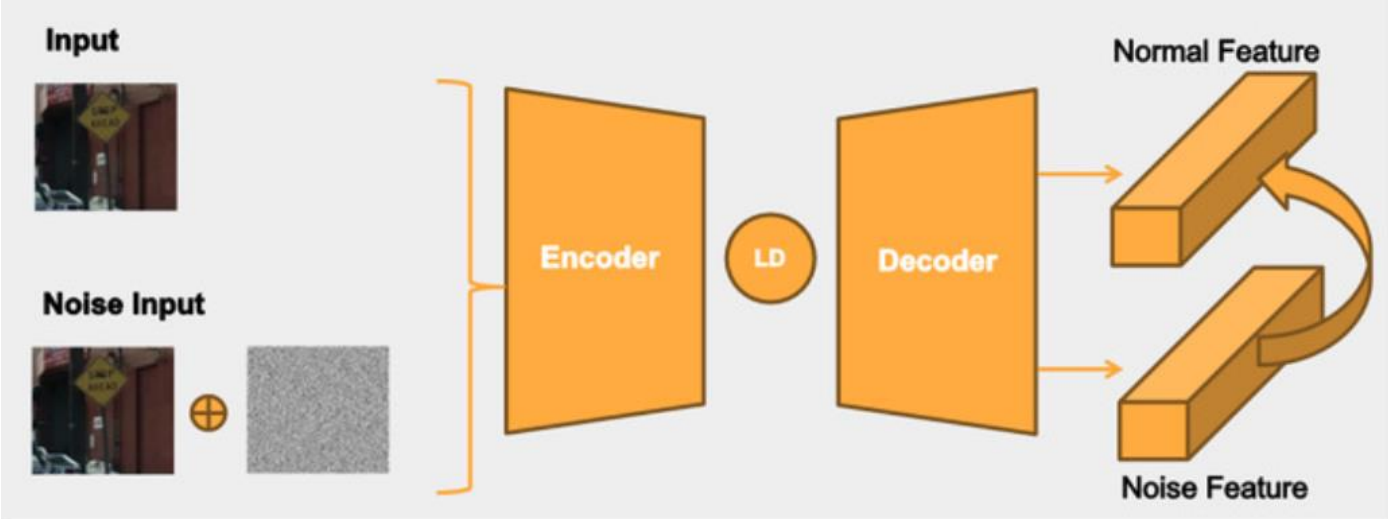
기여: Object Detection

1. 고난도 상황에 대한 그룹핑 기법을 활용한 적대적 학습 기법 제안
2. 당시 SOTA(FCOS) 대비 약 4배 빠른 FPS 및 약0.6% 높은 Localization 성능(AP75) 달성

Mutli-Type Corruption 문제:

- 장시간 운행, 터널 주행 등으로 인한 **카메라 노이즈**
- 객체 움직임, 고속 주행 등으로 생기는 **블러**
- **악천후**
- 카메라, 코덱, 네트워크 오류 등으로 인해 발생하는 **영상 품질 차이**

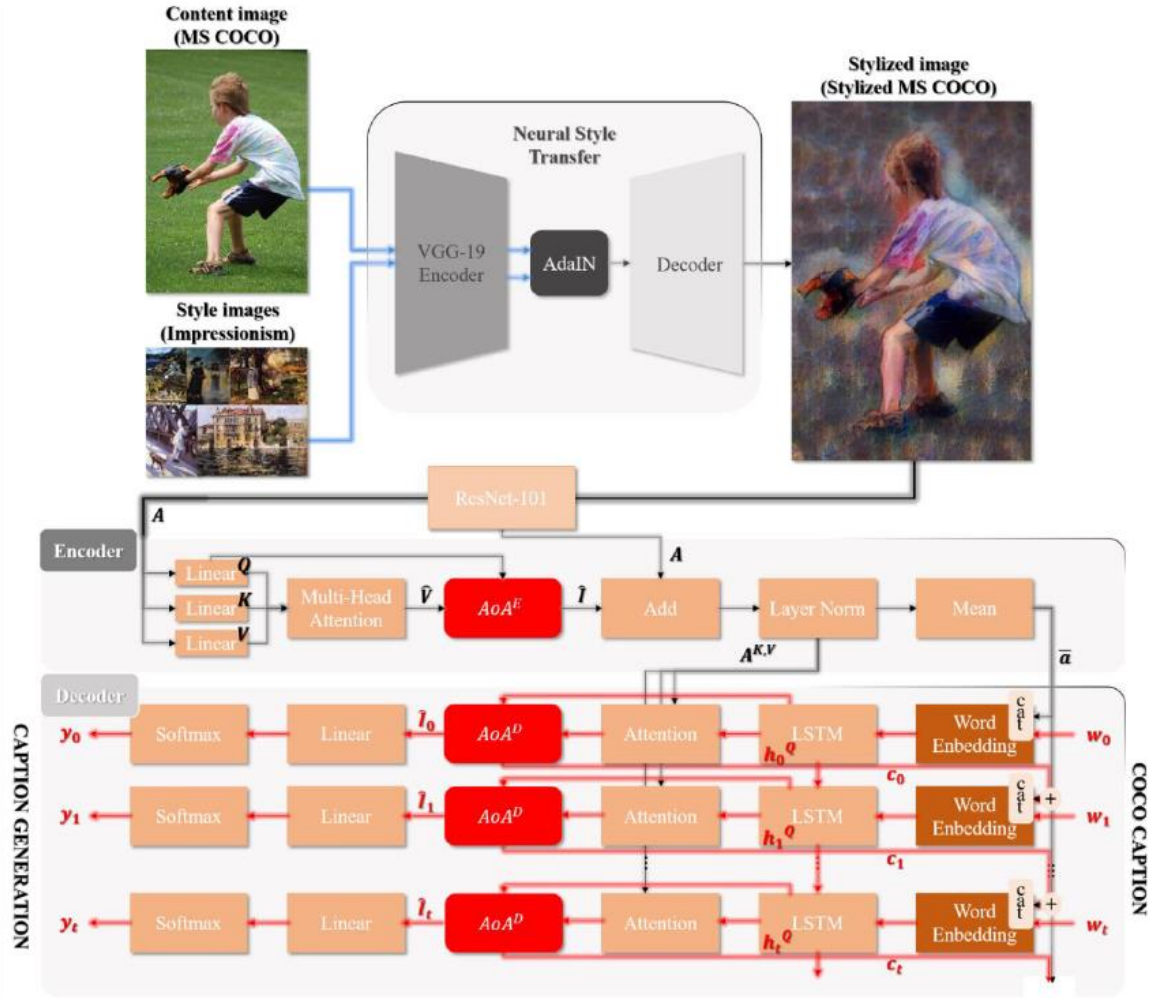
제안 모델 구조



시각장애인의 전시예술품 관람을 위한 도슨트 생성 모델 연구 (2020)

기여: Style Transfer 를 통한 데이터셋 제작

1. “시각 장애인 전시 예술품 도슨트 데이터의 부재” 라는 문제 제기
2. 일상 이미지 데이터와 인상주의 화풍 구도 사이의 유사점을 파악하여, 일반 이미지를 인상주의 화풍으로 변환하여 데이터를 제작할 것을 제안
3. AutoEncoder 구조와 AdaIN 모듈을 이용한 Style Transfer 모듈 구현
4. 시각 장애인 전시 예술품 도슨트 생성 모델 설계 참여



제안한 시각 장애인 전시 예술품 도슨트 생성 모델

제작한 모델을 통한 인상주의 화풍 예시 및 도슨트 생성 예:
papers/captioning-results.md at master · ailever/papers · GitHub

안저영상을 이용한 대표 안과질환들의 분류/검출/분할 및 미래 병변 예상 영상합성을 위한 설명가능한 딥러닝 기법개발 (2020 - 2023)

기여: Multi-class Image Classification

1. 안저 주요 5개 질병의 **임의 동시 발생을 포함한 영상의 질병 분류**
2. 안저 합병증 분류 모델 개발 (Accuracy: 94.7%, Specificity: 98.3%)
3. **XAI:** Heatmap 기반 딥러닝 모델 설명 기법 적용을 통한 모델 분석 제공
4. **XAI:** Heatmap 기반 딥러닝 모델 설명 기법 정량/정성적 평가의 손실 없이 추론 시간 최대 0.5배 단축

Task 1: 안저 주요 병변의 동시 존재 여부

연구: Classification, Class Imbalanced classification

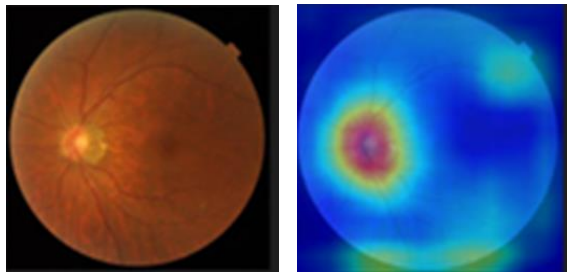
Normal [0/1]
Soft drusen [0/1]
Hard drusen [0/1]
Pigmentation & Exudate [0/1]
Other Abnormality [0/1]

1. 손실 함수 설계 시, 각 class에 대한 가중치를 batch 단위로 변화를 줌
2. 특징 공간에서의 re-sampling 을 적용한 feature augmentation 수행

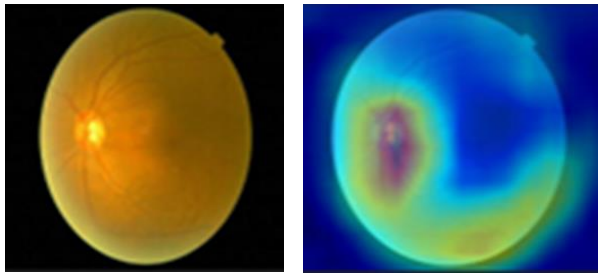
Task 2: 안저 주요 병변 분류 결과에 대한 Heatmap explanation 제작

연구: Class Activation Map, Matrix Decomposition

: Singular Value Decomposition 을 적용하여 channel-wise feature redundancy 를 추정하는 척도 개발



[정상 진단]



[비정상 진단 (하부 조직 경색)]

이동통신망 데이터의 효율적 군집화를 위한 딥러닝 기반 AI 엔진 개발
(산학과제, 2021)

기여: 비지도 학습 기반 딥 클러스터링을 통한 VoC 군집 분석

1. 네트워크 데이터 통계 분석 및 전처리
2. Categorical 데이터 피처 추출을 위한 인코더 제작
3. Deep Learning 기반 군집화 엔진 개발 (Purity: 86%)
4. 데이터 축적 기간에 따른 VOC 군집 특성 변화 분석

멀티모달 의료 데이터 기반 뇌출혈 질환 정밀 분석 및 심각도 예측 기술 개발
(산학과제, 2022 - 2024)

기여: 모델 개발 자문 및 의사 결정 요인 분석 방법 제공

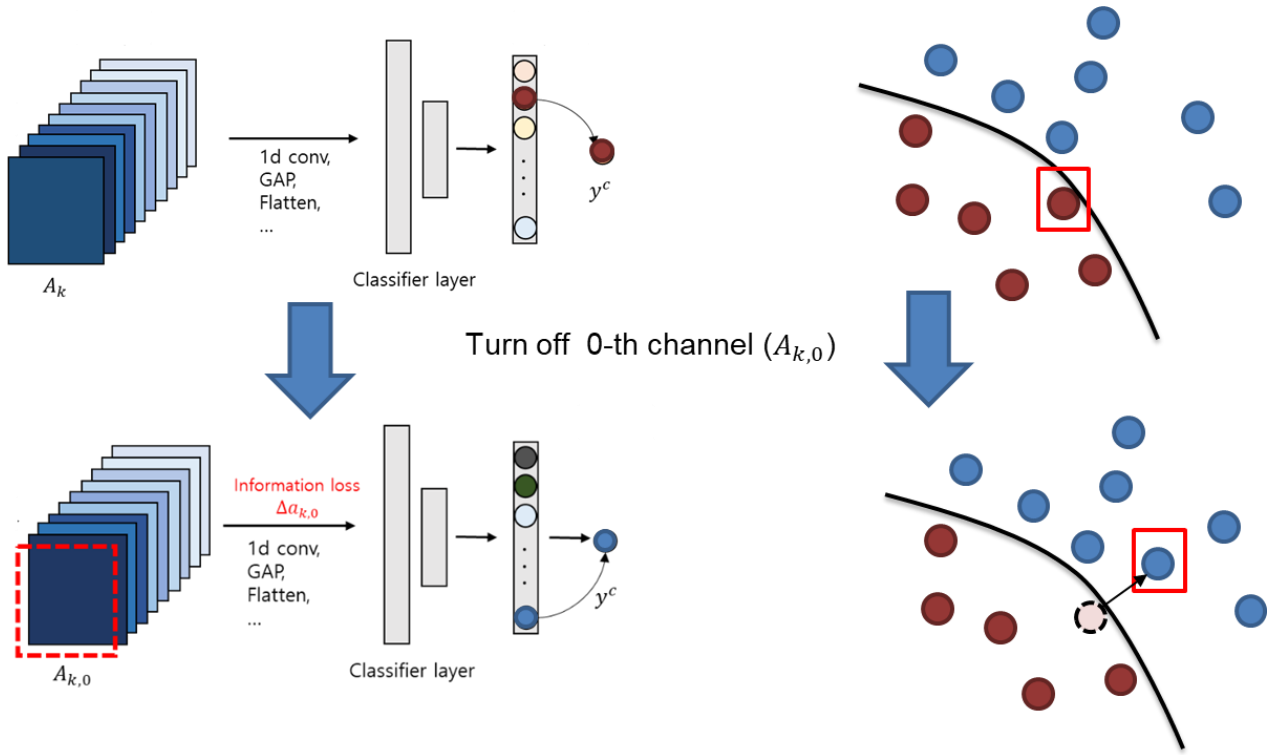
1. 매월 의료 관련 최신 AI 기술 세미나 진행
2. 설명가능 인공지능 기법을 통한 딥러닝 기반 뇌출혈 심각도 예측 모델의 의사 결정 요인 분석 제공
3. 영상 정보와 비 영상 정보 (EMR 데이터) 를 결합한 멀티모달 뇌출혈 심각도 예측 모델 제안
4. 뇌출혈 영역 semantic segmentation 모델 자문을 통한 제품 성능 개선

   [U-Net + Swin Transformer]

부록 - 실험 상세

POST-HOC MODEL ANALYSIS

- k 번째 뉴런 a_k 의 중요도 :
(N개 이미지에 대한 a_k 의 decision 변동 유발 횟수, 출력 값의 평균 변화도)



Dataset:

1. Animals with Attributes 2 (AwA2):
50 종류의 동물 분류를 목적으로 하는 37,322 장의 이미지

   - Train : Validation : Test = 7 : 1 : 2
   - Model: VGG16 (Accuracy 91.82%)

2. ImageNet

   - Model: Inception V3 pre-trained model
   (Accuracy 76%)

Validation set 의 50% 만을 이용하여 NeuronShapley와 제안 방법에서 각 뉴런의 중요도를 계산하였음

TEST CASE PRIORITIZATION

Dataset & Model:

ID	Dataset	Model	Train Acc (%)	Test Acc (%)
A	CIFAR10	MobilenetV2	99.51	93.38
B	CIFAR10	ResNet50	98.69	93.81
C	MNIST	ResNet18	98.03	97.66
D	FMNIST	WideResNet50	94.01	87.7
E	CIFAR100	EfficientNet V2-S	99.52	88.2
F	CIFAR10-C	ResNet50	-	77.29
G	MNIST-C	ResNet18	-	70.19
H	FMNIST-C	WideResNet50	-	47.83

- 오분류 검출률 측정 방법:

$$\mathrm{ATPF}(\%) = 100 \times \frac{1}{N_{\text{fail}}} \sum_{n=1}^{N_{\text{fail}}} \frac{N_{\text{err,n}}}{n}$$

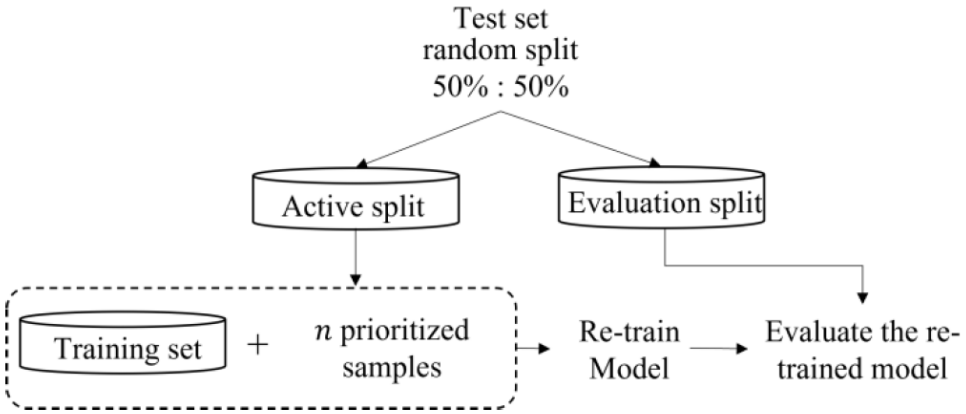
N_{fail} = test set 내 model failure의 총 개수
$N_{\text{err,n}}$ = n번째 sample 내에서 TCP metric이 감지한 model failure의 수

각 실험에 따른 오분류 검출률

ID	Gini [10]	DSA [8]	MLSA [9]	NBC [5]	NLC [6]	FD+ [7]	DA [14]	DeepFPC
A	54.53	53.85	52.3	51.835	34.818	70.37	78.88	**87.26**
B	52.62	47.58	48.48	56.071	40.109	35.46	71.61	**78.81**
C	51.8	63.79	56.94	7.51	3.13	46.58	64.36	**69.06**
D	59.13	55.66	41.63	9.45	11.79	35.48	**70.13**	68.56
E	**61.65**	60.11	61.39	10.62	12.26	28.81	60.4	58.18
F	66.33	29.13	31.17	52.74	25.29	57.57	78.0	**83.61**
G	58.68	55.82	56.96	71.23	49.32	66.83	85.41	**86.08**
H	73.58	71.29	73.37	71.18	70.16	49.06	92.39	**95.11**

TEST CASE PRIORITIZATION - 2

- Active Learning 설정

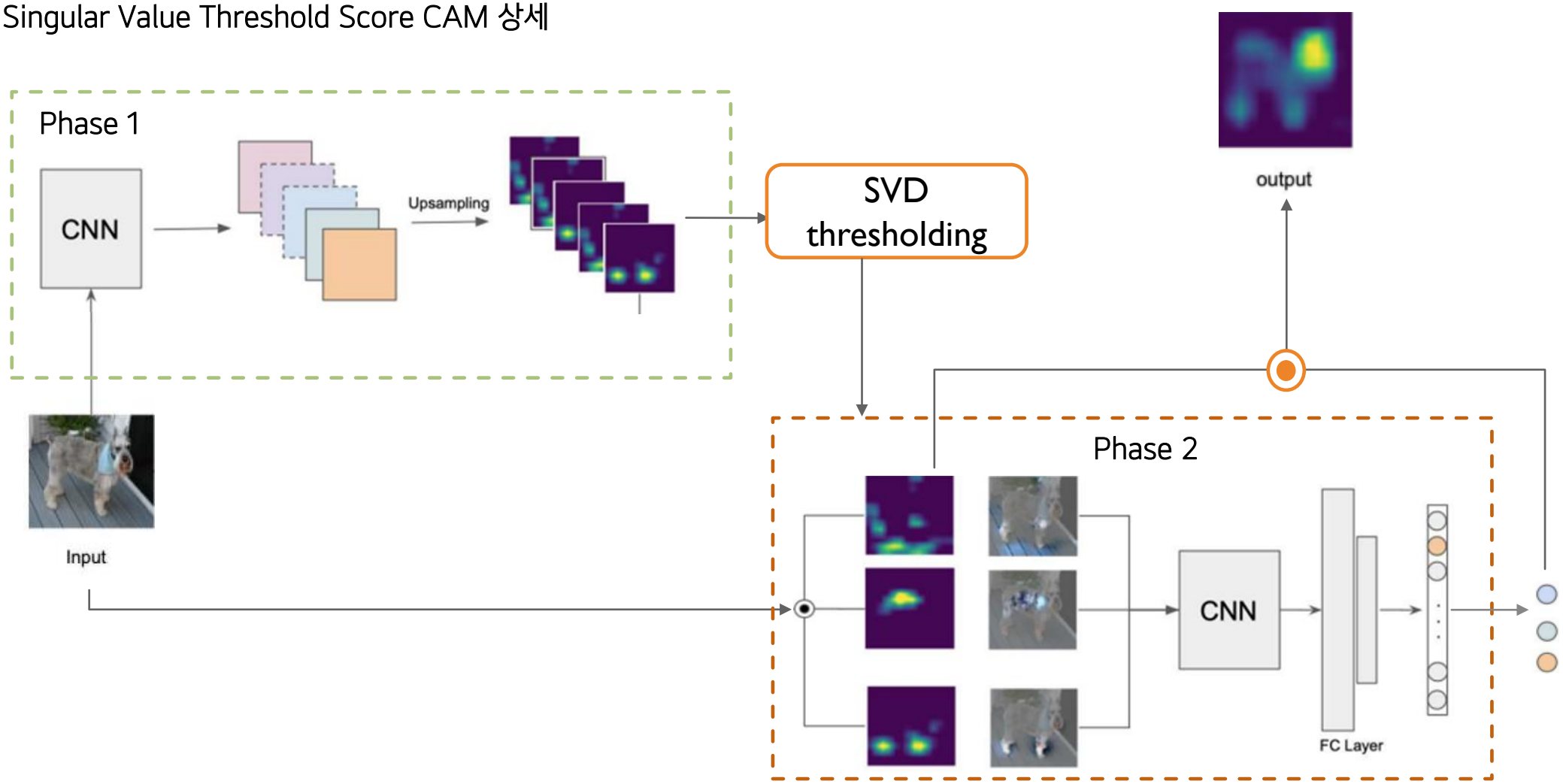


- Active Learning 실험 결과

TIP Metrics		Type of Evaluation Split	MNIST—ResNet18		FMNIST—WideResNet50	
			Type of Active Split		Type of Active Split	
			Nominal	OOD	Nominal	OOD
Neuron Coverage	NBC [5]	Nominal	90.28	67.48	91.12	51.04
		OOD	92.98	79.56	87.64	62.94
	NLC [6]	Nominal	91.94	70.38	87.70	41.38
		OOD	89.5	73.54	91.04	61.44
	FD+ [7]	Nominal	91.88	71.06	89.8	**52.44**
		OOD	91.18	72.98	87.1	60.5
Monitor-Based	DA [14]	Nominal	**93.92**	**74.36**	**92.32**	**52.34**
		OOD	93.6	**80.6**	88.34	65.18
	DeepFPC	**Nominal**	**92.2**	**74.16**	**92.48**	**52.44**
		OOD	**93.8**	**81.60**	**88.65**	**65.39**

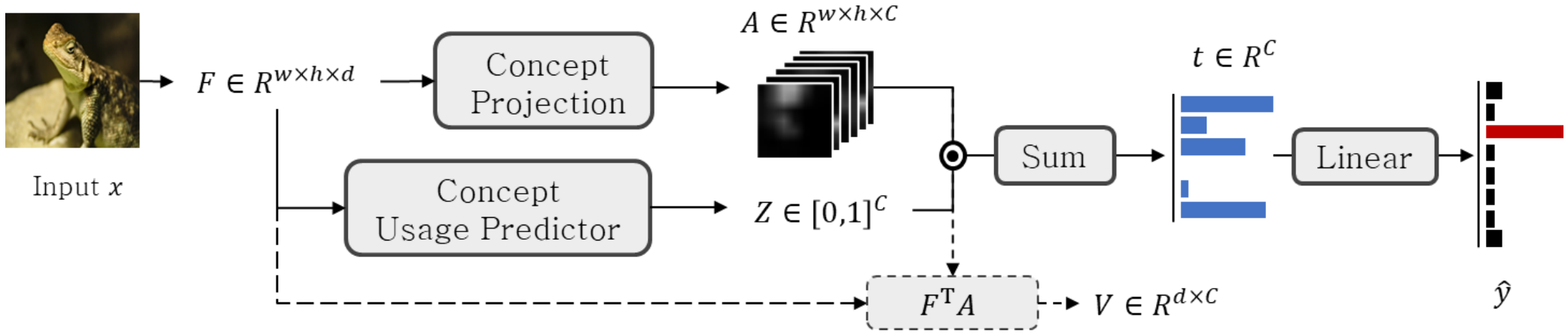
EXPLAINABLE AI - VISUALIZATION

- Singular Value Threshold Score CAM 상세



EXPLAINABLE AI – INTERPRETABLE DNN

- Algorithm Overview



Hypothesis: $\mathrm{p(Z)} \sim \mathbf{Bernoulli}(\theta)$

Method: Train a linear layer $W_Z \in R^{d \times C}$

1) Summarize $F \in R^{(h \times w \times d)} \rightarrow F \in R^d$
2) $p'(Z) \sim Bernoulli(p'(z) \mid \sigma(W_Z^T F))$

Train W_Z : KL divergence (let posterior to be close to the prior)

- t_κ: spatial summarization of $Z \cdot A$
  - ➔ Concept summarization
  - ➔ Input of classifier
- $V \in R^{d \times C}$: Average of F over (w, h) dimension weighted by $Z \cdot A$
  - ➔ Aggregation of image feature and concepts
  - ➔ Concept Regularization

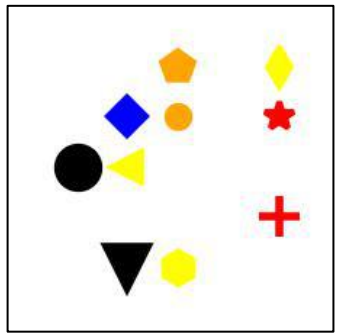
EXPLAINABLE AI – INTERPRETABLE DNN

[Classification 성능 평가 지표] : Accuracy

Accuracy (%)	Blackbox	PdiscoNet	BOTCL	제안 방법
BlackBox	**99.98 %**	81.91 %	99.31 %	99.82 %
AwA Classes = 50 Concepts = 20	84.39 %	88.13 %	87.94 %	**88.79 %**
CUB200 Classes = 200 Concepts = 50	74.03 %	73.35 %	74.62 %	**74.66 %**
ImageNet Classes = 1000 Concepts = 50	**69.758 %**	59.65 %	60.15 %	61.48 %

EXPLAINABLE AI – INTERPRETABLE DNN

- Interpretability 에 대한 정량 평가 방법



Dataset: Toy Data 제작 [Synthetics]

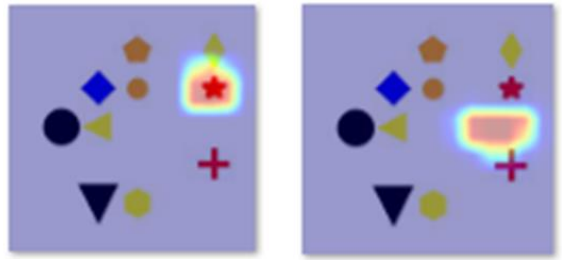
: 빈 바탕에 15 가지의 모양(5개 만이 정답과 관련 있음)들을 임의로 찍음
: 20,000 개의 이미지 제작, 9 : 1 로 train/test 분리

Label	Definition
ω_1	$\sim(s.1 \cdot s.3) + s.4$
ω_2	$s.2 + s.3 + s.5$
ω_3	$s.2 \cdot s.3 + s.4 \cdot s.5$
ω_4	$s.2$ xor $s.3$
ω_5	$s.2 + s.5$
ω_6	$\sim(s.1 + s.4) + s.5$
ω_7	$(s.2 \cdot s.3)$ xor $s.5$
ω_8	$s.1 \cdot s.5 + s.2$
ω_9	$s.3$
ω_{10}	$(s.1 \cdot s.2)$ xor $s.4$
ω_{11}	$\sim(s.3 + s.5)$
ω_{12}	$s.1 + s.4 + s.5$
ω_{13}	$s.2$ xor $s.3$
ω_{14}	$\sim(s.1 \cdot s.5 + s.4)$
ω_{15}	$s.4$ xor $s.5$

1. Concept Discovery 모델이 정답과 관련한 모양을 찾아낼 수 있는가?

➔ 정답 관련 모양과 concept k의 activation 영역 사이의 겹침 정도가 50% 이상 일 때,
해당 모양은 concept k 에 의해 **cover** 된다

$$h_{j\kappa} = \begin{cases} 1, & |s_j \cap a_\kappa|/|s_j| > \gamma \\ 0, & \text{otherwise} \end{cases} \qquad \text{Coverage}_{s\kappa} = \mathbb{E}[h_{s\kappa}]$$



[그림] 모양이 concept 에 의해 cover 된 경우 (좌)와 그렇지 않은 경우 (우)의 예

EXPLAINABLE AI – INTERPRETABLE DNN

[Interpretability 에 대한 정량 평가 지표]

- **Completeness**: concept activation 들이 각 모양을 얼마나 완벽하게 cover 하는가?

- **Purity**: 하나의 concept 이 한 개의 모양만을 다루고 있는가?

- **Oracle Impurity Score**: Feature level 에서 purty
- **Niche Impurity Score**: Feature level 에서 concept 사이의 correlation

Models	Completeness (↑)	Purity (↑)	OIS (↓)	NIS (↓)
PdiscoNet	0.312	0.365	0.456	0.627
BOTCL	0.177	0.351	0.446	0.581
Proposal w/o Z	**0.371**	0.628	0.405	0.583
Proposal	**0.371**	**0.631**	**0.391**	**0.563**