Vision AI

2022 arXiv Trends

2022-06

hutom

no.	Paper Title	Research group
1	SIoU Loss: More Powerful Learning for Bounding Box Regression	Zhora Gevorgyan
2	Multimodal Masked Autoencoders Learn Transferable Representations	Google Brain
3	Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images	1. Tecnologico de Monterrey, School of Engineering and Sciences, Mexico 2. CHU Nancy, Service d'urologie de Brabois, Nancy, France 3. Centre de Recherche en Automatique de Nancy, Universite de Lorraine, France

SIoU Loss: More Powerful Learning for Bounding Box Regression

Zhora Gevorgyan

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI)
ACM classes: I.2; I.4
Cite as: arXiv:2205.12740 **[cs.CV]**
(or arXiv:2205.12740v1 **[cs.CV]** for this version)
https://doi.org/10.48550/arXiv.2205.12740

Submission history

From: Zhora Gevorgyan [view email]
[v1] Wed, 25 May 2022 12:46:21 UTC (559 KB)

https://arxiv.org/pdf/2205.12740.pdf

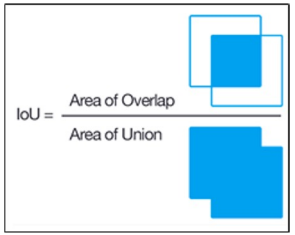
IoU Family

- **IoU (Intersection over Union; Jaccard overlap)**
- **GIoU (Generalized-IoU)**
  - Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression (CVPR 2019)
- **DIoU (Distance-IoU)**
  - Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression (AAAI 2020)
  - code
- **CIoU (Complete-IoU)**
  - DIoU 논문에서 동시에 제안.
- **Alpha-IoU**
  - Alpha-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression
- ICIoU
  - Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

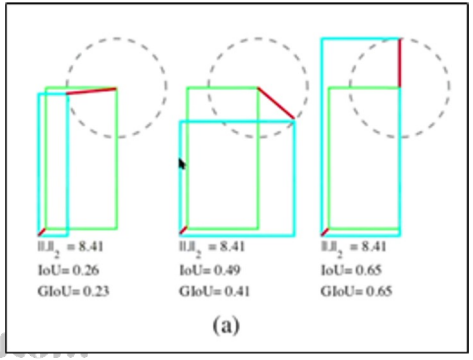
IoU (Intersection over Union; Jaccard overlap)

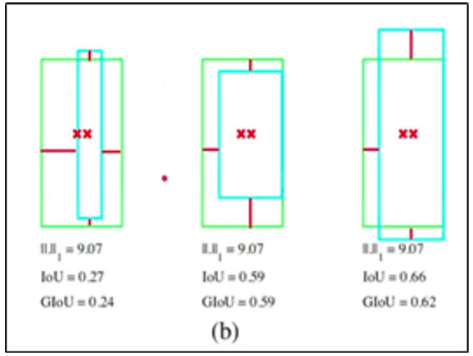
- Object Detection 분야에서 예측 Bounding Box 와 Ground Truth 가 일치하는 정도를 0과 1 사이의 값으로 나타낸 값
- 즉, 두 box 영역의 (교집합) / (합집합)
- Object Detection Loss function 에서 IoU 의 필요성
  - {1 - IoU} 를 loss 로 사용

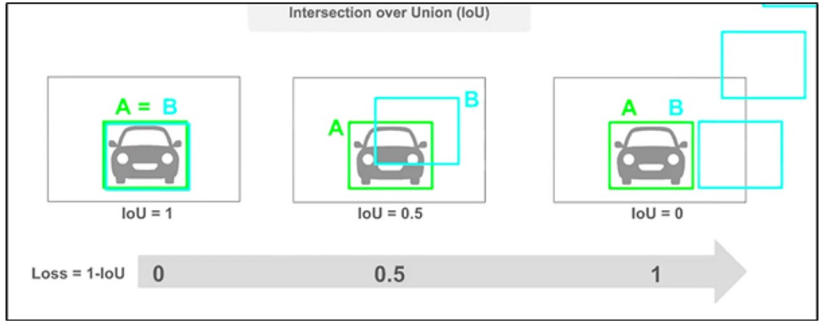
$$\mathcal{L}_{IoU} = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}.$$



- IoU 한계점
  - (a) L2 loss (=MSE) 는 일정하지만, box 의 겹침 정도를 나타내는 IoU 값은 변함
  - (b) L1 loss 는 일정하지만, IoU 값은 변함
  - 세 번째 예시처럼 두 박스가 겹치지 않을 경우, 어느 정도의 오차로 교집합이 생기지 않은 것인지 알 수 없어, gradient vanishing 문제 발생

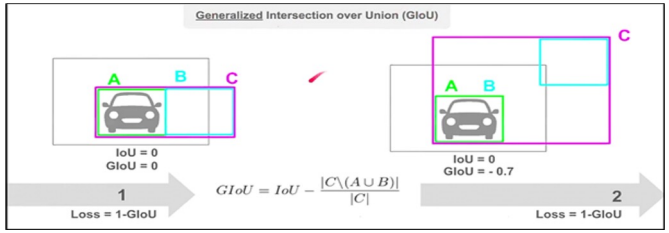






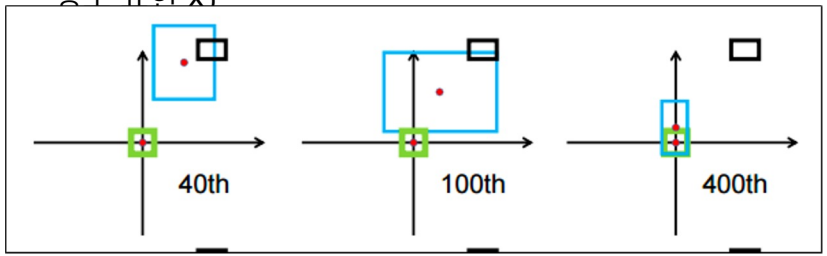
GIoU (Generalized-IoU)

- Bbox 와 GT 를 모두 포함하는 최소 크기의 C 박스를 활용
- C box 는 A와 B box를 포함하는 가장 작은 box, C\(AUB) 는 C box 영역에서 A와 B box 의 합집합을 뺀 영역.
- 기존 IoU 값에서 C box 중 A, B 모두와 겹치지 않는 영역의 비율을 뺀 값이 GIoU
  - GIoU 는 클수록 좋음



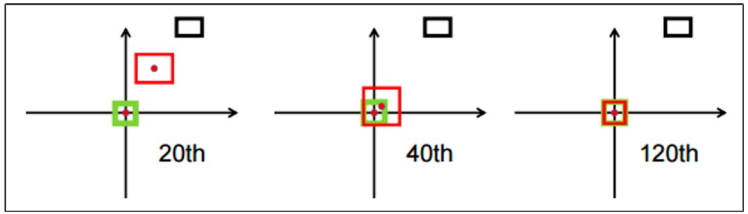
- 왼쪽 예시는 IoU=0, C＼(A ∪ B) = 0 → GIoU=0
- 오른쪽 예시는 IoU=0, C＼(A ∪ B) = 0.7 → GIoU=-0.7

- GIoU 한계점 (1-GIoU 를 loss 로 사용하는 경우)
  - Iteration 에 따른 GIoU loss 의 bounding box 예측 과정 (초록: GT, 파랑:Bbox)
    - GT와의 overlap 을 위해 Bbox 영역이 넓어지고, overlap 된 후 IoU 를 높이기 위해 Bbox 영역을 줄이는 방식으로 수행
  - 겹치지 않는 박스에 대한 gradient vanishing 문제는 개선했지만 수렴 속도가 느리고 부정확하게 박스 예측 (horizontal, vertical 정보 표현 X)



DIoU (Distance-IoU)

- IoU 와 함께 중심 좌표 활용
- Iteration 에 따른 DIoU loss 의 bbox 예측 과정 (초록: GT, 빨강:Bbox) 을 보면, GT 와의 overlap 을 위해 Bbox 영역을 넓히는 GIoU 와 달리 DIoU 는 중심 좌표를 비교하여 Bbox 자체가 GT 쪽으로 이동



$$\mathcal{R}_{DIoU} = \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2},$$

Equation 5

$$\mathcal{L}_{DIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2}.$$

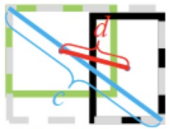


Figure 5: DIoU loss for bounding box regression, where the normalized distance between central points can be directly minimized. c is the diagonal length of the smallest enclosing box covering two boxes, and $d = \rho(\mathbf{b}, \mathbf{b}^{gt})$ is the distance of central points of two boxes.

- Loss (DIoU) 는 기존 IoU loss 에 중심 좌표를 고려하는 수식을 추가함.
- Notation
  - ρ는 Euclidean distance
  - b 는 Bbox 의 중심 좌표, b^gt 는 GT의 중심 좌표
  - c 는 Bbox 와 GT 를 포함하는 최소 박스인 C 박스의 대각 길이

DIoU (Distance-IoU)

- Object detection 에서 DIoU 를 loss 로 사용하는 경우
  - (왼쪽 이미지) 두 box 가 겹쳐진 영역의 크기가 동일하고 Bbox 의 위치만 달라졌을 경우, IoU, GIoU 는 Bbox 의 위치를 고려하지 않아 Loss 값이 변하지 않음. 중심 좌표를 활용하는 DIoU 는 해당 좌표가 변함에 따라 loss 값도 변함.
  - (오른쪽 이미지) 두 box 가 겹치지 않았을 때, GIoU 처럼 Bbox 의 영역을 넓히지 않고 중심 좌표를 통해 박스의 거리 차이를 최소화 함으로써 수렴 속도 향상. DIoU 를 NMS 에 적용 했을 때 동일한 class 의 GT 가 여러 개 겹쳐 있는 경우, IoU 에 비해 robust 함

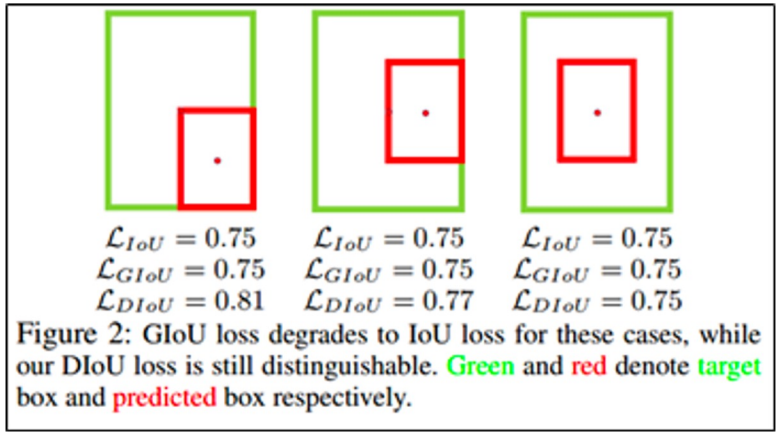
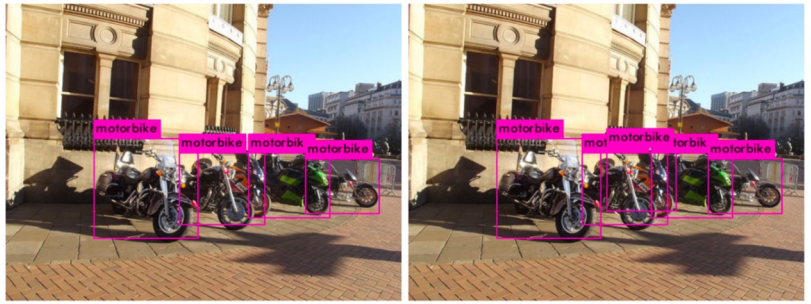


Figure 2: GIoU loss degrades to IoU loss for these cases, while our DIoU loss is still distinguishable. Green and red denote target box and predicted box respectively.



$\mathcal{L}_{CIoU}$ + NMS $\mathcal{L}_{CIoU}$ + DIoU-NMS

Figure 8: Detection example from MS COCO 2017 using YOLO v3 (Redmon and Farhadi 2018) trained on PASCAL VOC 07+12.

CIoU (Complete-IoU)

- DIoU 를 제안한 논문에서 동시에 제안된 방법
- Bbox 에 대한 좋은 loss 는 1) overlap area (겹치는 부분), 2) central point distance (중심점 사이의 거리), 3) aspect ratio (종횡비; 높이 너비 비율) 세 요소를 고려해야 함. 따라서 overlap area 와 central point distance 를 고려하는 DIoU 에 추가적으로 aspect ratio 를 고려하는 CIoU 를 제안함.

Therefore, based on DIoU loss, the CIoU loss is proposed by imposing the consistency of aspect ratio,

$$\mathcal{R}_{CIoU} = \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v, \tag{8}$$

Then the loss function can be defined as

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v. \tag{10}$$

또한, α는 다음과 같이 정의되며 겹치는 영역의 값이 높은 우선순위를 갖게 하여, 겹치지 않았을 때 더 빠른 수렴을 가능하게 한다.

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \qquad \alpha = \frac{v}{(1-IoU)+v},$$

Alpha-IoU

- 성능
  1) Can surpass existing IoU-based losses by a noticeable performance margin
  2) Offer detectors more flexibility in achieving different levels of bbox regression accuracy by modulating *alpha*
  3) More robust to small datasets and noisy bboxes

3.2 α-IoU Losses

The vanilla IoU loss is defined as $\mathcal{L}_{\text{IoU}} = 1 - IoU$. We first apply the Box-Cox transformation[3] [2] and generalize the IoU loss to an α-IoU loss:

$$\mathcal{L}_{\alpha\text{-IoU}} = \frac{1 - IoU^{\alpha}}{\alpha}, \alpha > 0. \tag{1}$$

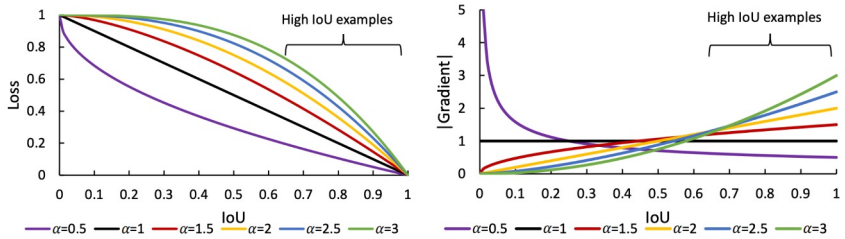


Figure 1: Correlation between IoU and $\mathcal{L}_{\alpha\text{-IoU}} = 1 - IoU^{\alpha}$ (left) and its absolute gradient $|\nabla_{\text{IoU}}\mathcal{L}_{\alpha\text{-IoU}}|$ (right) with different $\alpha \in [0.5, 3]$. According to both plots, $\mathcal{L}_{\alpha\text{-IoU}}$ reweights all objects adaptively and distinctively for $0 < \alpha < 1$ vs. $\alpha > 1$ ($\alpha = 1$ marks the IoU loss).

- Figure 1
  - (왼쪽) Loss_Alpha-IoU
  - (오른쪽) Gradient 절대값
- alpha 1 이하와 이상으로 다른 경향을 보임.
- alpha = 1 인 loss IoU loss 와 동치

Alpha-IoU

- alpha-IoU formula generalize

With the above α-IoU formula, we can now generalize the commonly used IoU-based losses including $\mathcal{L}_{\text{IoU}}$, $\mathcal{L}_{\text{GIoU}}$, $\mathcal{L}_{\text{DIoU}}$, and $\mathcal{L}_{\text{CIoU}}$ using the same power parameter α for the IoU and penalty terms:

$$
\begin{aligned}
\mathcal{L}_{\text{IoU}} = 1 - IoU &\Longrightarrow \mathcal{L}_{\alpha\text{-IoU}} = 1 - IoU^{\alpha}, \\
\mathcal{L}_{\text{GIoU}} = 1 - IoU + \frac{|C \setminus (B \cup B^{gt})|}{|C|} &\Longrightarrow \mathcal{L}_{\alpha\text{-GIoU}} = 1 - IoU^{\alpha} + (\frac{|C \setminus (B \cup B^{gt})|}{|C|})^{\alpha}, \\
\mathcal{L}_{\text{DIoU}} = 1 - IoU + \frac{\rho^2(\boldsymbol{b}, \boldsymbol{b}^{gt})}{c^2} &\Longrightarrow \mathcal{L}_{\alpha\text{-DIoU}} = 1 - IoU^{\alpha} + \frac{\rho^{2\alpha}(\boldsymbol{b}, \boldsymbol{b}^{gt})}{c^{2\alpha}}, \\
\mathcal{L}_{\text{CIoU}} = 1 - IoU + \frac{\rho^2(\boldsymbol{b}, \boldsymbol{b}^{gt})}{c^2} + \beta v &\Longrightarrow \mathcal{L}_{\alpha\text{-CIoU}} = 1 - IoU^{\alpha} + \frac{\rho^{2\alpha}(\boldsymbol{b}, \boldsymbol{b}^{gt})}{c^{2\alpha}} + (\beta v)^{\alpha},
\end{aligned}
\tag{4}
$$

Alpha-IoU

- Experiments

Table 1: The performance of YOLOv5s, YOLOv5x and DETR models trained using different localization losses on PASCAL VOC and MS COCO benchmarks. Results are obtained on the test set of PASCAL VOC 2007 and the val set of MS COCO 2017. mAP denotes $mAP_{50:95}$; $mAP_{75:95}$ denotes the mean AP over $AP_{75}, AP_{80}, \cdots, AP_{95}$. "rela. improv." stands for the relative improvement. $\alpha = 3$ is used for all α-IoU losses in all experiments.

Method	Loss	PASCAL VOC						MS COCO					
		AP_{50}	AP_{75}	AP_{85}	AP_{95}	**mAP**	**$mAP_{75:95}$**	AP_{50}	AP_{75}	AP_{85}	AP_{95}	**mAP**	**$mAP_{75:95}$**
YOLOv5s	$\mathcal{L}_{IoU}$	78.81	58.04	35.07	2.34	52.74	32.45	55.51	38.59	23.58	2.07	36.29	21.82
	$\mathcal{L}_{\alpha\text{-}IoU}$	78.62	58.78	38.16	3.64	53.61	34.46	55.25	39.69	25.85	3.35	37.01	23.66
	rela. improv.	**-0.24%**	**1.27%**	**8.81%**	**55.56%**	**1.65%**	**6.21%**	**-0.47%**	**2.85%**	**9.63%**	**61.84%**	**1.98%**	**8.43%**
	$\mathcal{L}_{DIoU}$	78.19	57.77	34.89	2.36	52.30	32.17	55.67	39.01	23.56	2.03	36.36	21.95
	$\mathcal{L}_{\alpha\text{-}DIoU}$	78.33	59.24	38.46	3.50	53.76	34.66	55.84	39.49	25.49	3.30	36.74	23.34
	rela. improv.	**0.18%**	**2.54%**	**10.23%**	**48.31%**	**2.79%**	**7.72%**	**0.31%**	**1.23%**	**8.19%**	**62.56%**	**1.05%**	**6.32%**
YOLOv5x	$\mathcal{L}_{IoU}$	85.24	70.08	53.08	10.88	63.95	46.78	67.36	52.15	38.22	9.31	48.42	34.42
	$\mathcal{L}_{\alpha\text{-}IoU}$	84.83	70.20	53.75	13.74	64.25	48.06	67.72	52.61	38.62	9.76	48.67	34.72
	rela. improv.	**-0.48%**	**0.17%**	**1.26%**	**26.29%**	**0.47%**	**2.73%**	**0.53%**	**0.88%**	**1.05%**	**4.83%**	**0.52%**	**0.87%**
	$\mathcal{L}_{DIoU}$	85.04	71.05	53.71	11.11	64.21	47.30	67.54	52.03	38.02	8.58	48.38	34.16
	$\mathcal{L}_{\alpha\text{-}DIoU}$	84.90	71.34	54.23	13.85	64.49	48.40	67.42	52.65	39.28	10.29	48.81	35.42
	rela. improv.	**-0.16%**	**0.41%**	**0.97%**	**24.66%**	**0.44%**	**2.32%**	**-0.18%**	**1.19%**	**3.31%**	**19.93%**	**0.89%**	**3.68%**
DETR	$\mathcal{L}_{IoU}$	76.50	53.85	29.54	1.62	49.78	28.82	59.38	41.67	26.13	3.52	39.23	24.37
	$\mathcal{L}_{\alpha\text{-}IoU}$	76.22	55.03	32.30	2.28	51.12	31.08	59.61	42.65	28.57	5.09	40.18	26.44
	rela. improv.	**-0.37%**	**2.19%**	**9.34%**	**40.74%**	**2.69%**	**7.84%**	**0.39%**	**2.35%**	**9.34%**	**44.60%**	**2.42%**	**8.49%**
	$\mathcal{L}_{DIoU}$	76.26	54.09	29.23	1.56	49.91	28.68	59.28	41.62	26.09	3.54	39.25	24.48
	$\mathcal{L}_{\alpha\text{-}DIoU}$	76.44	54.89	31.48	2.44	50.96	30.60	59.38	42.34	28.23	5.36	39.94	26.05
	rela. improv.	**0.24%**	**1.48%**	**7.70%**	**56.41%**	**2.10%**	**6.69%**	**0.17%**	**1.73%**	**8.20%**	**51.41%**	**1.76%**	**6.41%**

Table 2: The performance of Faster R-CNN (ResNet-50-FPN) with $1\times$ schedule and single scale training on MS COCO using different localization losses. Results are obtained on the val set of MS COCO 2017. mAP denotes $mAP_{50:95}$; $mAP_{75:95}$ denotes the mean AP over $AP_{75}, AP_{80}, \cdots, AP_{95}$. AP_s, AP_m, and AP_l denote the AP for small, medium, and large objects, respectively. † marks the reproduced results from the MMDetection toolbox [6], while * marks the results in the original papers. "–" represents the missing results in papers. $\alpha = 3$ is used for all α-IoU losses in all experiments. The top two best results in every column are **boldfaced**.

Loss	AP_{50}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}	**mAP**	**$mAP_{75:95}$**	AP_s	AP_m	AP_l
$^\dagger\ell_1$	58.13	40.45	33.56	23.39	11.09	1.24	37.37	21.95	21.20	40.96	48.13
$^\dagger\mathcal{L}_{IoU}$	58.12	41.23	34.03	24.43	12.42	1.61	37.88	22.74	21.61	41.63	49.11
$^\dagger\mathcal{L}_{GIoU}$	58.18	41.00	33.52	24.13	11.97	1.51	37.62	22.43	21.49	41.07	48.90
$^\dagger\mathcal{L}_{BIoU}$	58.05	40.57	33.54	23.85	11.10	1.19	37.43	22.05	21.57	41.00	48.17
$^*\mathcal{L}_{IoU}$	–	40.79	–	–	–	–	37.93	–	21.58	40.82	50.14
$^*\mathcal{L}_{GIoU}$	–	41.11	–	–	–	–	38.02	–	21.45	41.06	50.21
$^*\mathcal{L}_{DIoU}$	–	41.11	–	–	–	–	38.09	–	21.66	41.18	50.32
$^*\mathcal{L}_{CIoU}$	–	41.96	–	–	–	–	38.65	–	21.32	41.83	**51.51**
$^*\mathcal{L}_{Focal\text{-}EIoU}$	**59.10**	**42.40**	–	–	–	–	38.90	–	21.20	41.10	50.20
*Autoloss	58.60	41.80	–	–	–	–	38.50	–	22.00	**42.20**	50.20
$\mathcal{L}_{\alpha\text{-}IoU}$	58.81	41.94	34.81	**25.36**	**13.27**	1.81	38.96	23.44	**22.14**	42.11	50.36
$\mathcal{L}_{\alpha\text{-}GIoU}$	59.01	42.00	**35.13**	25.14	13.09	**2.03**	39.18	**23.46**	22.05	**42.19**	50.08
$\mathcal{L}_{\alpha\text{-}DIoU}$	**59.27**	**42.18**	**35.25**	**25.47**	**13.32**	1.95	**39.43**	**23.65**	**22.10**	42.10	**50.43**
$\mathcal{L}_{\alpha\text{-}CIoU}$	59.09	41.92	35.01	25.08	13.04	**1.98**	**39.25**	23.41	21.94	41.88	50.01

Alpha-IoU

- Experiments

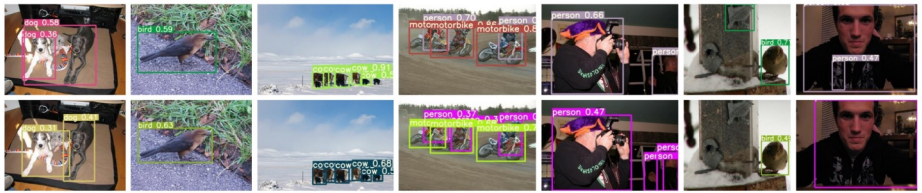


Figure 4: Example results on the test set of PASCAL VOC 2007 using YOLOv5s trained by $\mathcal{L}_{\text{IoU}}$ (top row) and $\mathcal{L}_{\alpha\text{-IoU}}$ with $\alpha = 3$ (bottom row). $\mathcal{L}_{\alpha\text{-IoU}}$ performs better than $\mathcal{L}_{\text{IoU}}$ because it can localize objects more accurately (image 1 and 2), thus can detect more true positive objects (image 3 to 5) and fewer false positive objects (image 6 and 7).

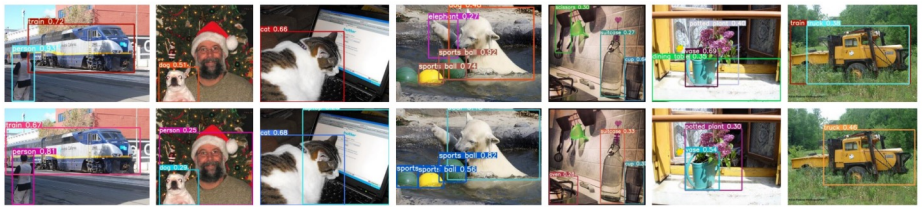


Figure 5: Example results on the val set of MS COCO 2017 using YOLOv5s trained by $\mathcal{L}_{\text{IoU}}$ (top row) and $\mathcal{L}_{\alpha\text{-IoU}}$ with $\alpha = 3$ (bottom row). $\mathcal{L}_{\alpha\text{-IoU}}$ performs better than $\mathcal{L}_{\text{IoU}}$ because it can localize objects more accurately (image 1), thus can detect more true positive objects (image 2 to 5) and fewer false positive objects (image 4 to 7). Note that $\mathcal{L}_{\alpha\text{-IoU}}$ detects both more true positive and fewer false positive objects in image 4 and 5 than $\mathcal{L}_{\text{IoU}}$.

1. object detection accurately
2. More True Positive objects
3. Fewer False Positive objects

ICIoU

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 13, 2021, accepted July 22, 2021, date of publication July 26, 2021, date of current version August 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100414

ICIoU: Improved Loss Based on Complete Intersection Over Union for Bounding Box Regression

XUFEI WANG[1,2] AND JEONGYOUNG SONG[2], (Member, IEEE)

[1]Key Laboratory of Industrial Automation, School of Mechanical Engineering, Shaanxi University of Technology, Hanzhong 723000, China
[2]Department of Computer Engineering, Pai Chai University, Daejeon 35345, South Korea

Corresponding author: Jeongyoung Song (jysong@pcu.ac.kr)

This work was supported in part by the Shaanxi Provincial Key Laboratory of Industrial Automation Research Program under Grant 18JS020.

ABSTRACT An object detector based on convolutional neural network (CNN) has been widely used in the field of computer vision because of its simplicity and efficiency. The average accuracy of CNN model detection results in the object detector is greatly affected by the loss function. The precision of the localization algorithm in the loss function is the main factor affecting the result. Based on the complete intersection over union (CIoU) loss function, an improved penalty function is proposed to improve the localization accuracy. Specifically, the algorithm more comprehensively considers matching bounding boxes between prediction with ground truth, using the proportional relationship of the aspect ratio from both bounding boxes. Under the same aspect ratio of the two bounding boxes, the influence factors of the prediction box on localization accuracy were considered. In this way, the function of the penalty function is strengthened, and localization accuracy of the network model improved. This loss function is called Improved CIoU (ICIoU). Experiments on the Udacity, PASCAL VOC, and MS COCO datasets have demonstrated the effectiveness of ICIoU in improving localization accuracy of network models by using the one-stage object detector YOLOv4. Compared with CIoU, the proposed ICIoU improved average precision (AP) by 0.57% and AP75 by 0.12% on Udacity, AP by 0.26% and AP75 by 1.28% on PASCAL VOC, and AP by 0.06% and AP75 by 0.65% on MS COCO.

INDEX TERMS Bounding box regression, localization accuracy, loss function, object detection.

SIoU (SCYLLA-IoU)

- Penalty metrics were redefined.
- Considering **the angle of the vector between the desired regression.**
- Results
  - CNN 과 데이터셋에 적용했을 때, **학습 속도와 inference accuracy 성능이 모두 향상.**
  - COCO-train/COCO-val results in improvements of +2.4% (mAP@0.5:0.95) and +3.6% (mAP@0.5) over other Loss Functions

- **Methods (SIoU loss function consists of 4 cost functions)**
  - Angle cost
  - Distance cost
  - Shape cost
  - IoU cost

SIoU (SCYLLA-IoU)

- **Methods (SIoU loss function consists of 4 cost functions)**

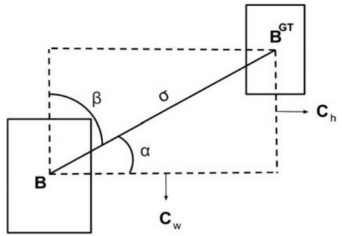


Figure 1. The scheme for calculation of angle cost contribution into the loss function.

Angle cost

$$\Lambda = 1 - 2 * sin^2\left(arcsin\,(x) - \frac{\pi}{4}\right),$$

Shape cost

The shape cost is defined as:

$$\Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta}$$

where

$$\omega_w = \frac{|w - w^{gt}|}{max(w,w^{gt})},\ \omega_h = \frac{|h - h^{gt}|}{max(h,h^{gt})}$$

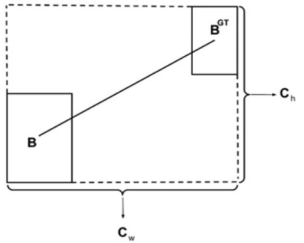


Figure 3. Scheme for calculation of the distance between the ground truth bounding box and the prediction of it.

Distance cost

Distance cost

The distance cost is redefined taking into account the angle cost defined above:

$\Delta = \sum_{t=x,y}(1 - e^{-\gamma\rho_t})$,

where

$$\rho_x = \left(\frac{b_{cx}^{gt} - b_{cx}}{c_w}\right)^2,\ \rho_y = \left(\frac{b_{cy}^{gt} - b_{cy}}{c_h}\right)^2,\ \gamma = 2 - \Lambda$$

Loss function

Finally let's define loss function

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2}$$

where

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|}$$

SIoU (SCYLLA-IoU)

- Results

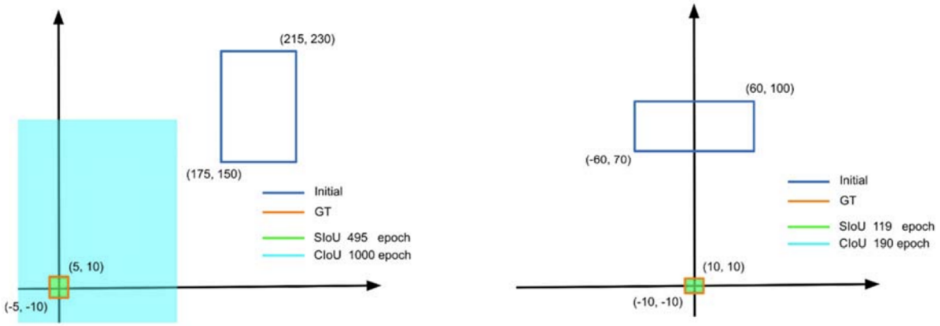


Figure 6. Example of simulation showing the convergence of boxes that are placed on axes versus the boxes that are further from axes. Clearly SIoU approach.

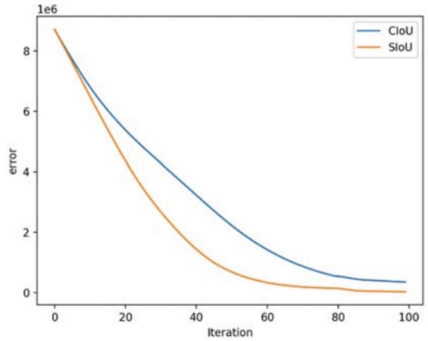


Figure 8. Plot of the errors from CIoU and SIoU losses through the training iterations.

SIoU (SCYLLA-IoU)

- Results

Table 1. Comparison of mAP metrics for Scylla-Net trained with CIoU loss, SIoU and SIoU applied to larger Scylla model.

Network/Loss	mAP@0.5	mAP@0.5:0.95
Scylla-Net-S/CIoU	66.4%	50.3%
Scylla-Net-S/SIoU	70.0%	52.7%
Scylla-Net-L/SIoU	74.3%	57.1%

Figure 11. Comparison of detections by Scylla-Net + SIoU and other models.

Multimodal Masked Autoencoders Learn Transferable Representations

Xinyang Geng[1]* Hao Liu[1,2]*,† Lisa Lee[2]
Dale Schuurmans[2] Sergey Levine[1] Pieter Abbeel[1]
[1]UC Berkeley, [2]Google Brain
*Equal contribution, listing is random †Project lead

https://arxiv.org/pdf/2205.14204.pdf

- **Multimodal data remains an open challenge**
  - V-L data, 각 모달리티 encoder 를 학습함으로서 contrastive learning approach 사용
    - data augmentation에 의해 sampling bias 존재
    - 해당 bias가 performance degrading 유발.
  - Image-text data, 범용적으로 사용하기에 unpaired data 가 많음 ⇒ contrastive learning approach를 사용하기에 민감

- We investigate whether a large multimodal model trained **purely via masked token prediction**
  - without using modality-specific encoders or contrastive learning
  - can learn transferable representations for downstream tasks

- **간단하고 확장가능한** Multimodal Network Architecture 제안
    - ⇒ **MultiModal Masked Autoencoder (M3AE)**

M3AE

- **Masked token prediction**을 통해 vision and language data 를 위한 encoder 학습
- Large-scale image text dataset 에 대해서 pretrained, **downstream tasks에 잘 generalizable 확인.**
- Scalability with large model size and training time **(= flexibility)**
- Image와 Language 사이로 부터 얻을수 있는 meaningful information 을 학습을 하는 것을 **qualitative analysis**
- 기존 Image-text multimodal approach?
    - CLIP, ALIGN 처럼 결국 Contrastive learning 사용 ⇒ **unpair data를 범용적으로 사용하기에 major limitation 존재.**
- To address this limitation, how to do it?
    - **Based on MAE , M3AE is trained purely via masked token prediction**
    - simple, scalable
    - **with using modality-specific encoder or contrastive learning approach**

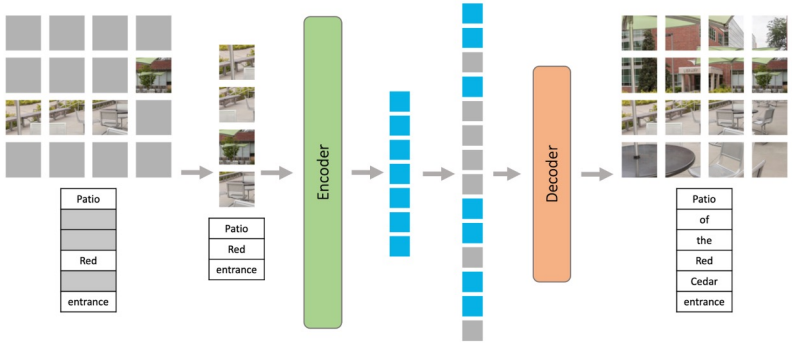


Figure 1: Multimodal masked autoencoder (M3AE) consists of an encoder that maps language tokens and image patches to a shared representation space, and a decoder that reconstructs the original image and language from the representation.

Multimodal Masked autoencoder (M3AE)

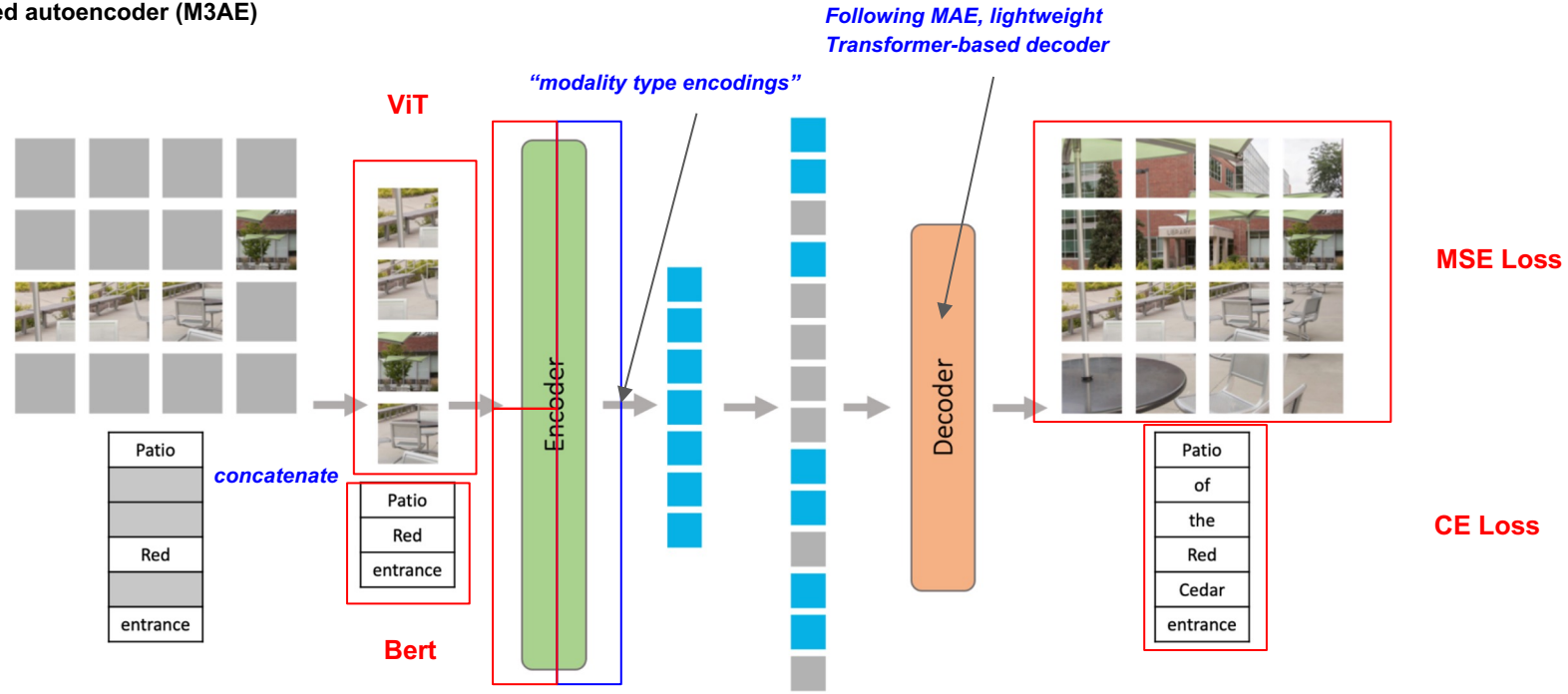


Figure 1: Multimodal masked autoencoder (M3AE) consists of an encoder that maps language tokens and image patches to a shared representation space, and a decoder that reconstructs the original image and language from the representation.

Multimodal Masked autoencoder (M3AE)

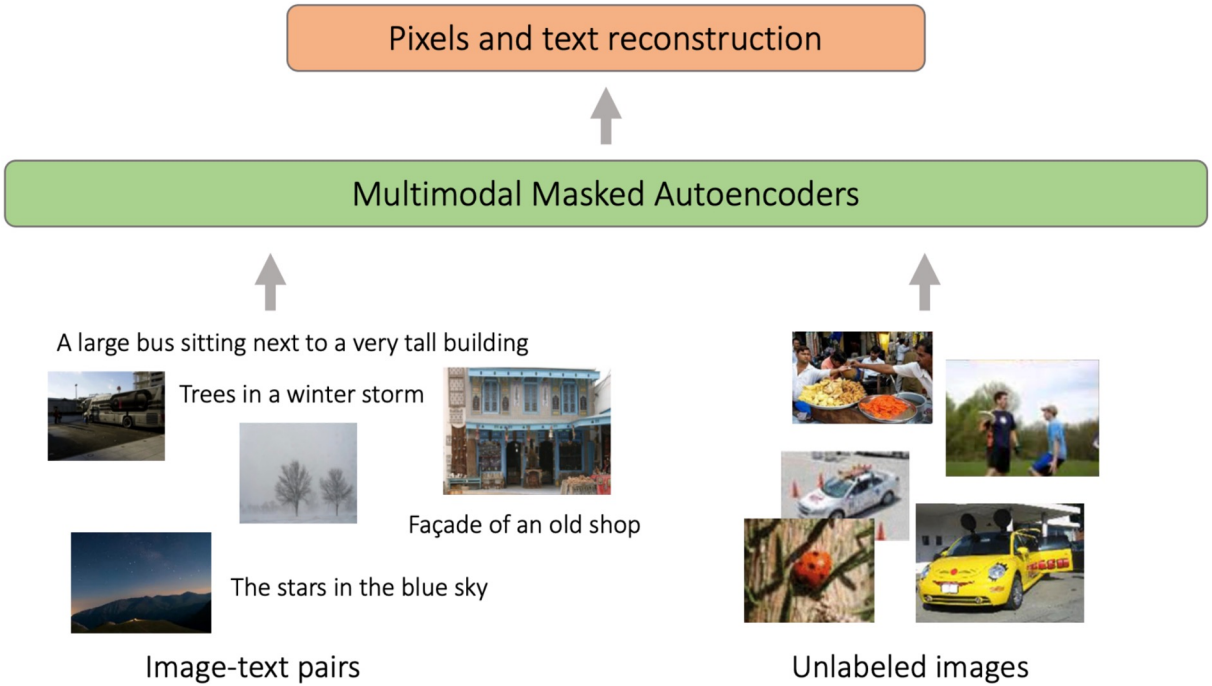


Figure 2: M3AE can learn representations from a flexible mixture of image-text pairs and unpaired images using a unified model without relying on data augmentations.

Multimodal Masked Autoencoders Learn Transferable Representations

Results

- M3AW is able to learn generalizable representations that transfer well to downstream tasks such as **image classification and OOD detection**
- Our strong results for M3AE demonstrate the **generalization** benefits of multimodal training for learning transferable representations **across datasets.**
- we find that M3AE performs best when we apply **a high mask ratio (75%)** on language while in contrast, language models like BERT conventionally use a low mask ratio (15%)
    - language data are highly semantic and information-dense
- We also provide qualitative analysis showing that the learned representation incorporates meaningful information from both image and language

Experiments

- **Can M3AE learn generalizable visual representations that transfer?**
- **Does the learned representation incorporate meaningful information from both images and language?**
- **Does M3AE scale well with model size and training time?**

- **Dataset**
    - pretrained on multimodal CC12M dataset
    - ImageNet-1k linear classification benchmark (compared to pre-training on MAE)

Model	MAE	M3AE	CLIP	Supv
Accuracy	45.2	59.4	69.0	81.8
M3AE text ratio	10%	20%	30%	100%
Accuracy	53.3	54.0	54.5	59.4

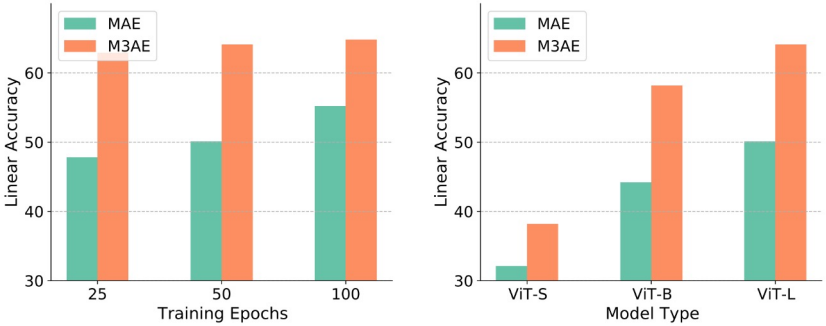
Results of linear classification ⇒

- *A lower percentage of paired image-text data contains less information*
- *model has to infer the relation between visual and language concepts based on limited paired data.*

Figure 3: Comparison of M3AE, MAE, and CLIP on ImageNet. M3AE significantly outperforms MAE. M3AE can flexibly leverage a combination of paired image-text data and unpaired image only data. All models are ViT-B. MAE and M3AE are pretrained on CC12M for 100 epochs.

Experiments

- linear classification AUC: MAE < M3AE with every training epoch (VIT-B/16) & model Type (Flexibility).
- high text mask ratio of 75% is best performance.



Epoch Scaling	25	50	100	Model Scaling	ViT-S	ViT-B	ViT-L
M3AE	62.9	64.1	64.8	M3AE	38.2	58.2	64.1
MAE	47.8	50.1	55.2	MAE	32.1	44.2	50.1

Figure 4: Left. ViT-L/16 with longer pre-training schedules (25/50/100 epochs). **Right**. Comparing ViT model variants of different capacities (ViT-S/B/L). All models are pre-trained for 50 epochs. We see that M3AE scales well with model size and training epochs, outperforming MAE in every setting.

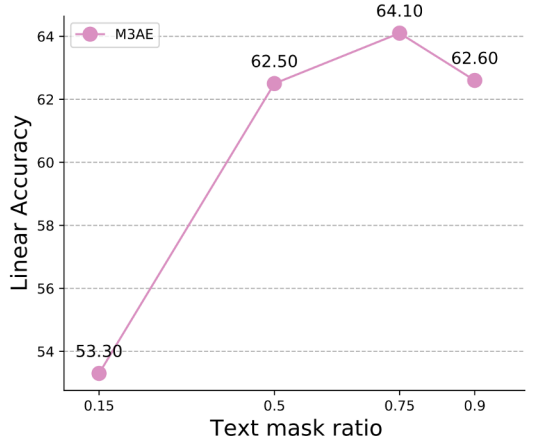


Figure 5: Comparing M3AE with different text mask ratio. We see that M3AE performs the best with a surprisingly high text mask ratio of 75%.

Experiments

- Mahalanobis outlier score
- CIFAR 100 (In dist.) vs CIFAR 10 (out dist.)

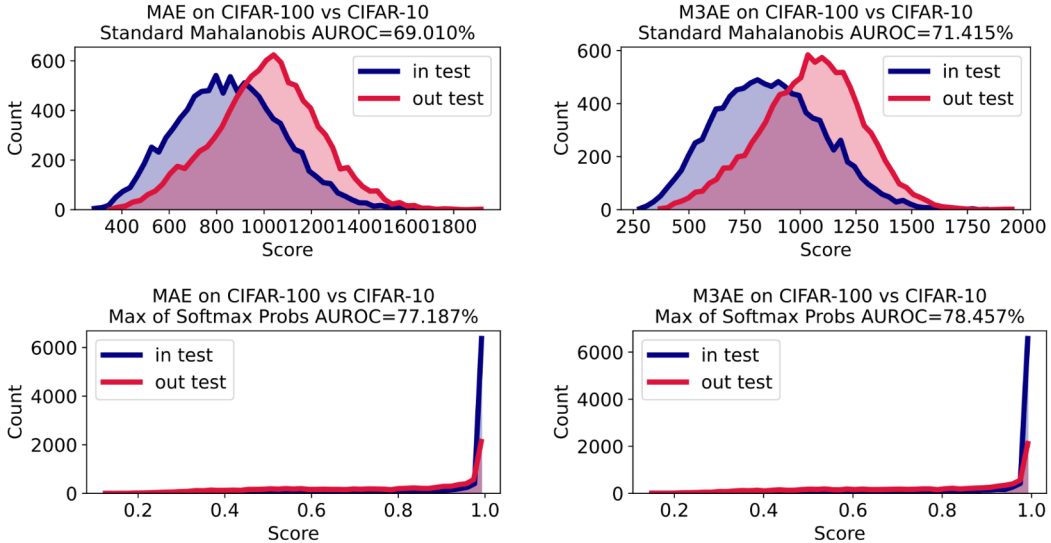


Figure 6: Out-of-distribution detection results on CIFAR-100 (in-distribution) and CIFAR-10 (out-of-distribution). **Upper** shows results based on Mahalanobis outlier score, M3AE achieves *71.4%* which is higher than MAE's *69.0%*. **Lower** shows results based on max over softmax score, M3AE achieves *78.5%* which is also higher than MAE's *77.2%*.

Experiments

- Visualization of cross-modal attention weights (visualize the **M3AE encoder attention between a given text token and all image patches**)

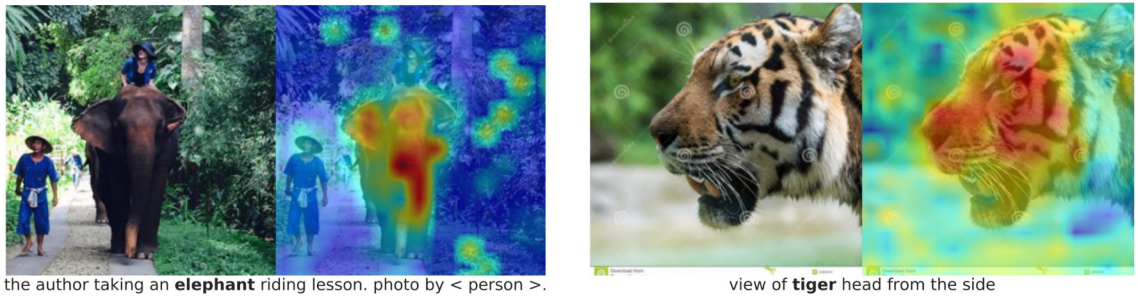


Figure 7: Visualization of attention between a given text token and image patches on CC12M dataset. The text token for which we visualize the attention is bolded. We see that the M3AE encoder is able to attend to the correct objects.

Figure 8: Visualization of attention between a given image patch and all text tokens on CC12M dataset The highlighted rectangle is the image patch for which we visualize the attention. Denser color of the text denotes higher attention. The visualization suggests that M3AE encoder is able to attend to the correct words corresponding to the image patch.

Experiments

- t-SNE visualization for learned representation of 10 classes on ImageNet validation set

Clustering analysis of representation. We perform t-SNE [40] visualizations of the learned representation of M3AE and MAE for 10 classes on ImageNet validation set in Figure 9. Compared to MAE, M3AE successfully clusters together images that correspond to the same semantic label.

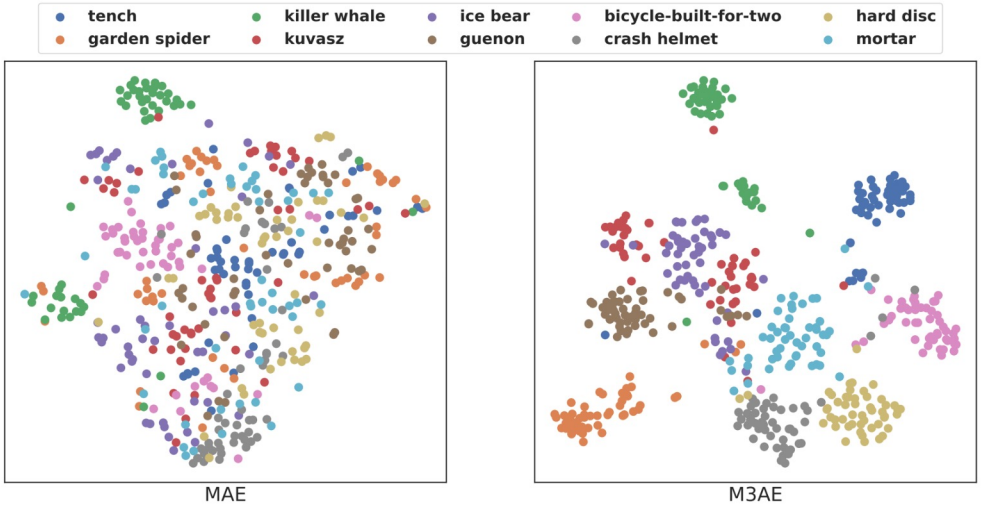


Figure 9: t-SNE visualization for learned representations of 10 classes on ImageNet validation set. Left is MAE and right is M3AE. The representation of M3AE clusters much stronger together with the semantic labels compared to MAE representations.

Multimodal Masked Autoencoders Learn Transferable Representations

Conclusion

- In this paper, M3AE를 통해 simple but effective model 제공
  - 기존 image-text의 접근방식인 multimodal representations contrastive objectives 를 더 이상 사용할 필요 없음.
  - downstream task에서도 well generalize 하도록 shared representations learning, Due to its flexibility and scalability.

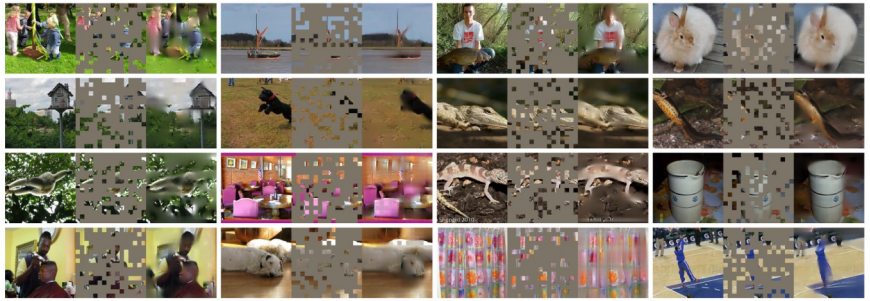


Figure 10: Masked image reconstruction on ImageNet validation images. For each triplet, we show the ground-truth (left), the masked image (mid) and our M3AE reconstruction (right).

ImageNet, CC12M dataset ⇒ Masked Image Reconstruction =>

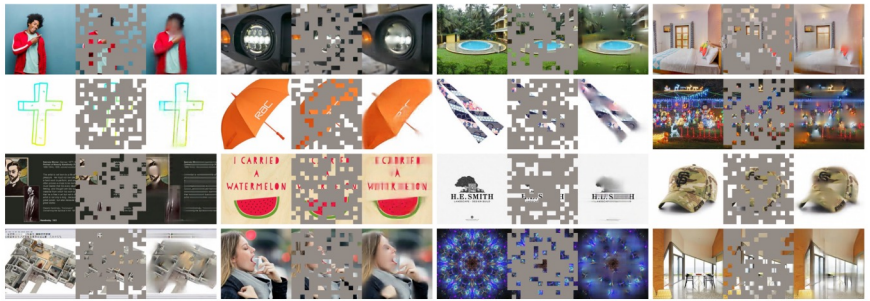


Figure 11: Masked image reconstruction on CC12M images. For each triplet, we show the ground-truth (left), the masked image (mid) and our M3AE reconstruction (right).

Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Daniel Flores-Araiza[1], Francisco Lopez-Tiro[1], Elias Villalvazo-Avila[1], Jonathan El-Beze[2], Jacques Hubert[2], Gilberto Ochoa-Ruiz[1], Christian Daul[3]

[1]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]CHU Nancy, Service d'urologie de Brabois, Nancy, France
[3]Centre de Recherche en Automatique de Nancy, Université de Lorraine, France

CVPR 2022 Workshop
https://arxiv.org/pdf/2206.00252.pdf

This paper investigates means for **learning part-prototypes (PPs)** that enable interpretable models and suggests a **classification for a kidney stone patch image** and provides explanations in a similar way as those used on the MCA method.

- **신장 결석 유형 식별 (Kidney Stones Classifier)** -> 형성 원인, 적절한 치료법의 조기 처방 -> 재발률을 줄일 수 있다.
- 신장 결석 유형 식별 방법: associated ex-vivo diagnosis (known as morpho-constitutional analysis, MCA) : 시간,비용,많은 경험
  - 머신 러닝 사용하면 높은 정확도를 산출, 설명할 수 없다.
  - 합리적인 증거를 기반으로 행동 방침을 제안하는 이해할 수 있는 computer-aided diagnosis(CAD)이 필요
  - **해석 가능한 모델을 가능하게 하는 part-prototypes (PPs)을 학습하기 위한 수단 조사**
  - **classification for a kidney stone patch image를 제안하고 MCA 방법에 사용되는 것과 유사한 방법으로 설명**

- **MCA**
  - **의사가 보고 결정하는 것.**
  - **그러기는 쉽지 않으니, 머신러닝을 활용해서 예측. 그리고 왜 그런 결과가 나왔는지 해석하자.**

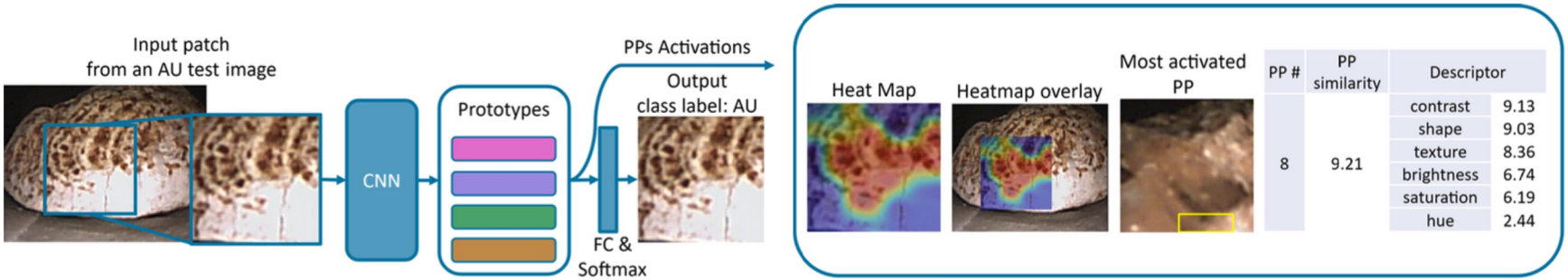
Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Learning part-prototypes (PPs) that enable interpretable models

- **interpretable models**
- 예측 결과 + 사람이 이해할 수 있는 형태로 추가적인 정보를 제공하도록 하는 머신러닝 알고리즘

prototype

- a data instance that is representative of all the data. (전체 데이터를 대표하는 data instance)
- part-prototypes 활용



(a) Model architecture

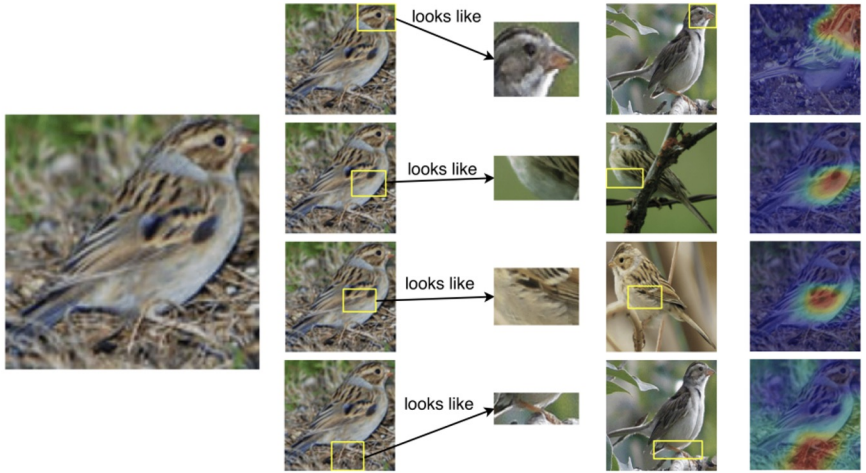
(b) Explanations

Figure 1: (a) Overall view of the proposal workflow, using ProtoPNet to obtain particular explanations for an input image. By use of PPs we provide explanations of the output classification, in tree different ways on (b), with a heatmap of the relevant parts on the input image, training images detected similar to the input, and measures of visual characteristics (descriptors) important for the activated PPs.

Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Materials and Methods

- [This Looks Like That: Deep Learning for Interpretable Image Recognition]
  - [2] ProtoPNet plus descriptors
- ProtoPnet 모델을 제안 + 사용한 이전 연구에서 제시된 방법론을 적용 [3,17]



Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

Figure 1: Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.

Materials and Methods

ProtoPNet 아키텍처는 표준 컨볼루션 신경망(CNN)으로 구성

- prototype layer
  - 미리 결정된 수의 클래스별 프로토타입으로 구성
  - 여기서는 클래스당 10개의 파트 프로토타입을 사용하며,
  - train을 위한 initialization and training procedure를 따르며,
  - ImageNet에서 사전 훈련된 VGG19(batch normalization)를 CNN 백본으로 사용
- a fully-connected layer.

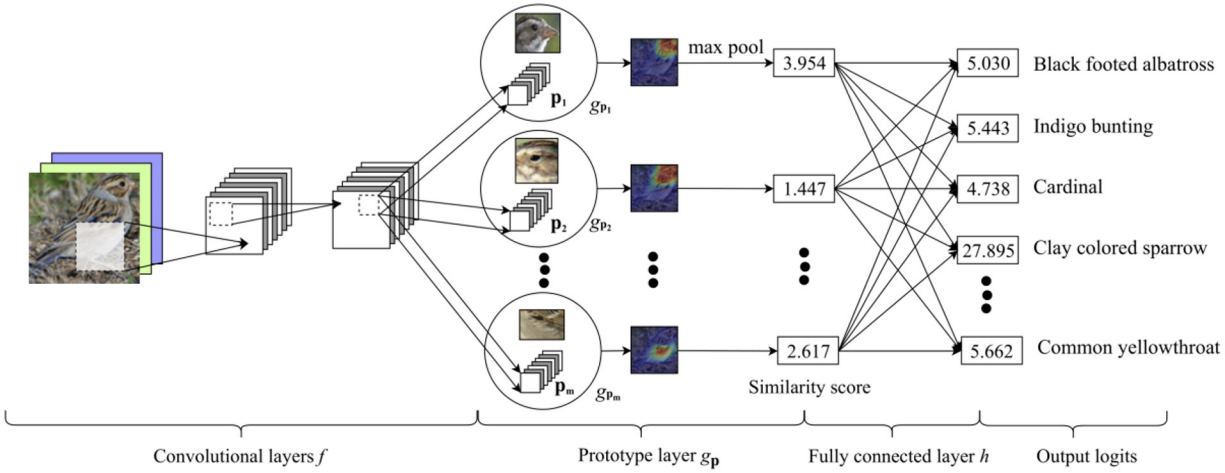


Figure 2: ProtoPNet architecture.

Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Materials and Methods

- [This Looks Like That: Deep Learning for Interpretable Image Recognition]

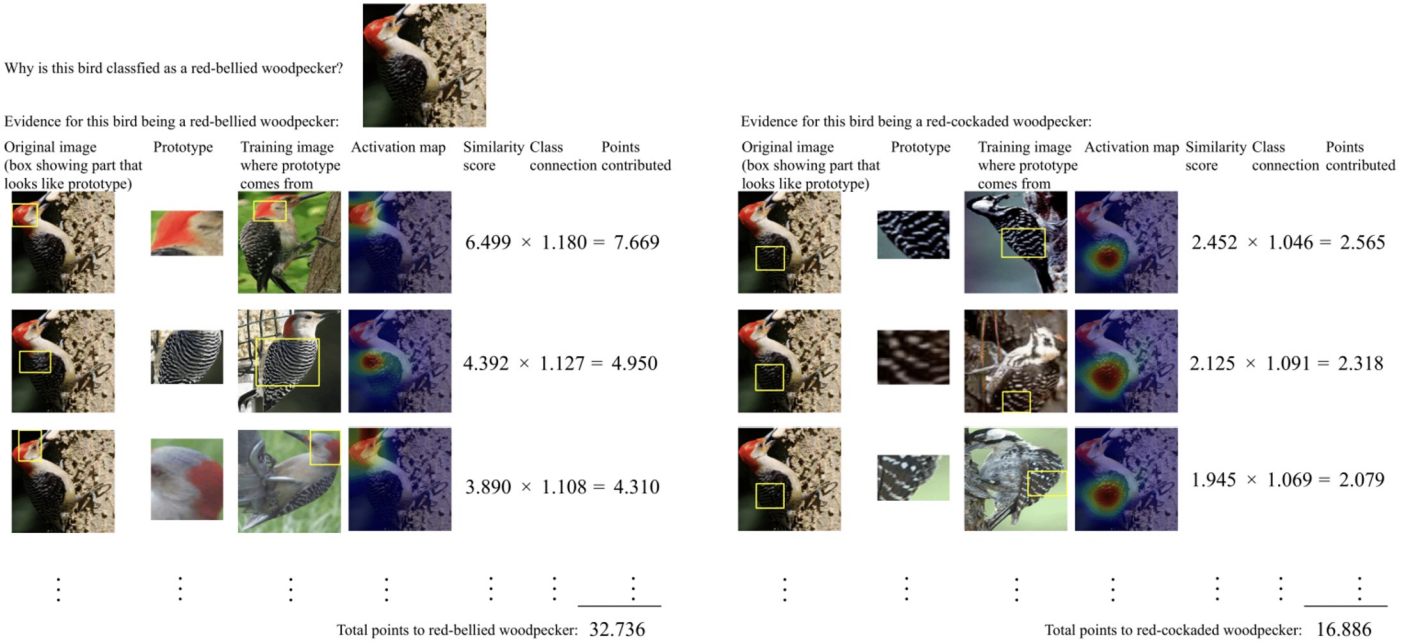


Figure 3: The reasoning process of our network in deciding the species of a bird (top).

Interpretable Deep Learning Classifier by Detection of Prototypical Parts on Kidney Stones Images

Introduction

MCA(=Morpho-Constitutional Analysis)
체외 샘플(내시경 수술 중 추출)의 신장 결석 유형 검사

- 돌의 시각적 특성에 대한 미시적 형태학적 검사
  - 분자 연구[4]와 밀접한 상관관계가 있으며 중요한 진단 정보를 보존할 수 있다.
  - (예: 표면 및 단면도의 크기, 형태, 색상, 질감 및 외관)
- 적외선 분광 광도 분석을 통해 석재의 결정 성분을 보다 정확하게 파악 가능

endoscopic (intra- operative) stone recognition (ESR) CAD

- 화면의 시각 보조 장치와 함께 내시경이 제공하는 비디오 신호 정보만을 기반으로 더 빠른 진단을 얻는 데 도움이 될 수 있다는 의견이 제시
- 비뇨기 내부의 신장 결석을 파편화하고 파괴할 수 있기 때문에 수술이 더 빠르고 traumatic이 적다.

자동화된 ESR을 수행하기 위한 딥러닝, 머신러닝 사용

- 기계 학습(ML) 모델
  - 신장 결석 보기(표면 및 단면)에서 특징(예: 색상 및 질감)의 효율적인 추출이 생체 내 내시경 영상에 대한 분류(정확도90%)에 상당한 영향
    (신장 결석의 형태학적 분석과 강한 상관 관계).
  - UMAP[13]와 같은 시각화에서 클러스터는 충분히 엄격하지 않음 -> 클러스터가 분류기에서 사용할 수 있는 최상의 기능이 아님
- 딥 러닝(DL) 기반 모델
  - classifier 및 UMAP의 엄격한 클러스터와 관련된 특징을 추출하는 데 우수한 결과(높은 정확도95%)
  - 추출한 특징에 대한 해석이 부족하여 이러한 모델은 임상 환경에서 그다지 유용하지 않다.

Interpretable Deep Learning Classifier by Detection of Prototypical Parts o

Dataset

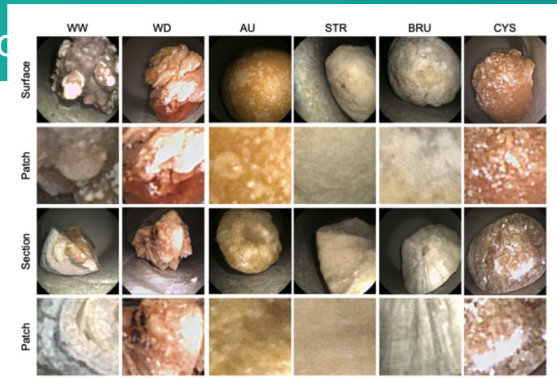
- **[1] Kidney stone dataset**

Ex-vivo (체외) 데이터 세트

- 305개의 신장 결석 이미지
- 비뇨기과 의사 Jonathan El Beze2가 수동으로 레이블을 부착

이 연구를 위해, ex-vivo 이미지 데이터 세트를 3가지로 분류해 사용

- 177개의 surface images,
- 128개의 section images
- third subset of 305 images of the six kidney stone types with the highest incidence
  - Urique(AU),
  - Brushite(BRU),
  - Cystine(CYS),
  - Struvite(STR),
  - Weddellite(WD),
  - Whewellite(WW).



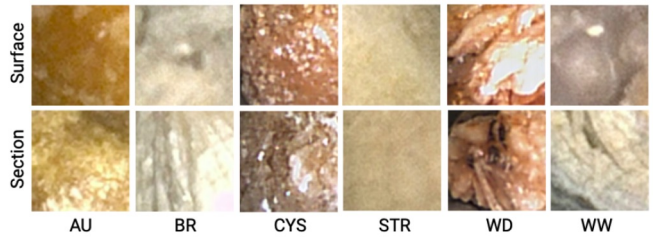


Figure 2: Examples of ex-vivo kidney stones generated patches.

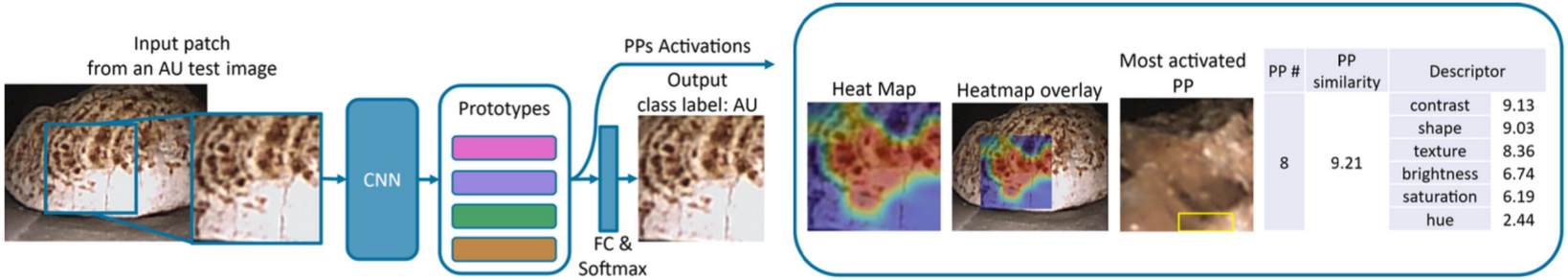
identification of kidney stones: 일반적으로 전체 이미지에 수행되지 않음

- train dataset (38400개 이미지 -> 1152000개 이미지)
  - 원본 이미지에서 200×200 픽셀의 패치 crop (최소 크기가 충분한 텍스처 및 색상 정보를 캡처할 수 있음)
    - 클래스(AU, BRU, CYS, STR, WD, WW) 및 뷰(surface, section, and mixed)당 총 2000개의 패치를 사용 가능
  - patch flipping, perspective distortions
  - The patches were also “whitened”
    - using the mean mi and standard deviation σi of the color values Ii
    - in each channel (Iiw = (Ii − miσi), with i = R, G, B).
- test dataset (9600개 이미지)

Materials and Methods
[2] ProtoPNet plus descriptors

Prototypical Part Network(ProtoPNet)를 사용

- 이미지의 여러 부분을 식별 가능
- 이미지의 일부가 특정 클래스의 learned part-prototypes (PPs) 처럼 보임 (그림 1).
- 이 유형의 모델은 이미지의 일부와 learned part-prototypes (PPs)사이의 combination of the similarity scores 에 대해 예측
- 이 capacity은 사용자가 비교적 쉽게 이해할 수 있는 예측을 산출하여 해석할 수 있게 함.



(a) Model architecture (b) Explanations

Figure 1: (a) Overall view of the proposal workflow, using ProtoPNet to obtain particular explanations for an input image. By use of PPs we provide explanations of the output classification, in tree different ways on (b), with a heatmap of the relevant parts on the input image, training images detected similar to the input, and measures of visual characteristics (descriptors) important for the activated PPs.

Results and Discussion

표 1: 백본 모델 및 AlexNet(참고 자료)의 평가 지표
ProtoPNet의 성능은 해당 해석 불가능한 백본 모델과 비교 가능

Table 1: Weighted average metrics comparison for section, surface, and mixed patches. ProtoPNet with VGG19 with batch normalization as the backbone (PPN-VGG19bn). VGG 19-layer model, configuration 'E', with batch normalization (VGG19bn).

Model	Accuracy			Precision			Recall			F1 score		
	Surface	Section	Mixed	Surface	Section	Mixed	Surface	Section	Mixed	Surface	Section	Mixed
PPN-VGG19bn	0.98	0.99	0.97	0.98	0.99	0.97	0.98	0.99	0.97	0.98	0.99	0.97
VGG19bn	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00
AlexNet	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97

Results and Discussion

그림 3: ProtoPNet의 각 출력 클래스에 대한 클래스 분리 가능성

출력 클래스의 별도 클러스터를 얻을 수 있다.
확대된 예제(파란색 원)에서 새로운 분류(노란색 점)의 투영이 동일한 등급(보라색 점)의 표본으로 둘러싸여 있음을 알 수 있으며,
이는 테스트 샘플의 정확한 분류에 대한 높은 확실성을 나타냄

동일한 클래스의 샘플로 둘러싸인 새로운 classification의 경우에 추가적인 global insight을 얻으며,
이는 정확한 분류에 대한 신뢰를 제공

여러 PP가 결국 explanation과 동일한 training patch임

descriptors의 사용은 각 PP와 가장 관련이 있는 characteristics에 대한 세부 정보를 제공함으로써 시각적으로 유사한 PP의 cases를 완화

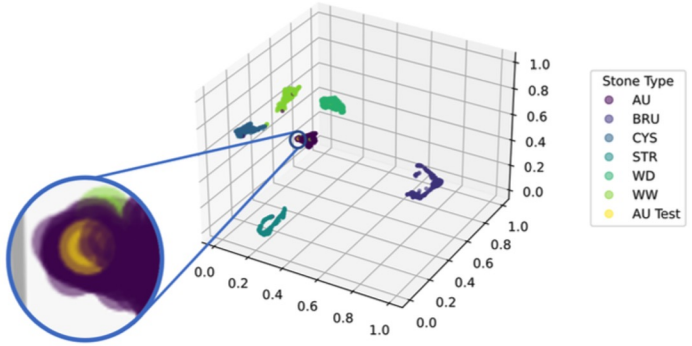


Figure 3: UMAP of the Part-Prototypes activations on section images. Our approach allows obtaining separate clusters of the output classes. On the zoomed example (blue circle) can be appreciated the projection of a new classification (yellow point) is surrounded by samples of the same class (purple points), indicating a high certainty on the correct classification of the test sample.

Conclusion and Future work

We showed that by training of PPs and extracting their descriptors we convert an uninterpretable VGG19 into an interpretable model.
This can facilitate the use of these models for ESR by a urologist.
However, mode collapse of the learned PPs is a limitation on the current implementations of ProtoPNets. To prevent this, better initialization procedures and loss function adjustments will be explored. Finally, we found indications of better class separability by use of PPs and their descriptors, to the point UMAP visualizations could be used to provide global context of the certainty of the output classification of a particular image.

- PPs를 훈련하고 descriptors를 추출하여 해석 불가능한 VGG19를 해석 가능한 모델로 변환
  비뇨기과 의사가 ESR에 이러한 모델을 사용하는 것을 용이하게 할 수 있음.

- mode collapse of the learned PPs는 ProtoPNets의 현재 구현에 대한 제한이다.
  더 나은 initialization와 손실 함수 조정 수행돼야 함

- 특정 이미지의 output classification의 확실성에 대한 글로벌 컨텍스트를 제공하기 위해 UMAP visualizations를 사용할 수 있을 정도로 PPs와 descriptors를 사용하여 더 나은 클래스 classification 가능성 확인

End of the Document