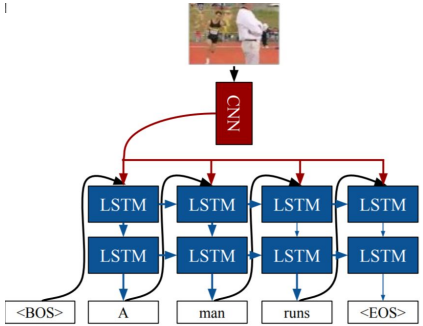
Google Research

Model interpretability as a tool for discovery

Subhashini Venugopalan

Background in Language and Vision

Image Captioning



LRCN: Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Donahue et. al. CVPR'15

Video Description

S2VT: Sequence to Sequence Video to Text.

Venugopalan et. al. ICCV'15

Describing "unseen" objects

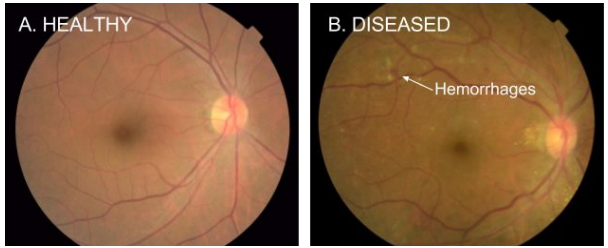
A **woodpecker** sitting on a tree branch in the woods.

A **orca** is riding a small wave in the water.

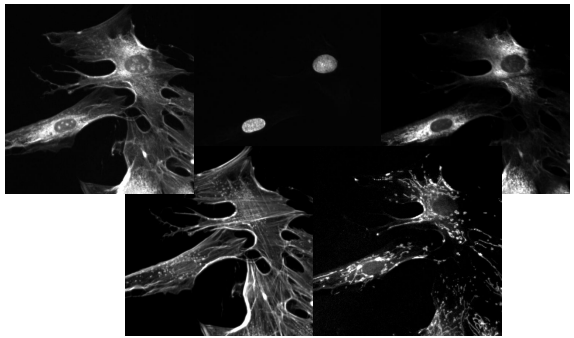
NOC: (Novel Object Captioner) Captioning Images with Diverse Objects

Venugopalan et. al. CVPR'17

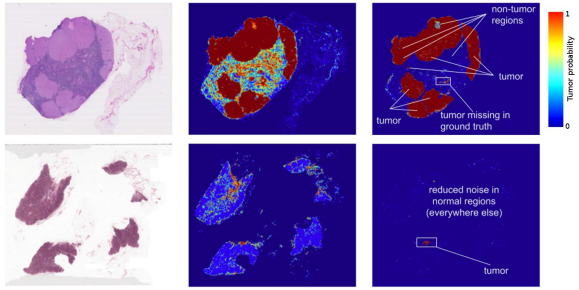
ML (transfer learning) for medically relevant problems



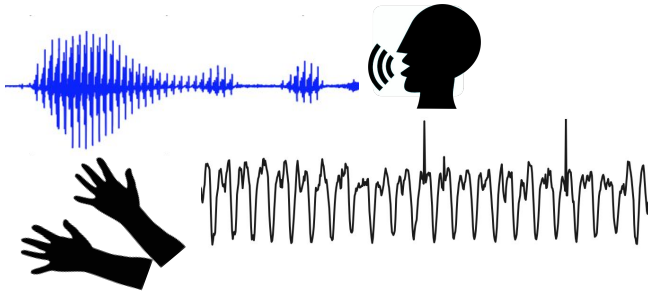
Deep Learning for Detection of Diabetic Retinopathy (Gulshan et. al. JAMA'16)



Searching for biomarkers from microscopy images. (Yang et. al. SLAS Discovery '19)



Detecting cancer metastases on pathology images. (Liu et. al. 2017)

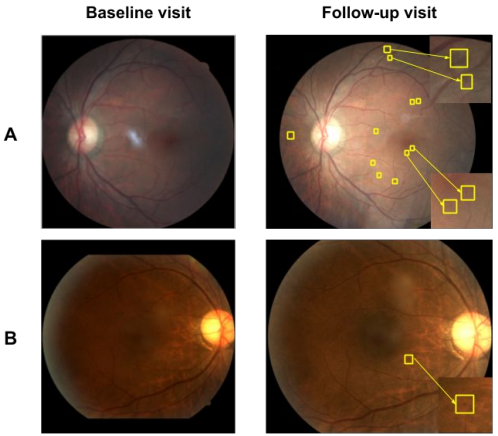


ALS disease progression from voice and accelerometer samples.

Better source of truth leads to novel signals, superhuman perf

future outcome

novel signal



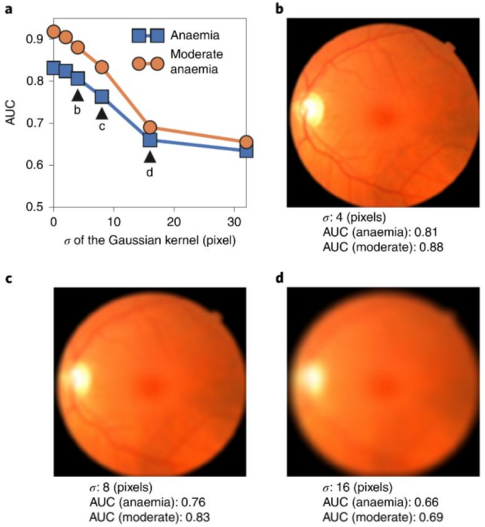
Predicting risk of developing diabetic retinopathy using deep learning

Ashish Bora, Siva Balasubramanian, Boris Babenko, Sunny Virmani, Subhashini Venugopalan, Akinori Mitani, Guilherme de Oliveira Marinho, Jorge Cuadros, Paisan Ruamviboonsuk, Greg S Corrado, Lily Peng, Dale R Webster, Avinash V Varadarajan, Naama Hammel, Yun Liu, Pinal Bavishi

Lancet Digital Health (to appear)

different modality label

novel signal



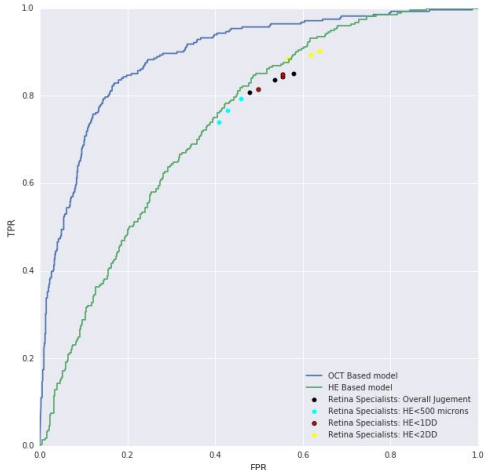
Detection of anaemia from retinal fundus images via deep learning

Akinori Mitani ✉, Abigail Huang, Subhashini Venugopalan, Greg S. Corrado, Lily Peng, Dale R. Webster, Naama Hammel, Yun Liu & Avinash V. Varadarajan

Nature Biomedical Engineering **4**, 18–27(2020) | Cite this article

different modality label

strongly exceeds

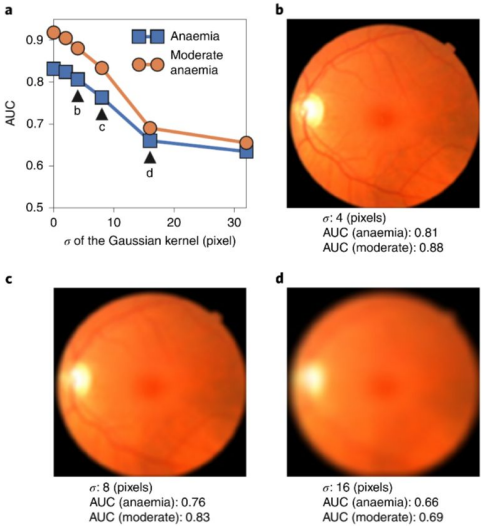


Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning

Avinash V. Varadarajan, Pinal Bavishi, Paisan Ruamviboonsuk, Peranut Chotcomwongse, Subhashini Venugopalan, Arunachalam Narayanaswamy, Jorge Cuadros, Kuniyoshi Kanai, George Bresnick, Mongkol Tadarati, Sukhum Silpa-archa, Jirawut Limwattanayingyong, Variya Nganthavee, Joseph R. Ledsam, Pearse A. Keane, Greg S. Corrado, Lily Peng ✉ & Dale R. Webster

Nature Communications **11**, Article number: 130 (2020) | Cite this article

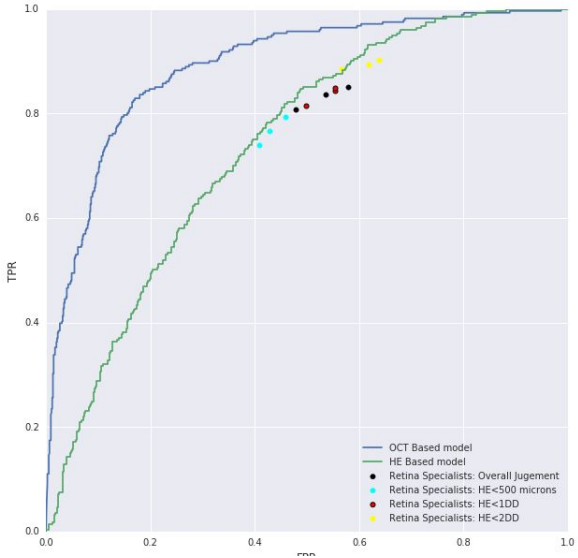
Learning from outcome - novel signals, outperforming humans



Detection of anaemia from retinal fundus images via deep learning

Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S. Corrado, Lily Peng, Dale R. Webster, Naama Hammel, Yun Liu & Avinash V. Varadarajan

Nature Biomedical Engineering **4**, 18–27(2020) | Cite this article



Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning

Avinash V. Varadarajan, Pinal Bavishi, Paisan Ruamviboonsuk, Peranut Chotcomwongse, Subhashini Venugopalan, Arunachalam Narayanaswamy, Jorge Cuadros, Kuniyoshi Kanai, George Bresnick, Mongkol Tadarati, Sukhum Silpa-archa, Jirawut Limwattanayingyong, Variya Nganthavee, Joseph R. Ledsam, Pearse A. Keane, Greg S. Corrado, Lily Peng & Dale R. Webster

Nature Communications **11**, Article number: 130 (2020) | Cite this article

* Background

* **Scientific Discovery by Generating Counterfactuals using Image Translation**

  Narayanaswamy*, Venugopalan*, et. al. MICCAI 2020

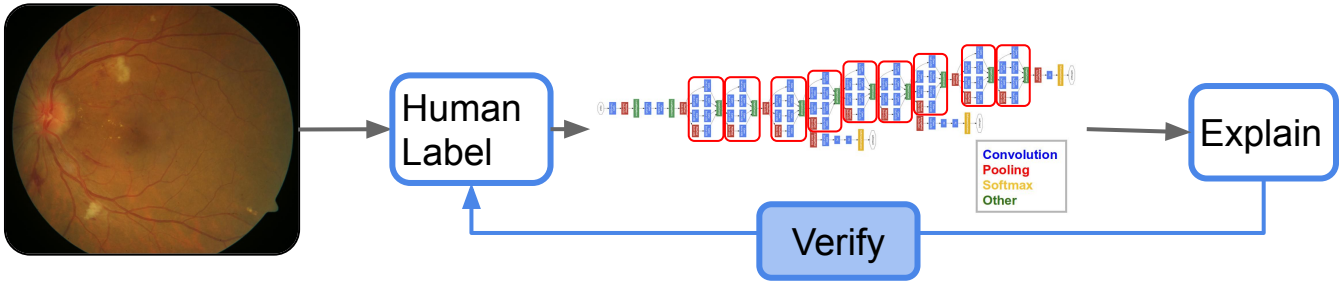
* Attribution in Scale and Space

  Xu, Venugopalan, Sundararajan CVPR 2020

* Predicting risk of developing diabetic retinopathy using deep learning

  Bora et. al. Lancet Digital Health 2021

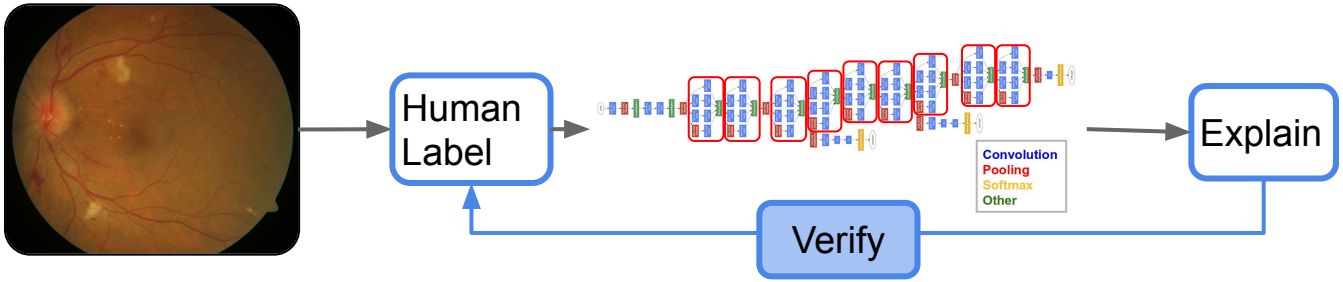
Prediction to Explanation

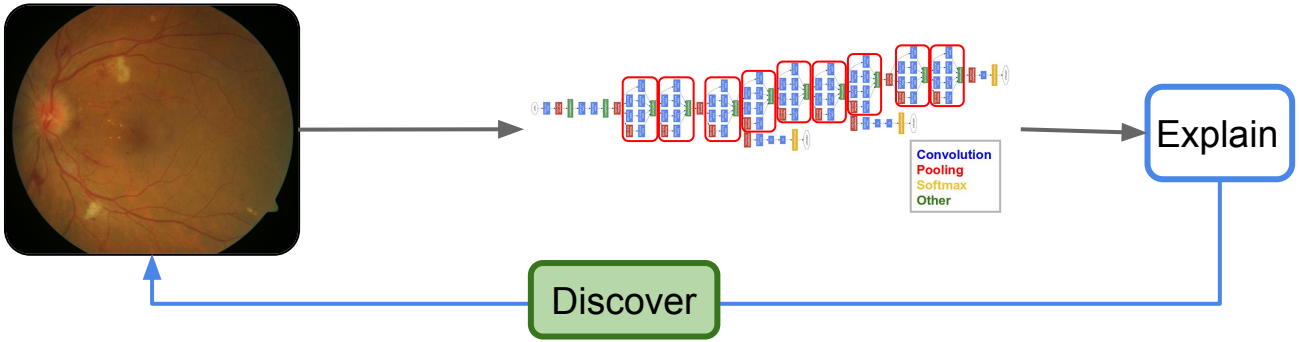


Validation → Is the model making predictions for the right reasons?

Does it conform to how humans choose a label? (Builds trust)

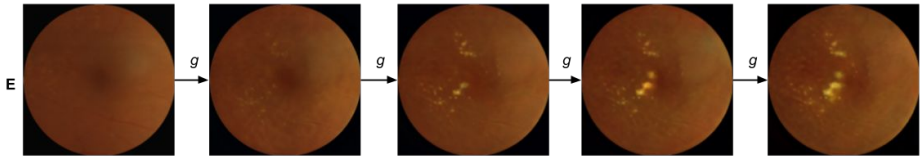
Prediction to Explanation to Discovery





Discovery? Identify properties to teach humans

Progressively exaggerate the kind of things that the model is looking at.



Translate what we can learn into a simple set of properties / features

to enable experts to perform better

or improve our understanding of the disease.

Diabetic Macular Edema - case study

Diabetic Macular Edema (DME)

- late stage of diabetic eye disease.
- characterized by retinal thickening in the macula.
- often results in vision loss.

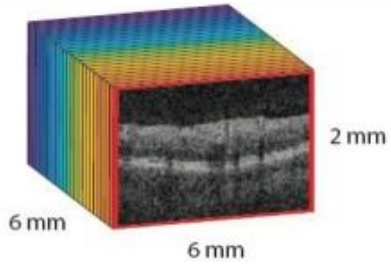
Diabetic Macular Edema - 2 sources of labels

Diabetic Macular Edema (DME)

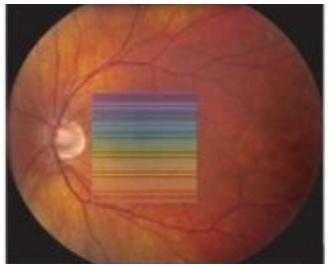
- late stage of diabetic eye disease.
- characterized by retinal thickening in the macula.
- often results in vision loss.

Diagnosis:

- referral based on fundus image. (>80% FP rate)
- confirmation based on measuring retinal thickness in optical coherence tomography (OCT) image (3D).



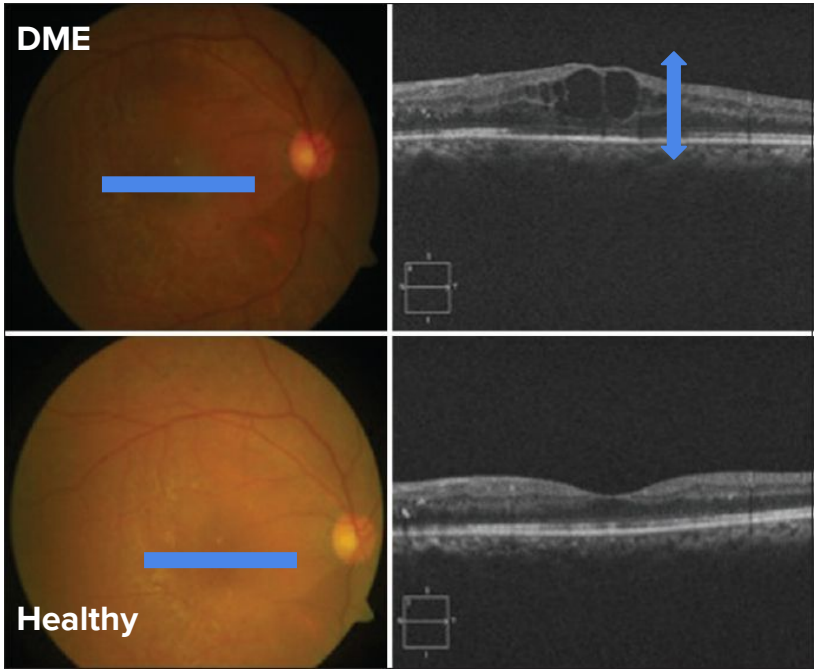
OCT Slices (3D)



Fundus photograph (2D)

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3785070/

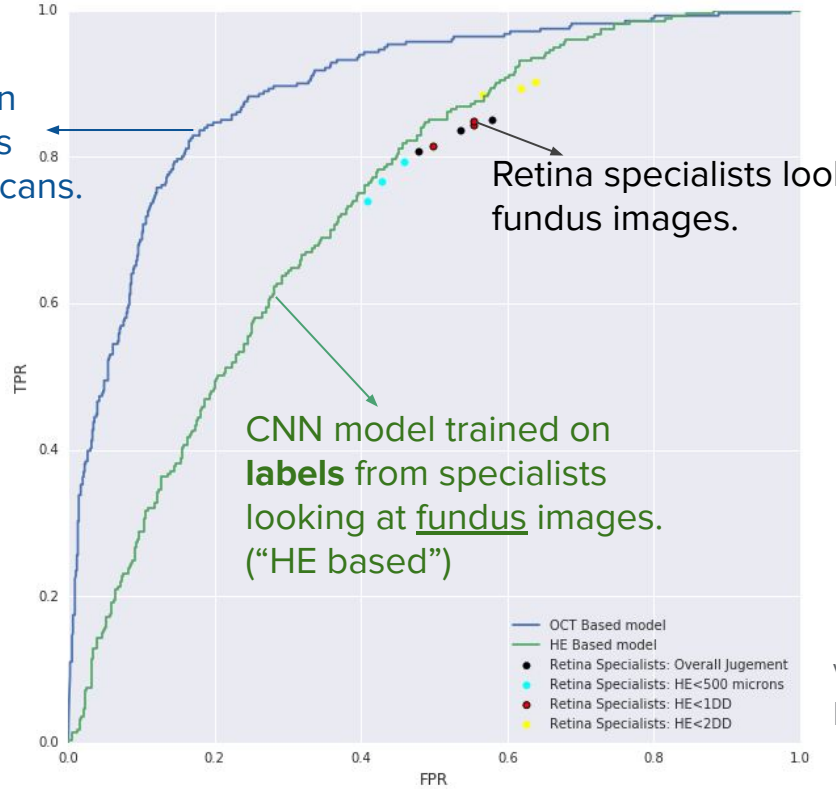
Diabetic Macular Edema - 2 sources of labels



Fundus photograph Center OCT Slice

CNNs see more from 2d fundus images than specialists

CNN model trained on **labels** from specialists looking at OCT (3D) scans. (“OCT based”)



Retina specialists looking at fundus images.

CNN model trained on **labels** from specialists looking at fundus images. (“HE based”)

Varadarajan et. al. Nature Comms. ‘20

Validation - Is the signal in a specific region?

Collect labels to ablate image regions and evaluate performance.

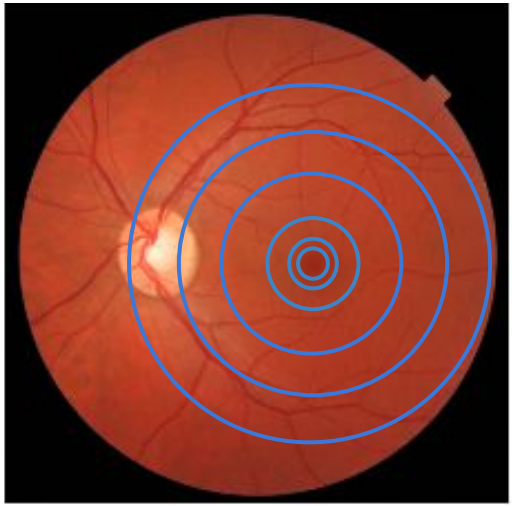
Optic Disc

Fovea / Macula

Validation - Is the signal in a specific region?

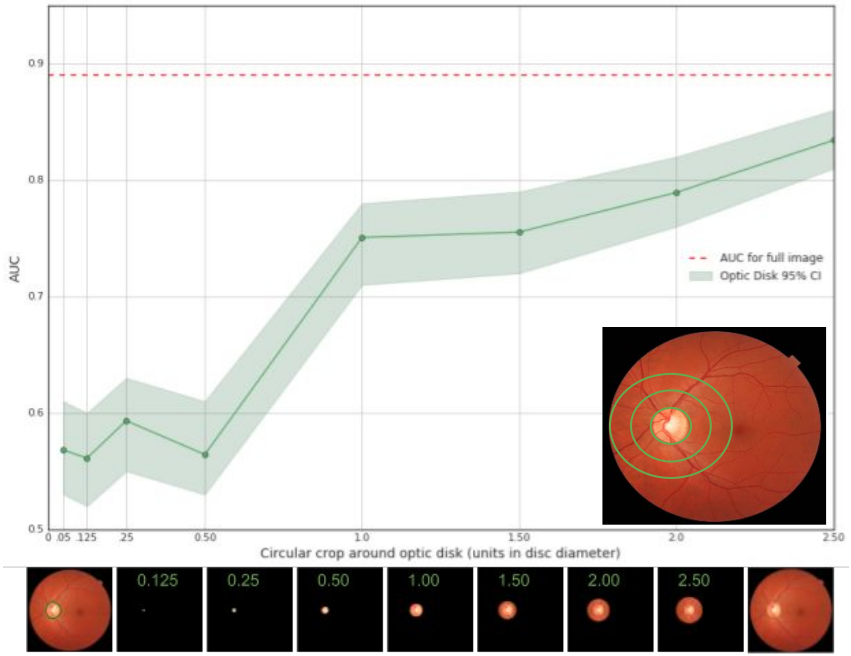
Extract crops in multiples of the optic disc diameter and train models.

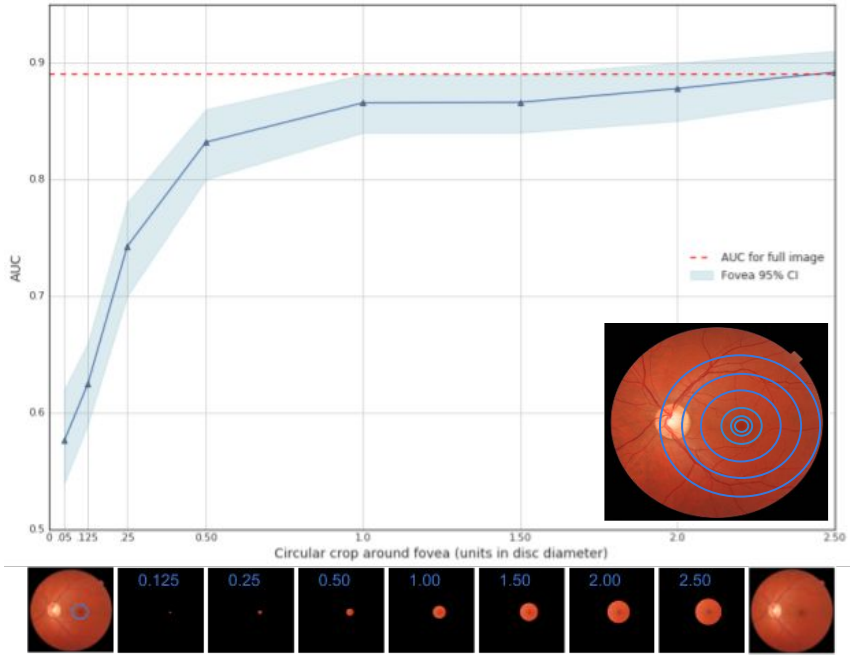
Optic Disc



Fovea / Macula

Region around fovea accounts for performance.





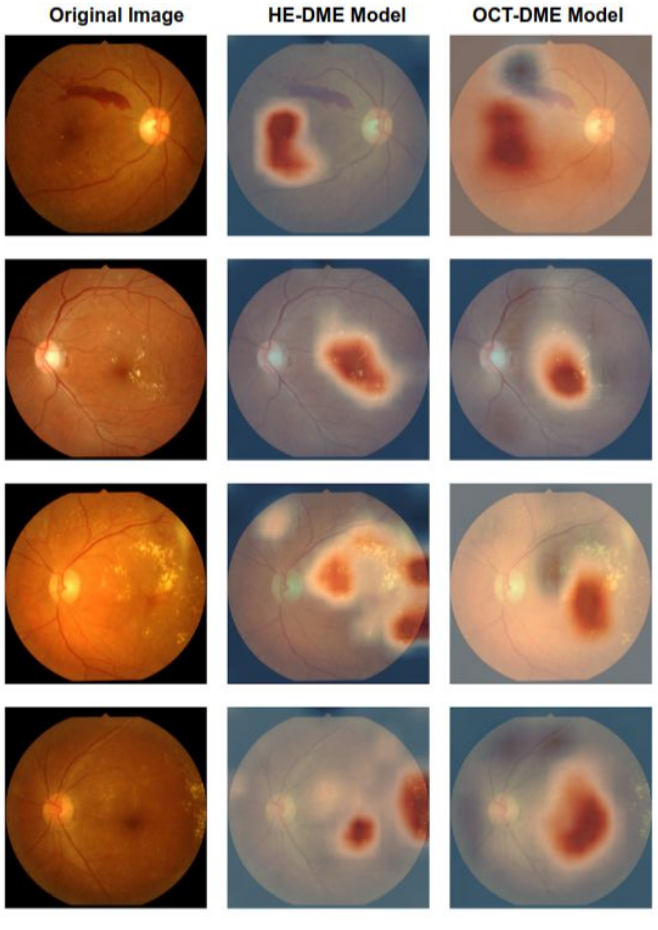
Saliency maps - Where is it looking?

Middle column - HE-DME: CNN trained on specialist labels on fundus images (2d)

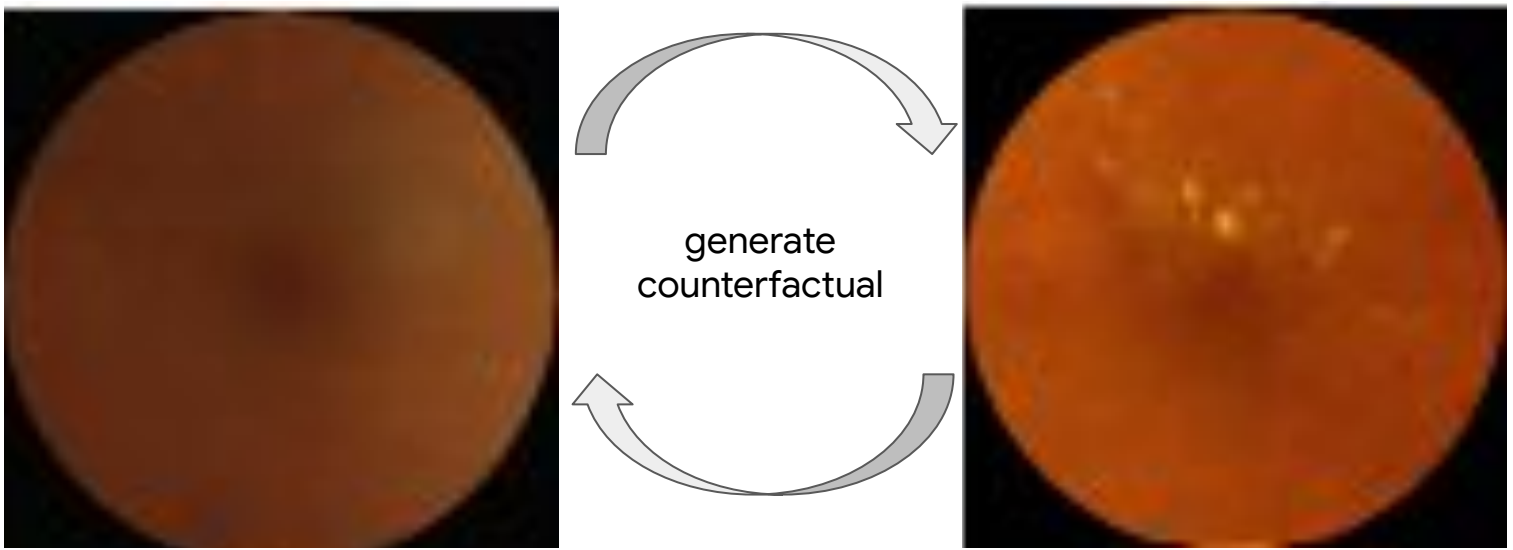
- Hard exudates (HE) yellow lesions

Last column - OCT-DME: CNN trained on specialist labels from OCT scans (3d)

- Mostly around macula

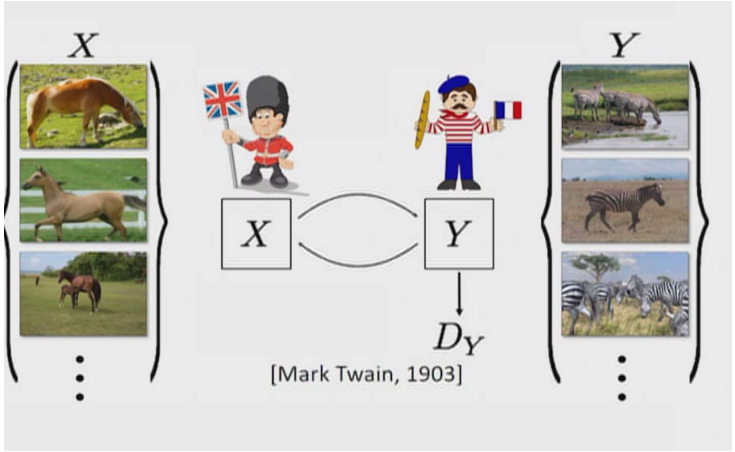


What is different about those regions?



No DME **DME**

Image to Image translation (cycleGANs) to the rescue.



Zhu et. al. ICCV'17 (UC Berkeley)

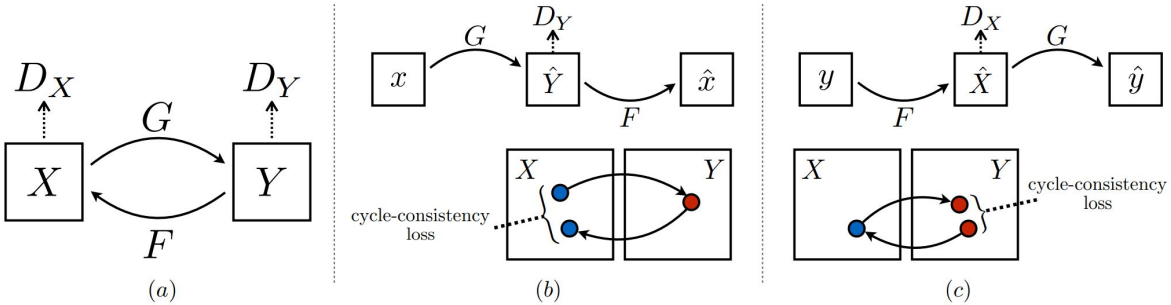
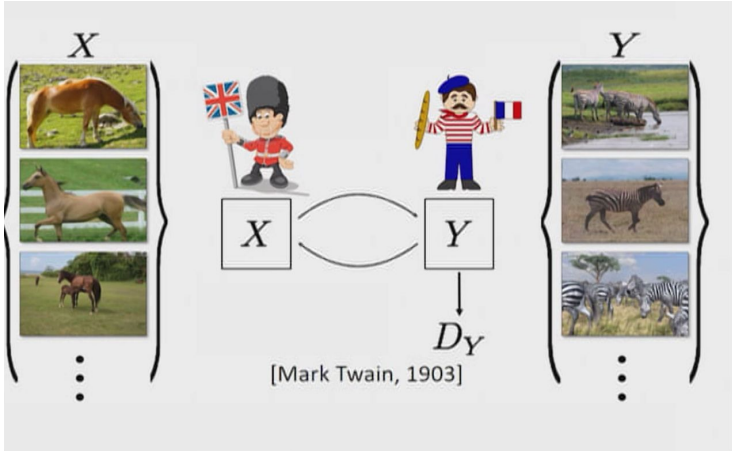
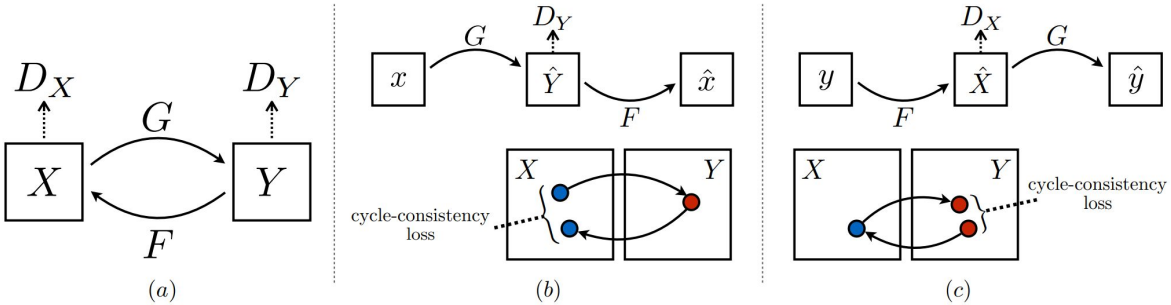


Image to Image translation (cycleGANs) to the rescue.



Zhu et. al. ICCV'17 (UC Berkeley)



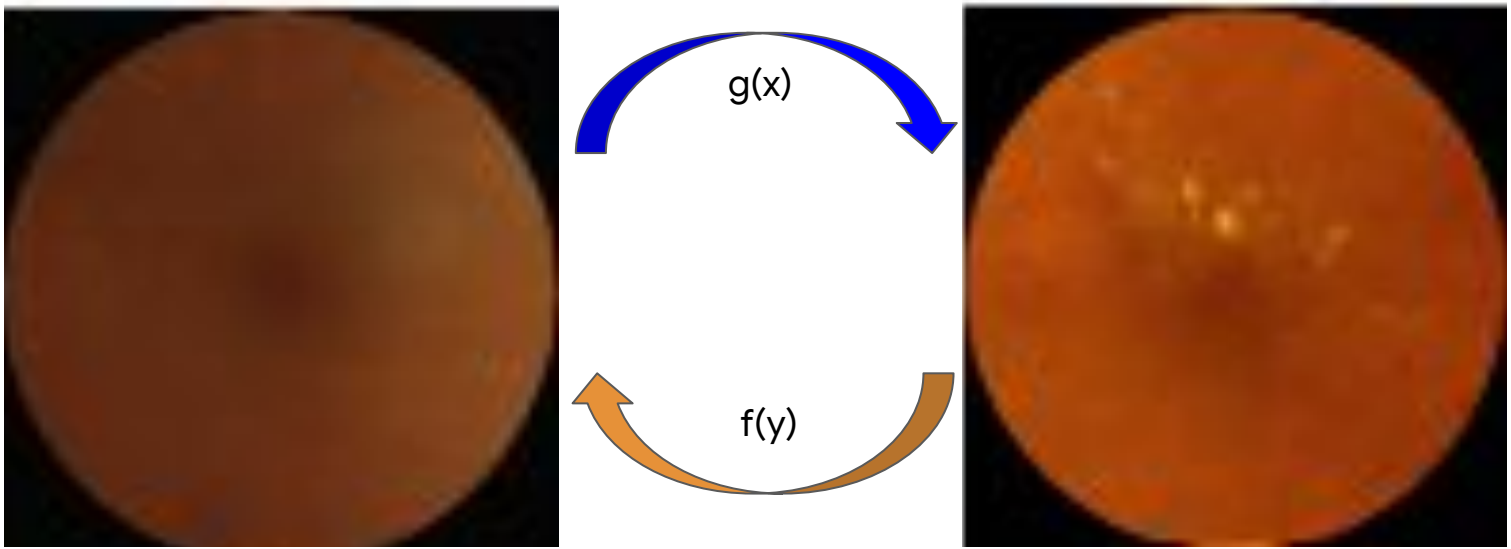
Add 1x1 conv.

Add residual connection.

(easy to copy)

CycleGANs to visualize the difference

Specifically apply it to region around macula.

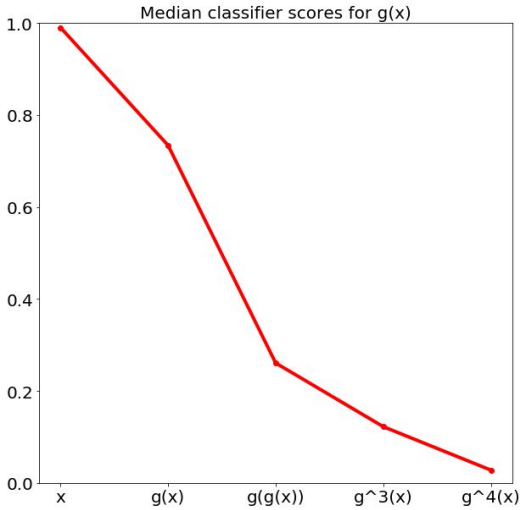


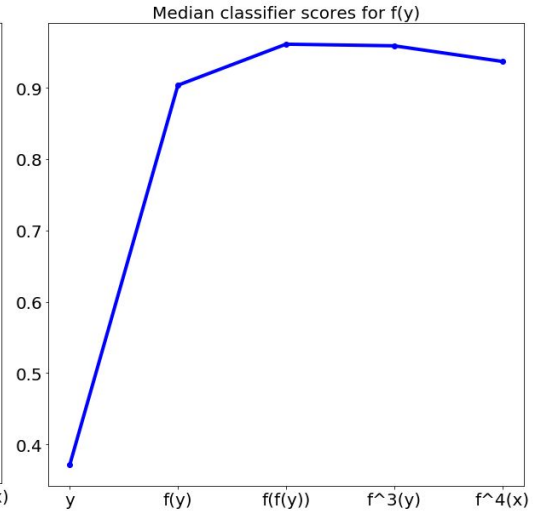
No DME **DME**

Verify that the cycleGAN is “faithful” to the model

Verify that the independently trained classifier AUC changes as expected when we convert/translate images from one class to another.

No DME





DME

Input	AUC [95% CI range]
(x, y)	0.804 [0.746 - 0.862]
$g(x), f(y)$	0.374 [0.294 - 0.450]
$g^2(x), f^2(y)$	0.180 [0.139 - 0.241]
$g^3(x), f^3(y)$	0.130 [0.087 - 0.186]
$g^4(x), f^4(y)$	0.106 [0.060 - 0.156]

Google Research

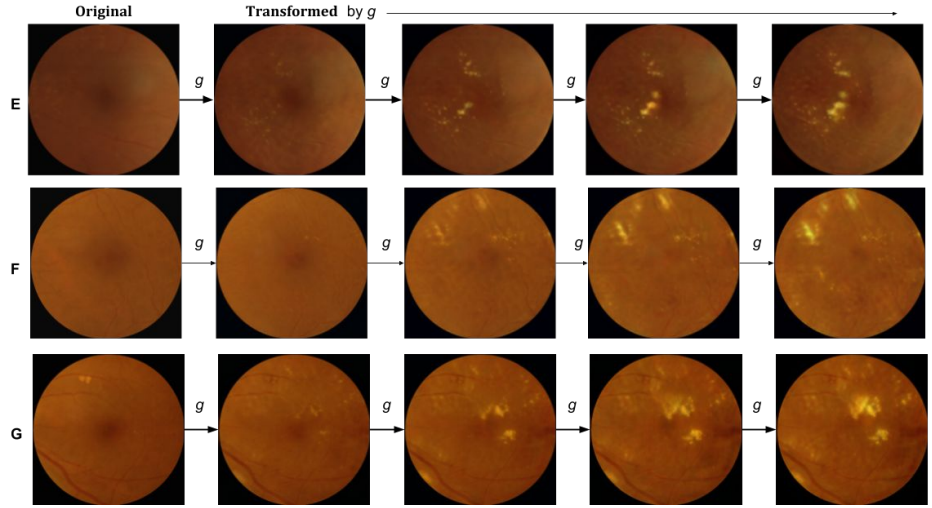
Observe transformations

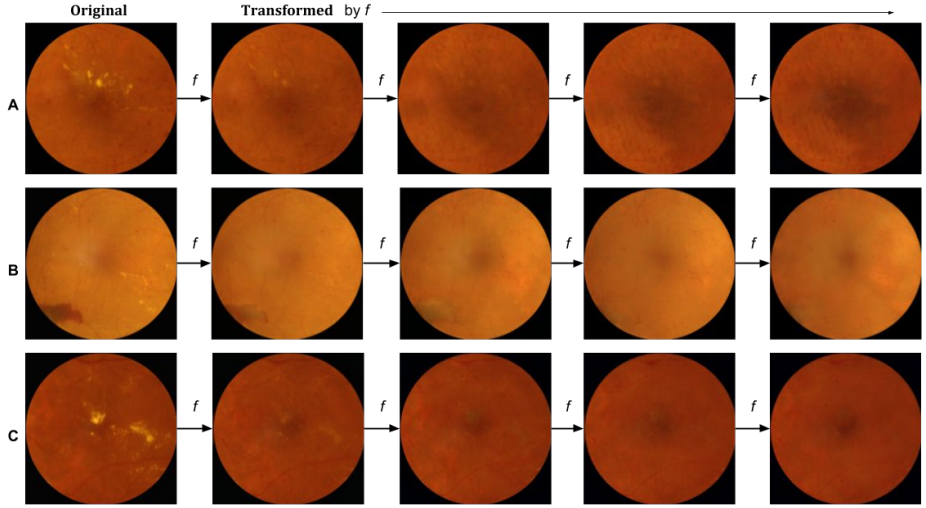
g(x) (i.e. No DME to DME):

1. Adds hard exudates (yellow lesions)
2. Lightens fovea

f(y) (i.e. DME to No DME):

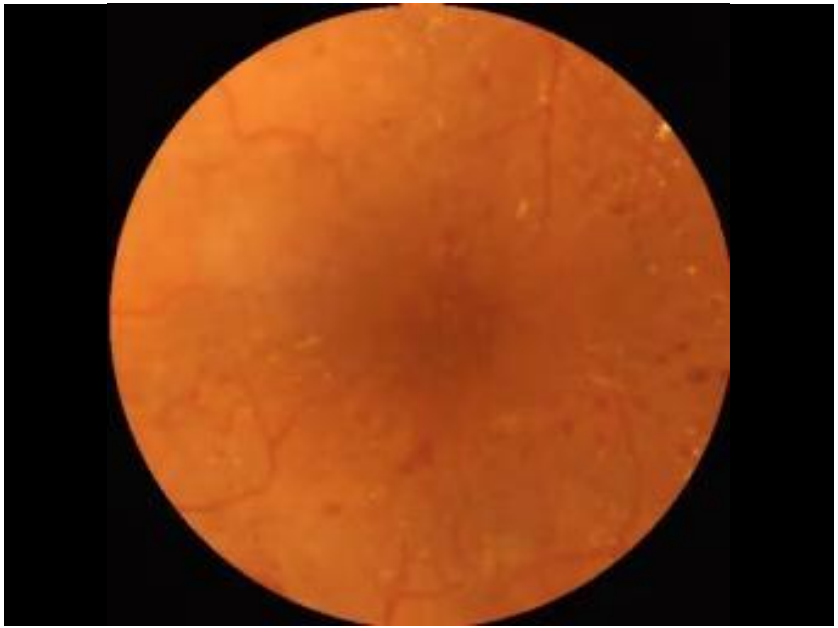
1. Removes hard exudates
2. Darkens fovea (very subtle)





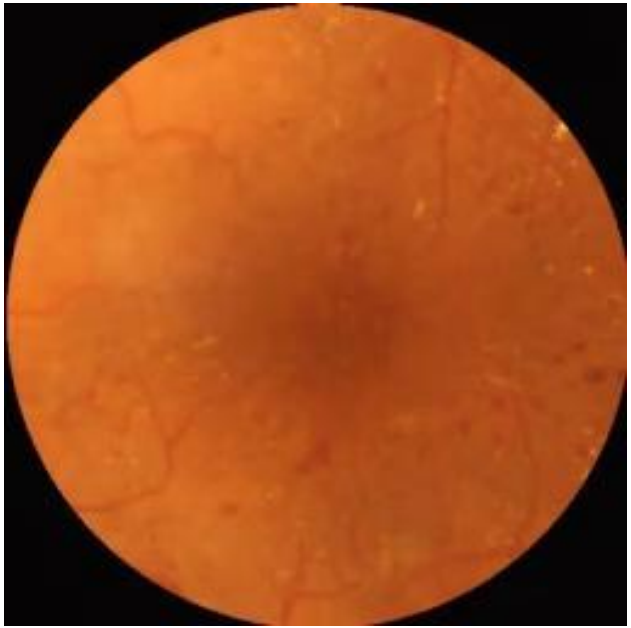
Observe transformations (loop)

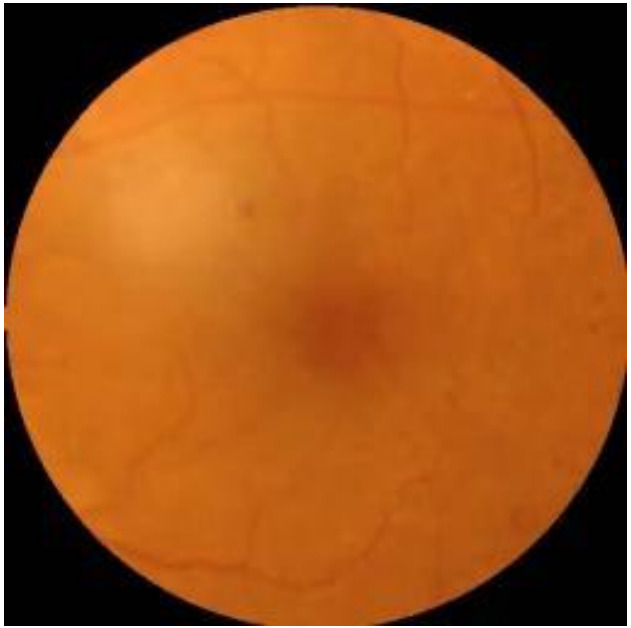
Alpha-blended images applying the transformation and back.



Observe transformations (loop)

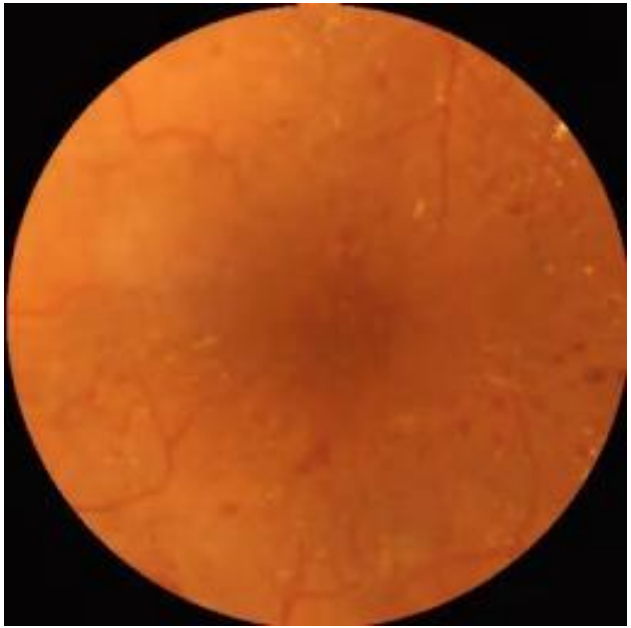
Alpha-blended images applying the transformation and back.

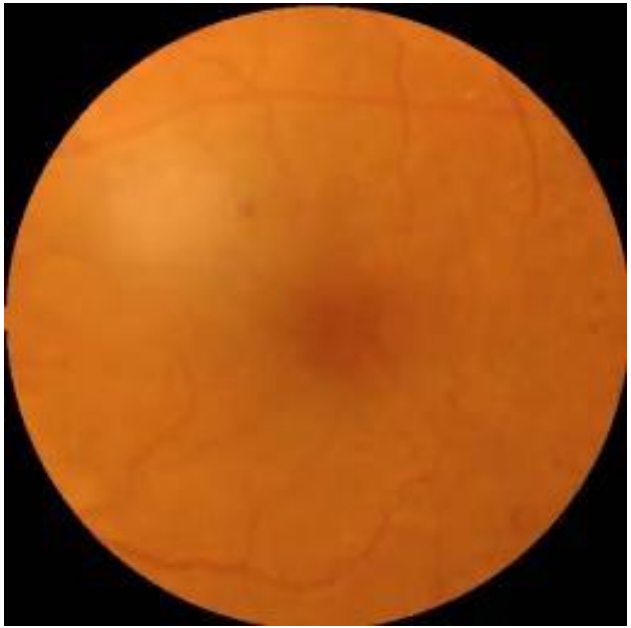




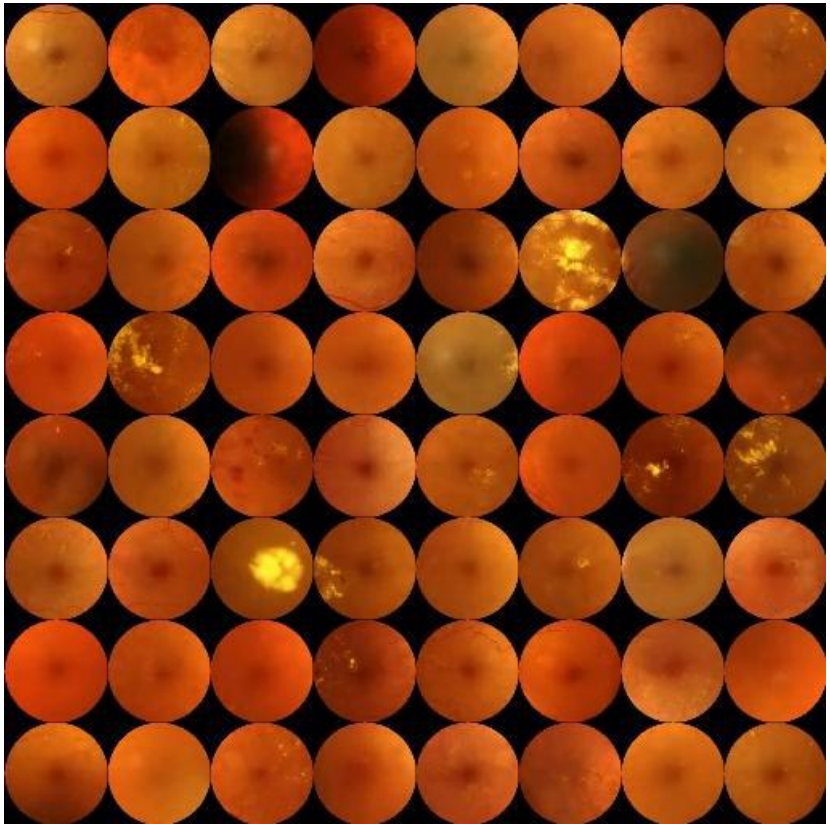
Observe transformations (loop)

Alpha-blended images applying the transformation and back.





Observe lots of transformations



Hand engineer features to explain performance

On the full image

- Are hard exudates (yellow lesions) present or not? --- Easy for humans
- Hand-engineered features - mean intensities of pixels (at concentric circles around fovea) [10 (x3) values]
  - fit functional form from weights of a linear model

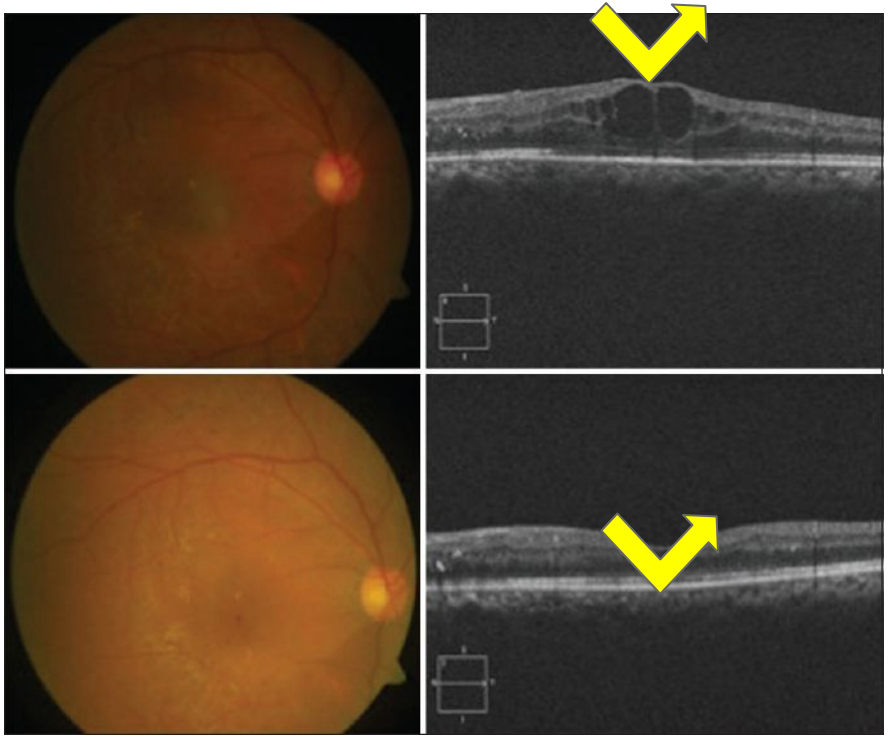
Features	AUC SVM	AUC MLP
Hand-engineered features alone	72.4 ± 0.0	76.3 ± 0.3
Presence of hard exudates alone	74.1 ± 0.0	74.1 ± 0.0
Hand-engineered features + hard exudates' presence	$\mathbf{81.4 \pm 0.0}$	$\mathbf{82.2 \pm 0.2}$
M(raw pixels single task on cropped image)	CNN: 84.7	

Light bounces differently of the surface

Hypothesis (validating in simulation)

- Deeper fovea leads to darker macula
- Flatter fovea leads to lighter macula

Essentially use prediction models, and explanations to generate hypotheses (hopefully leading to discoveries)



Appearing in MICCAI 2020

Google Research and Google AI Healthcare

Arun Narayanaswamy

Subhashini Venugopalan

Philip Nelson

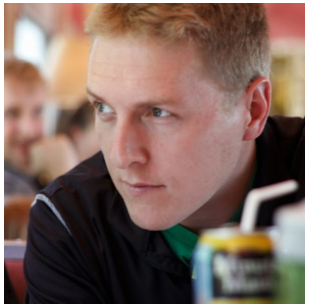
Michael Brenner

Rory Sayres

Lily Peng

Pinal Bavishi

Abigail Huang



Dale Webster

Avinash Varadarajan

* Background

* Scientific Discovery by Generating Counterfactuals using Image Translation

  Narayanaswamy*, Venugopalan*, et. al. MICCAI 2020

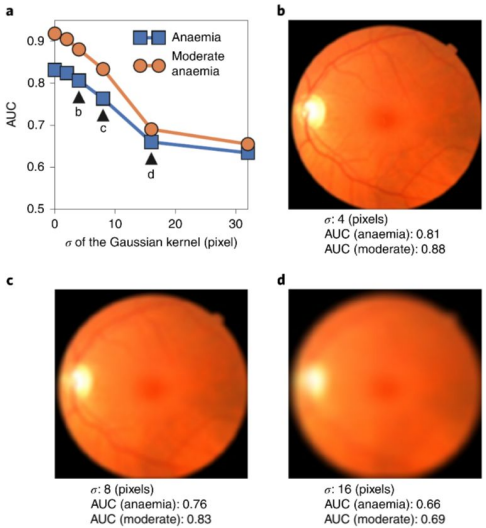
* **Attribution in Scale and Space**

  Xu, Venugopalan, Sundararajan CVPR 2020

* Predicting risk of developing diabetic retinopathy using deep learning

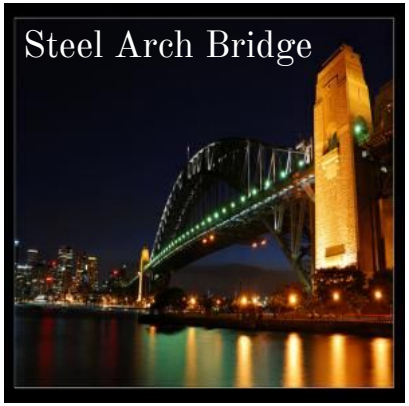
  Bora et. al. Lancet Digital Health 2021

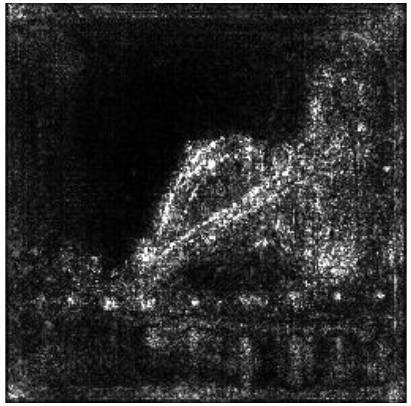
At what scale/frequency does a network recognize signal?

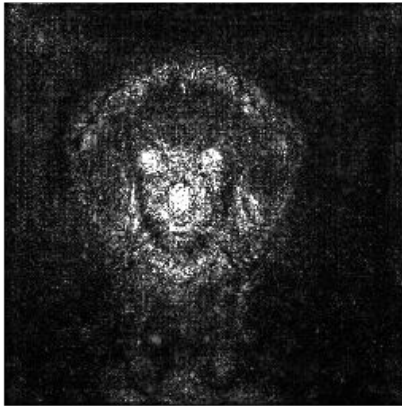


Mitani et. al. Nature BME ‘20

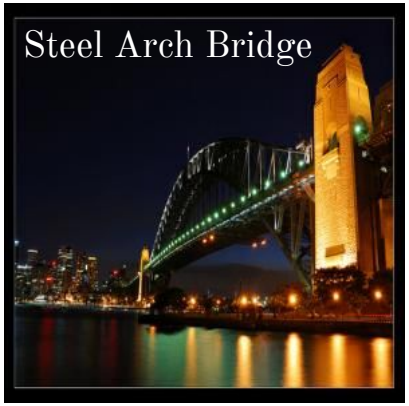
Existing explanations localize in space (i.e. pixels)

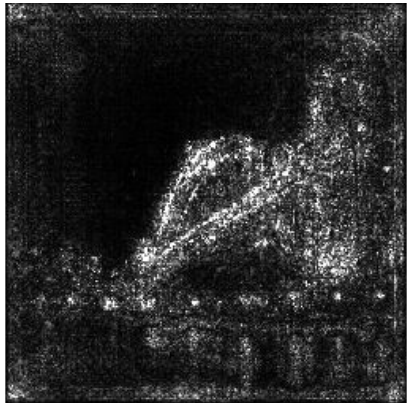




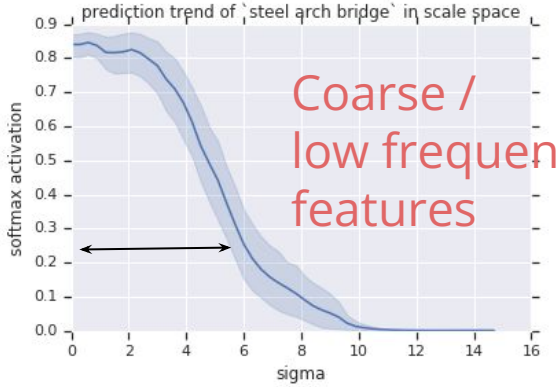


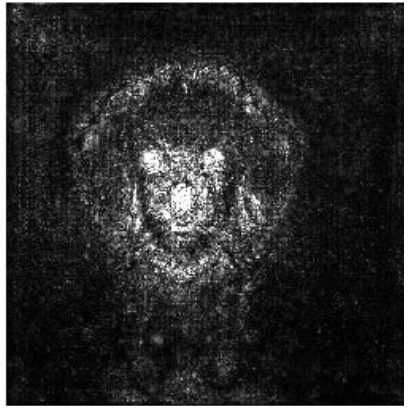
Can we also localize in scale/frequency?



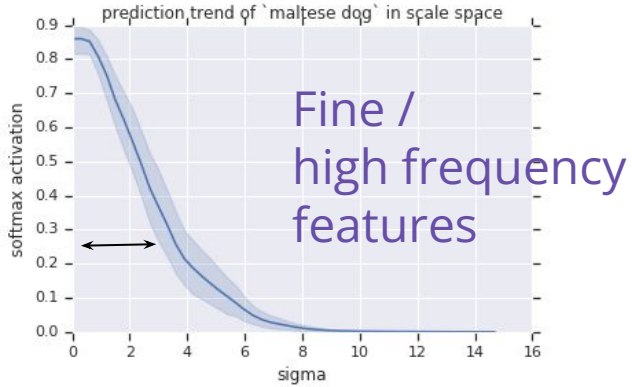


arch bridge





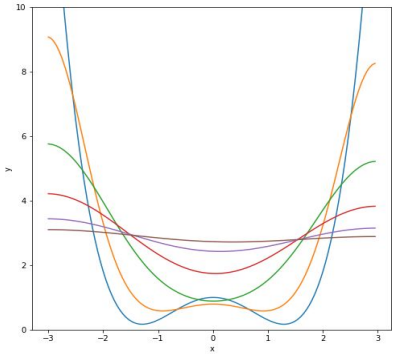
dog breed



Key idea from scale-space theory

To handle image structures at different scales, **represent an image as a one-parameter family of smoothed images**. Burt and Adelson ‘81, T. Lindberg ‘90

Gaussian blur perturbation only destroys information!
Ensures explanation is free of artifacts.



How? Integrate gradients, progressively reducing blur

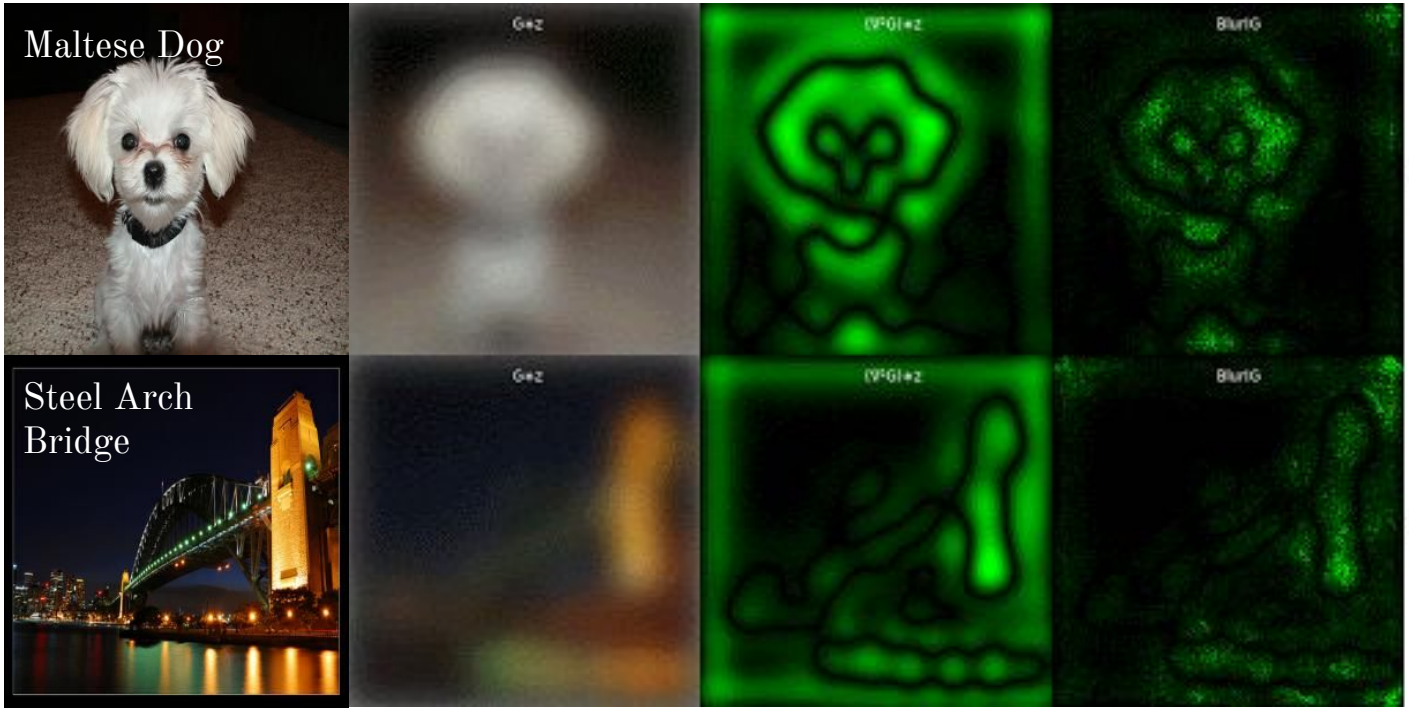
Original image

Gaussian filtered image

Gaussian derivatives localize attributions in scale

Model derivatives localize attributions in space

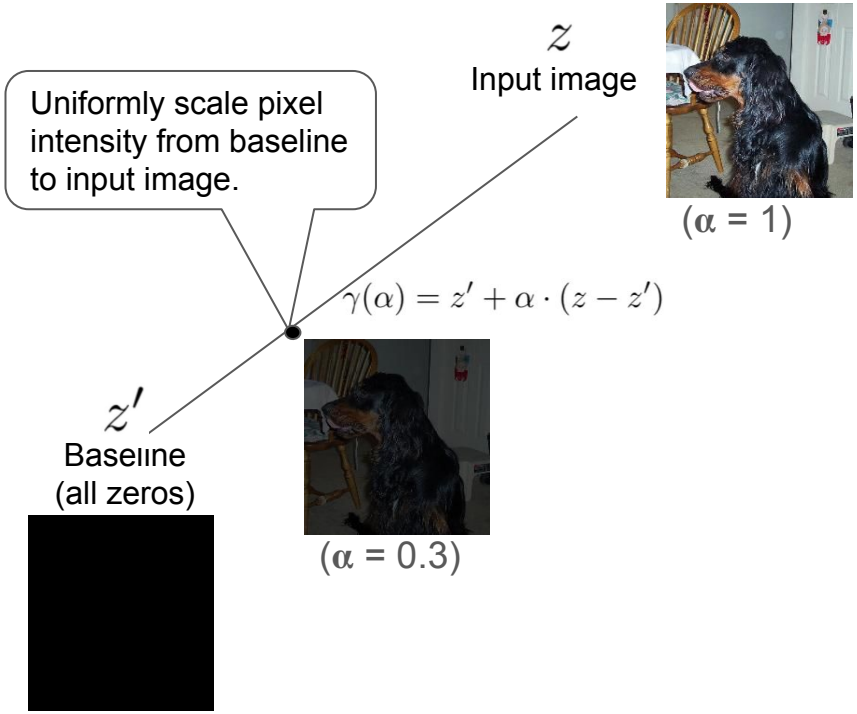
Evolution of prediction scores



Google Research

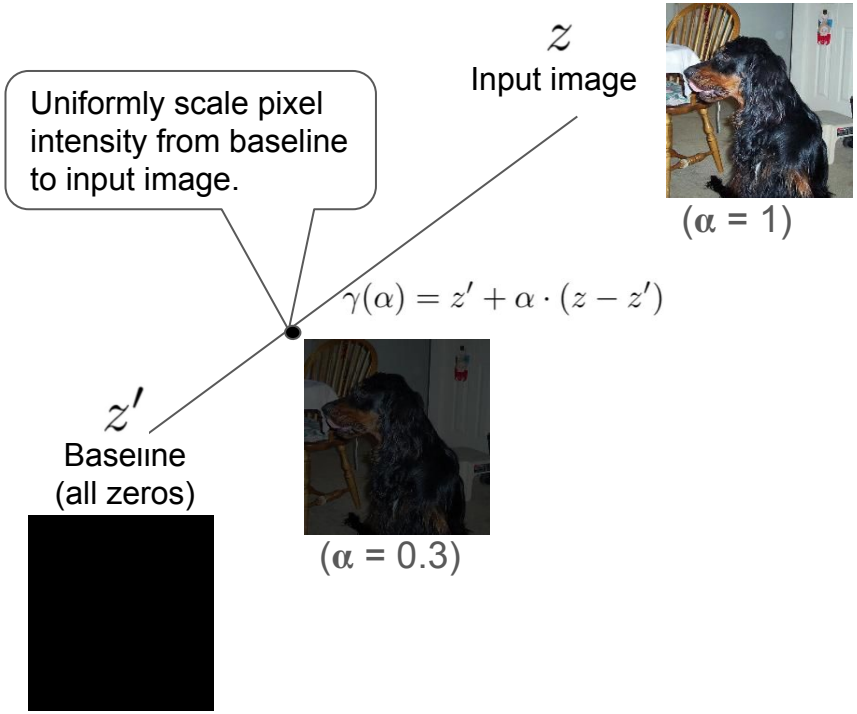
BlurIG:
Blur Integrated Gradients

Integrated Gradients: Intensity path



Baseline → “information-less” input

Integrated Gradients: Localize attributions in space



Baseline → “information-less” input

$$IG(\text{image}) = (z - z') \cdot \int_{\alpha=0}^{1} \frac{\partial F(z' + \alpha \cdot (z - z'))}{\partial \gamma(\alpha)} \, d\alpha$$

model

gradient

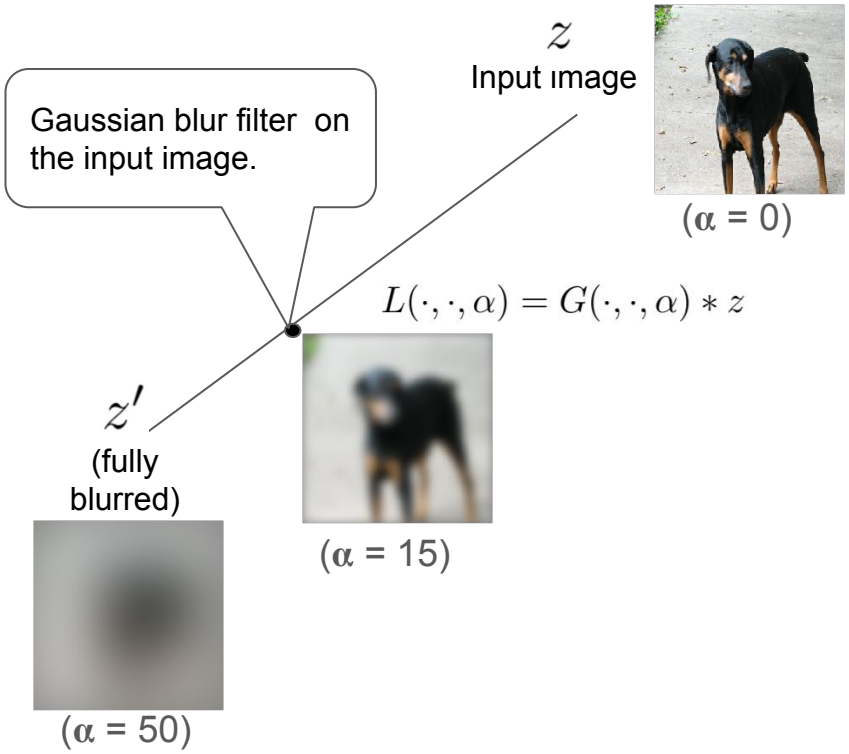
F is the prediction function for the label.

Attribution is per pixel, but I’m ignoring subscripts.

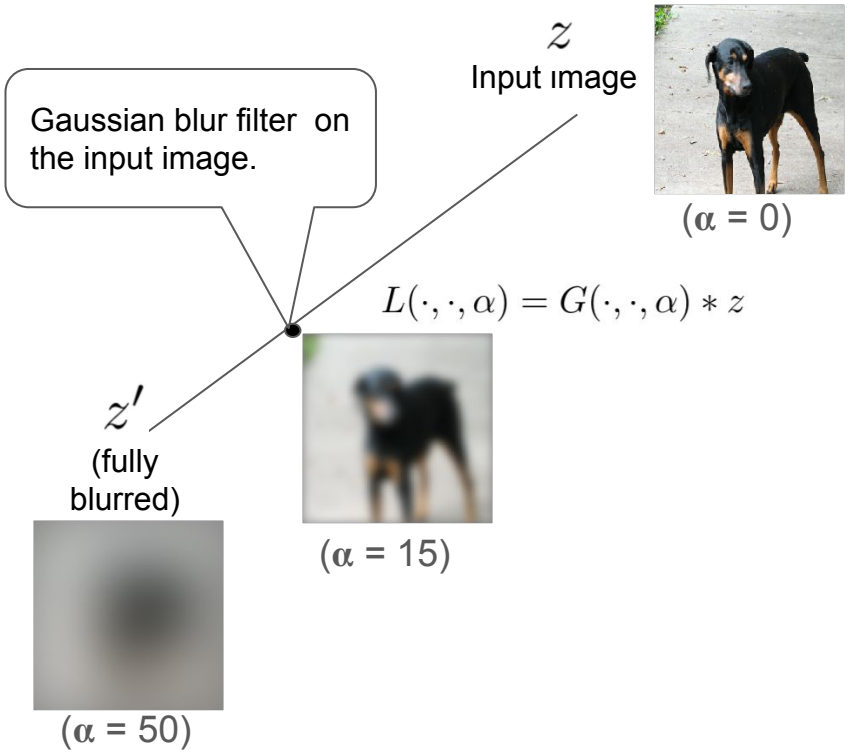
Why integral? Why not just gradient?

--Sundararajan et. al. ICML’17

BlurIG: Gaussian Blur path



BlurIG: Gaussian Blur path



$$\text{BlurIG}(x,y) ::= \int_{\alpha=\infty}^{0} \frac{\partial F(L(x,y,\alpha))}{\partial L(x,y,\alpha)} \frac{\partial L(x,y,\alpha)}{\partial \alpha}\, d\alpha$$

model

gradient LoG filtering

F is the prediction function for the label.

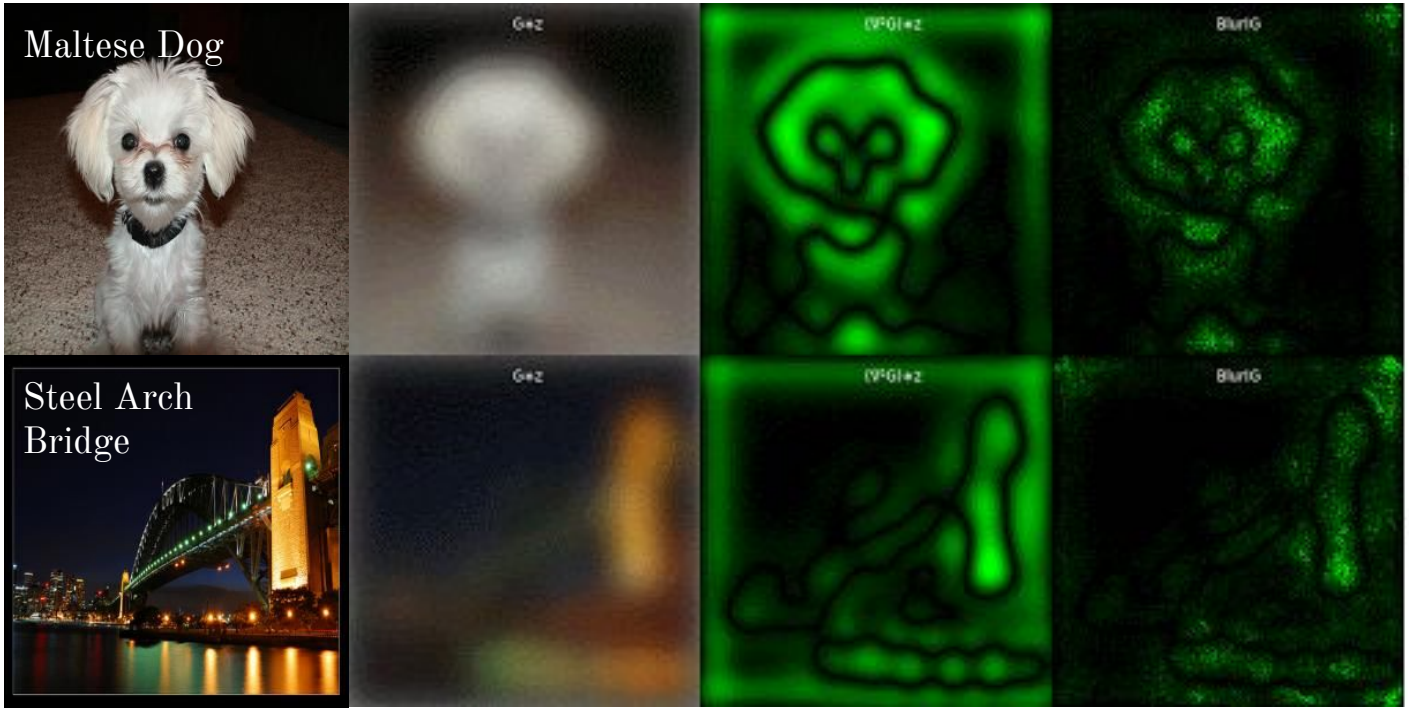
Attribution is per pixel indicated by (x, y).

LoG → Laplacian of Gaussian (blob detector!)

Google Research

BlurIG: Integrated gradients along the Gaussian blur path

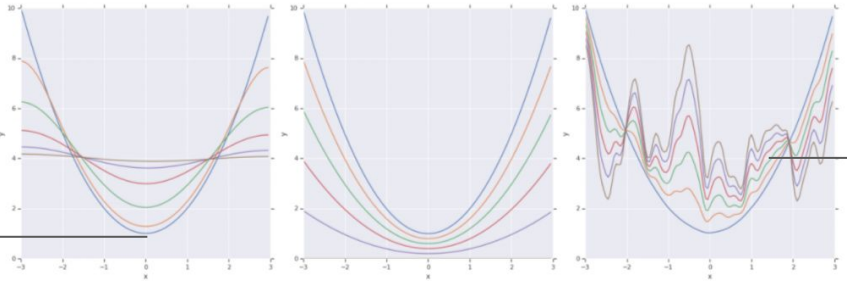
Original image	Gaussian filtered image	Gaussian derivatives localize attributions in scale	Model derivatives localize attributions in space	Evolution of prediction scores



Scale-space axioms - attractive properties for explanation

Gaussian blur perturbation only destroys information!
Does not introduce artifacts.

Blur path - Non-enhancement of local extrema.



(a) Gaussian blur (b) Intensity scaling (black baseline) (c) Intensity scaling (random baseline)

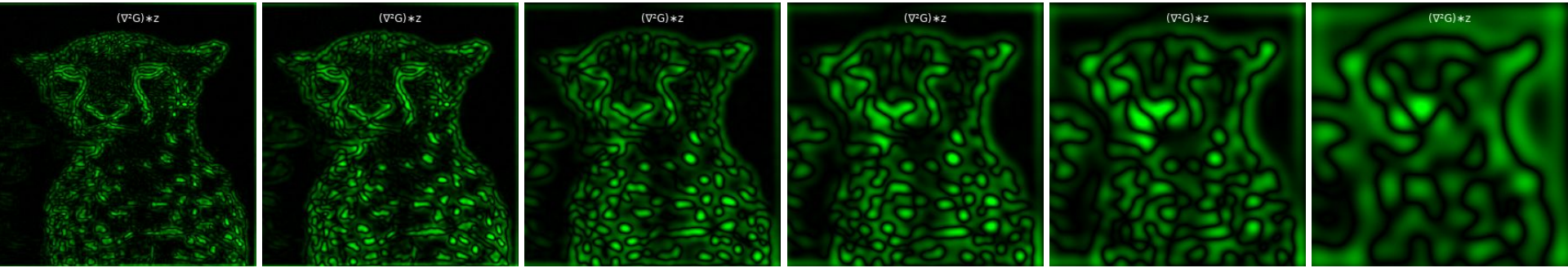
Intensity path - creation of local extrema!

Figure 2: Scale space for $x^2 + 1$ along the Gaussian blur, and intensity scaling (black and random baseline) paths.

Scale-space axioms - attractive properties for explanation

- **Gaussian blur perturbation only destroys information!**
  Does not introduce artifacts.

- **Gaussian operator and Laplacian of Gaussian operator enhance image features - edges, blobs, textures.**



Scale-space axioms - attractive properties for explanation

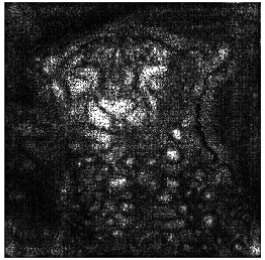
- Gaussian blur perturbation only destroys information!
  Does not introduce artifacts.

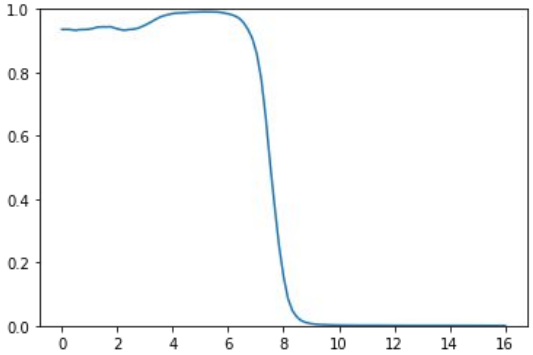
- Gaussian operator and Laplacian of Gaussian operator enhance image features - edges, blobs, textures.

- Gaussian blur path eliminates the need for an “information-less” baseline image for Integrated Gradients.

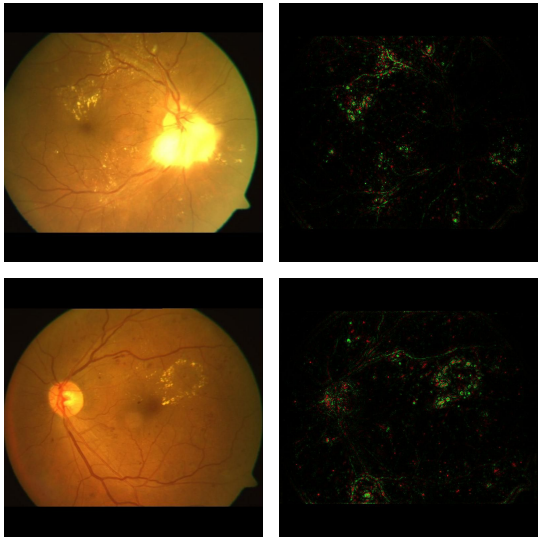
Applied to 3 classification tasks

IMAGENET

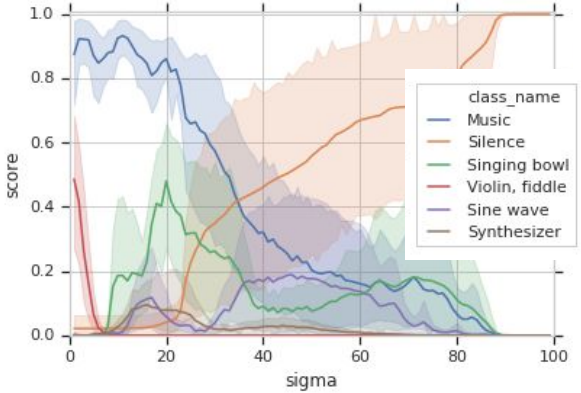




Diabetic Retinopathy

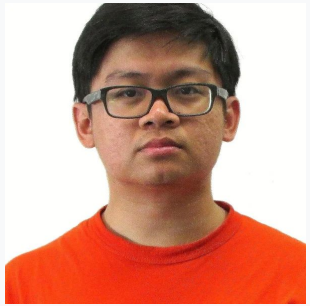


Audio identification



Links

Google Research



Shawn Xu

Subhashini Venugopalan

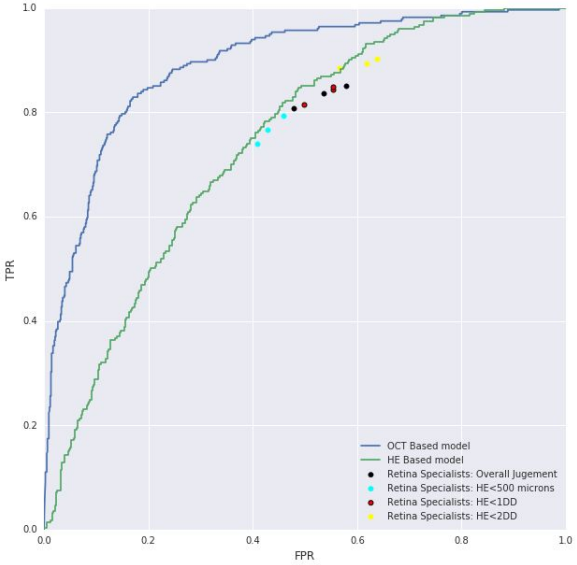
Mukund Sundararajan

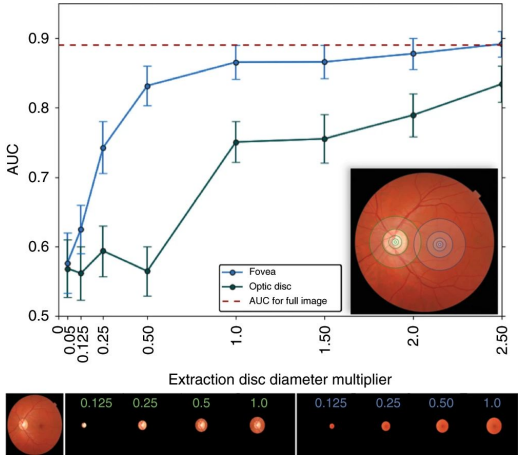
Paper: https://arxiv.org/abs/2004.03383

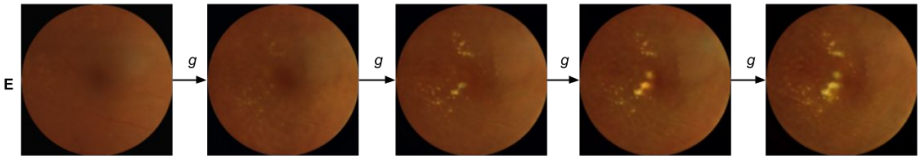
Code: https://github.com/PAIR-code/saliency

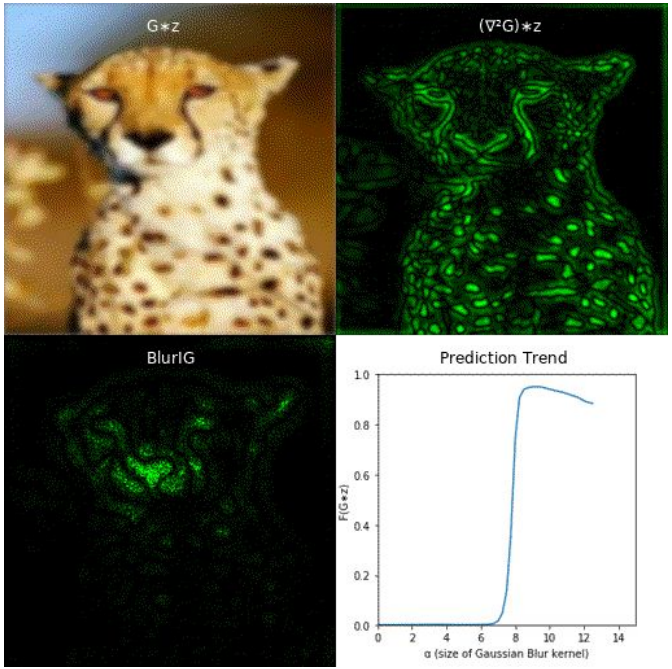
Video: https://www.youtube.com/watch?v=0iof_BMe1Q0

Questions?


TPR
FPR
OCT Based model
HE Based model
Retina Specialists: Overall Jugement
Retina Specialists: HE<500 microns
Retina Specialists: HE<1DD
Retina Specialists: HE<2DD

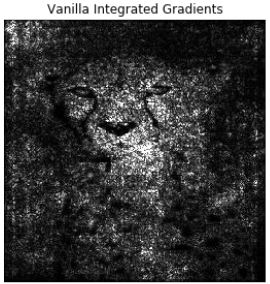

AUC
Fovea
Optic disc
AUC for full image
Extraction disc diameter multiplier
0.125
0.25
0.5
1.0
0.125
0.25
0.50
1.0

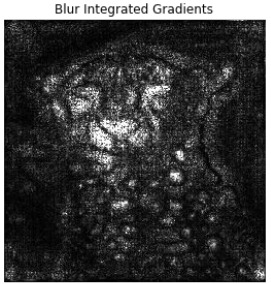

E
g
g
g
g

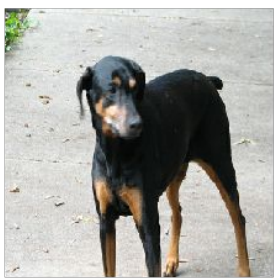

G∗z
(∇²G)∗z
BlurIG
Prediction Trend
F(G∗z)
α (size of Gaussian Blur kernel)

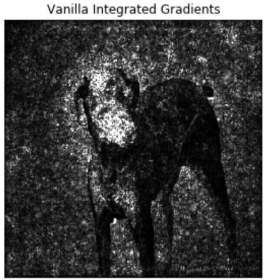
Google Research

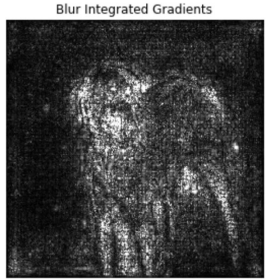
ImageNet (IG vs BlurIG)

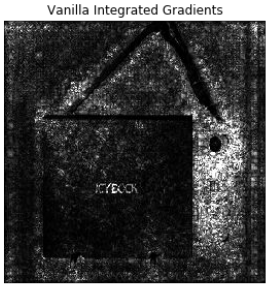


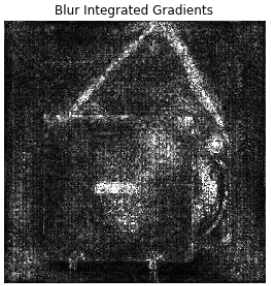


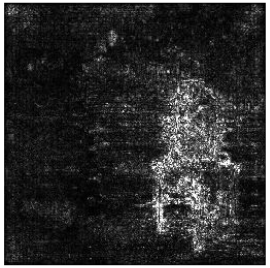












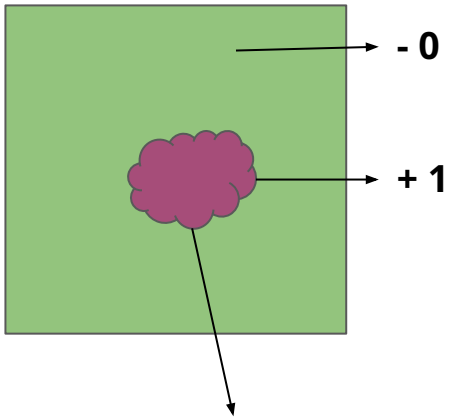
IG has color bias from choice of baseline.

BlurIG has a shape bias.

Quantitative Evaluation - Human Interpretability

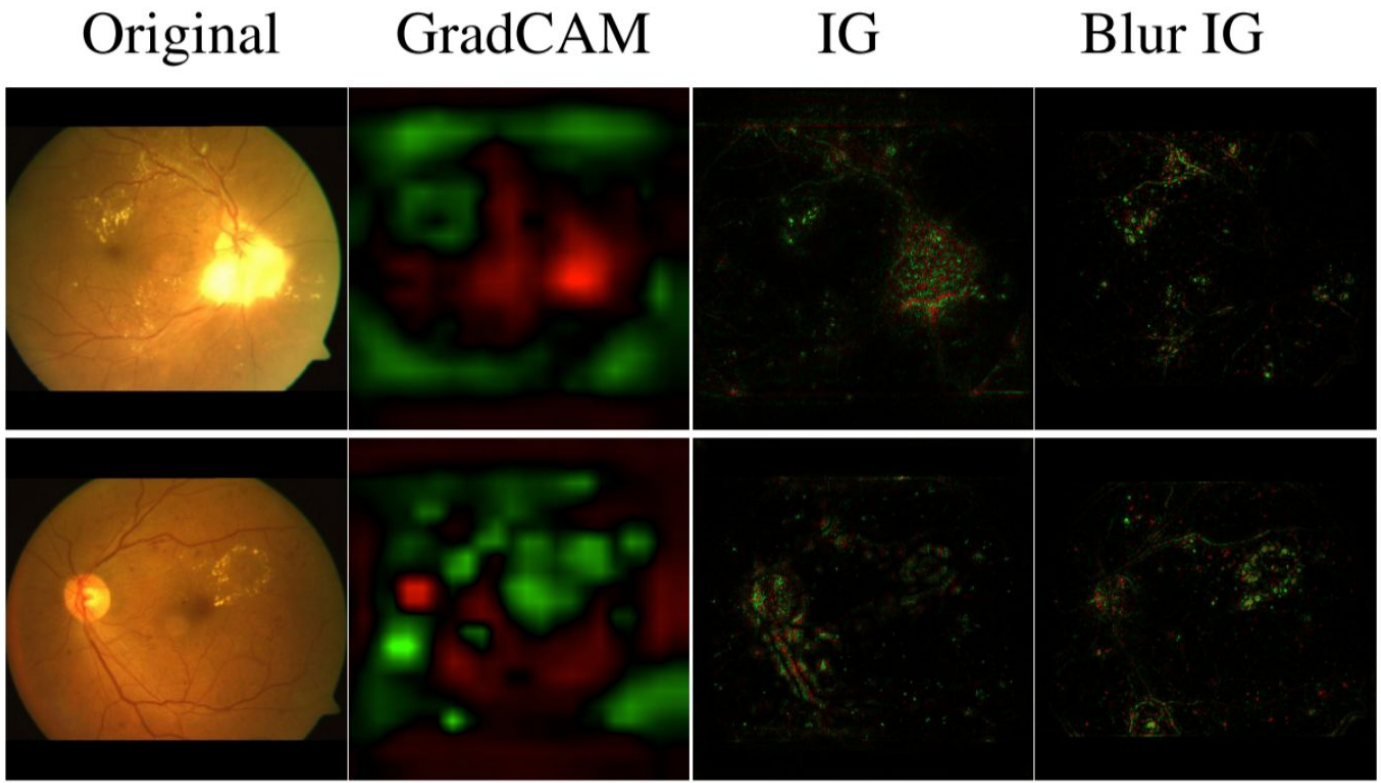
Attribution method	ImageNet AUC ↑	ImageNet F1 ↑	ImageNet MAE ↓	Diabetic Retinopathy AUC ↑	Diabetic Retinopathy F1 ↑	Diabetic Retinopathy MAE ↓
XRAI	0.836	0.786	0.149	0.805	0.285	0.068
GradCAM	0.742	0.715	0.194	0.817	0.249	0.058
IG (random-4)	0.709	0.674	0.223	0.827	0.344	0.060
IG (black)	0.710	0.674	0.219	0.828	0.307	0.062
IG (black+white)	0.729	0.681	0.216	0.818	0.296	0.062
Blur IG (ours)	0.738	0.693	0.209	0.831	0.293	0.061

Table 1: Average F1, AUC, and MAE scores for different explanation methods on images from ImageNet validation set (N=9684), and Diabetic Retinopathy dataset (N=141). (↑ indicates higher is better, ↓ indicates lower is better)



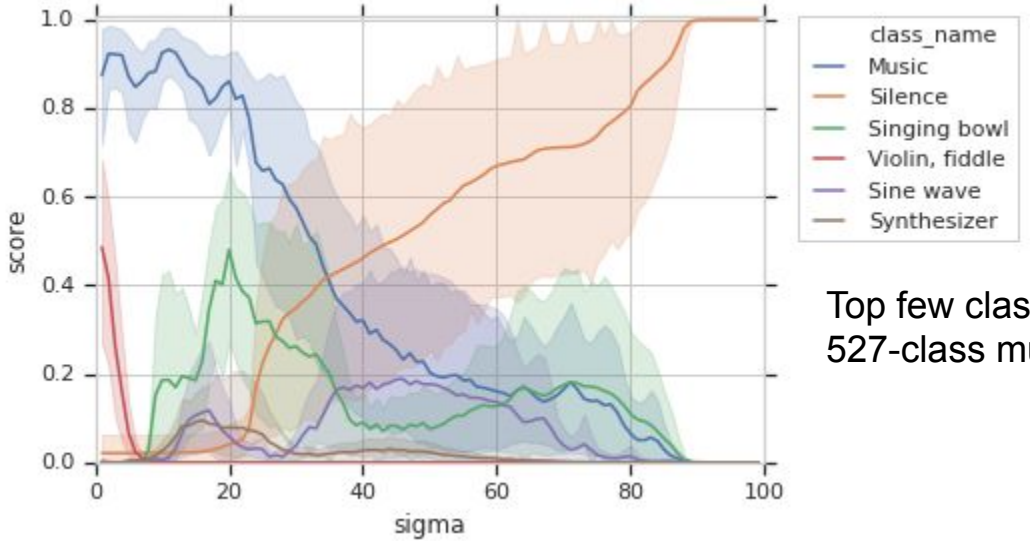
DR different from "natural" images

Diabetic Retinopathy



Green are positive attributions, and red negative.

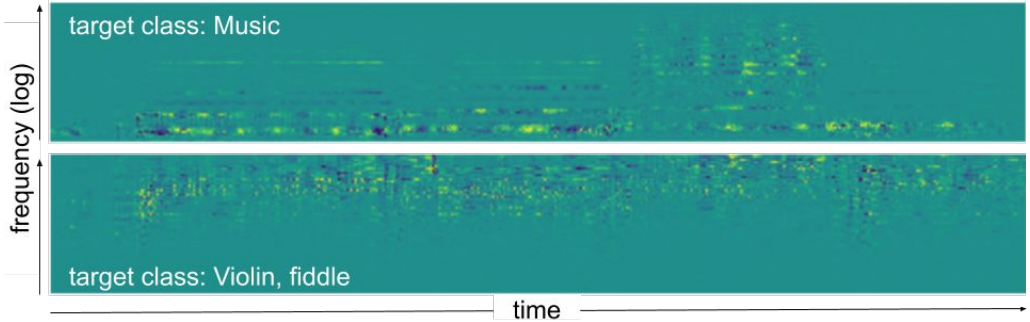
AudioSet - Predictions evolve to lower frequency classes.

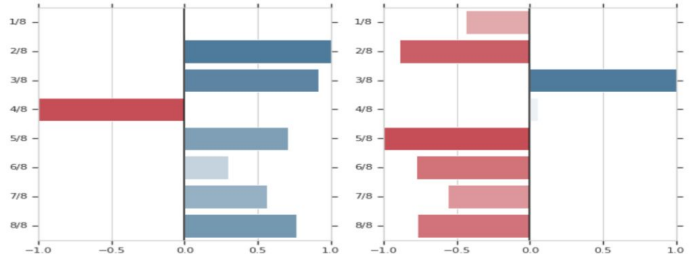


Top few classes from
527-class multilabel classifier

Visualizing the evolution of class prediction probabilities for prominent categories along the blur integration path from a few (7) violin audio samples. Y axis shows the confidence score, and X axis the sigma for the gaussian blur kernel. Color indicates the class. Initially model has higher confidence on violin and string instrument classes. With increased blur, confidence shifts towards singing bowl, sine wave, and then silence.

Class conditioning: Prediction is music not because it is violin!



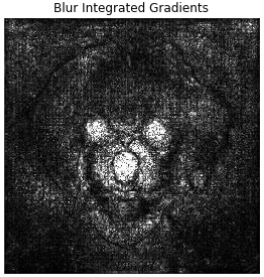


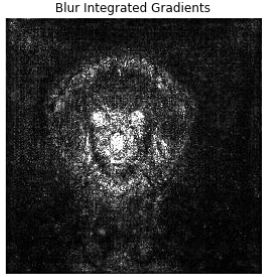
(a) Target class: 'Music' (b) Target class: 'Violin, fiddle'

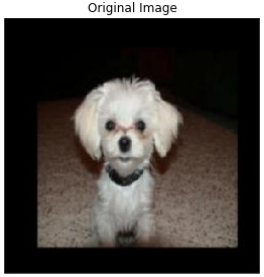
Figure 12: Aggregation of Blur IG contributions on 7 violin audio samples with target class as 'Music' (*left*) and target class as 'Violin, fiddle' (*right*). Y axis depicts the frequency bins, and X axis the integrated gradients. Blue indicates positive gradients/contributions and red the negative gradients. Explanation for class Music shows positive contributions from the lower frequencies while explanation for 'Violin, fiddle' shows positive contributions only from the higher frequencies.

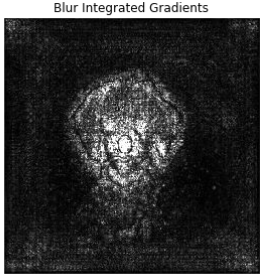
Google Research

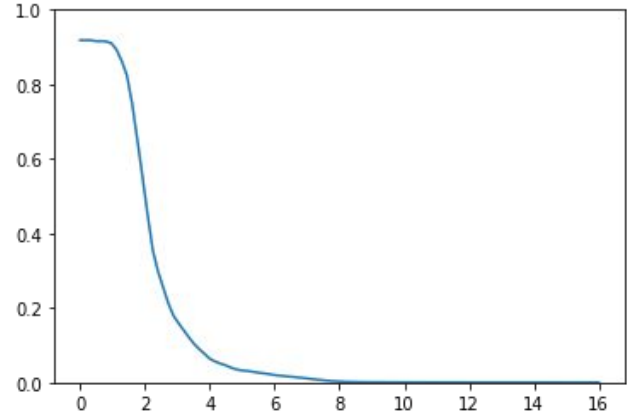
Effect of Crop and Zoom

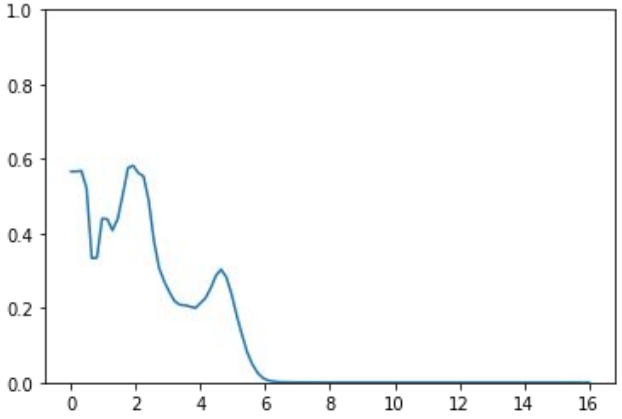


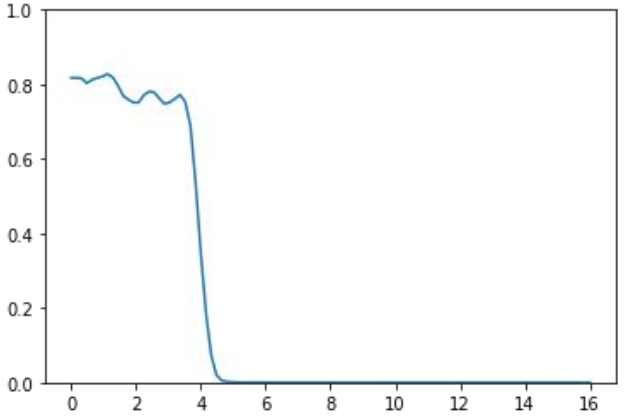






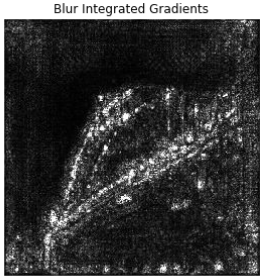


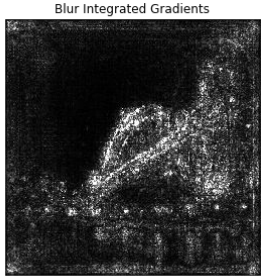


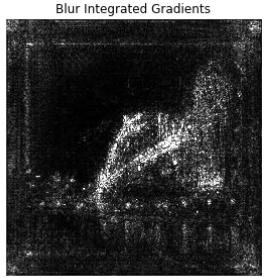


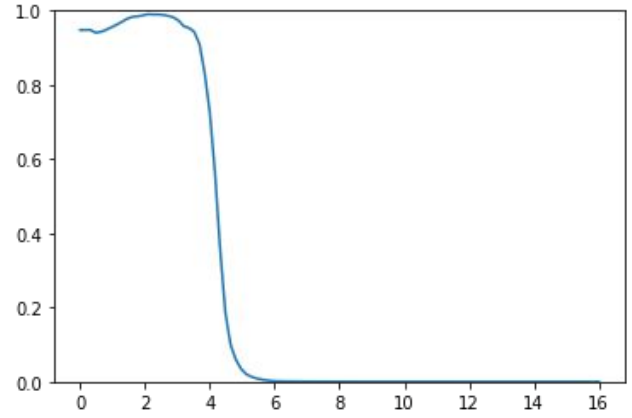
Google Research

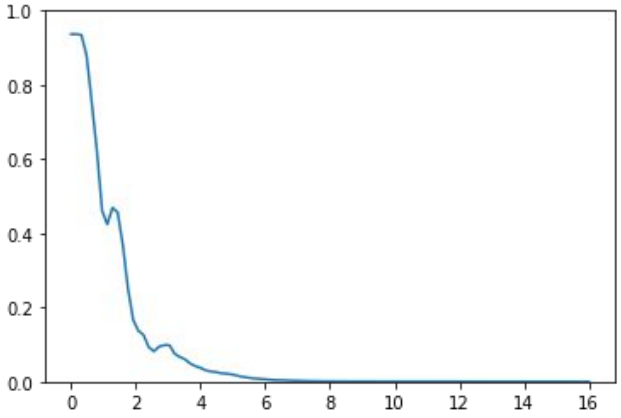
Effect of crop and zoom

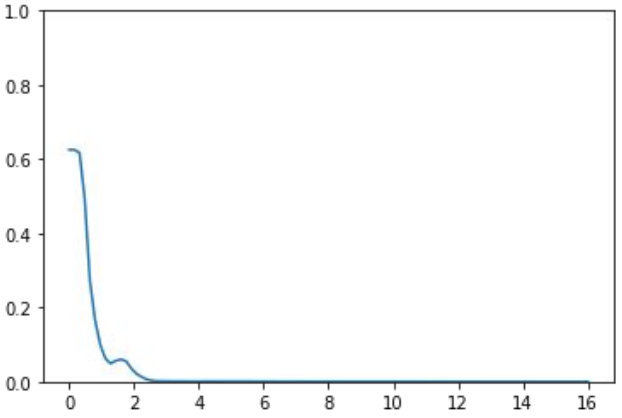






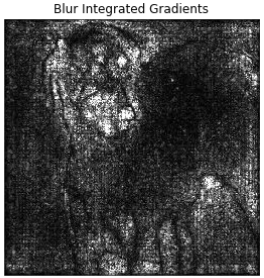


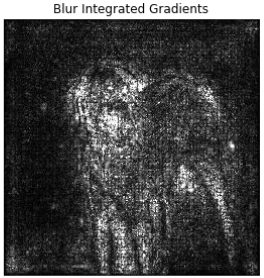


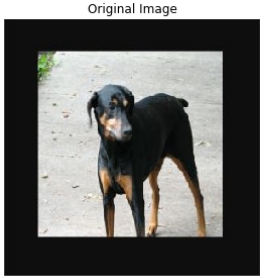


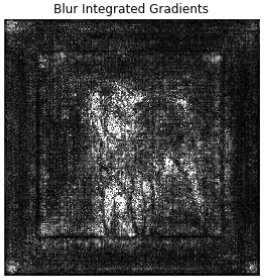
Google Research

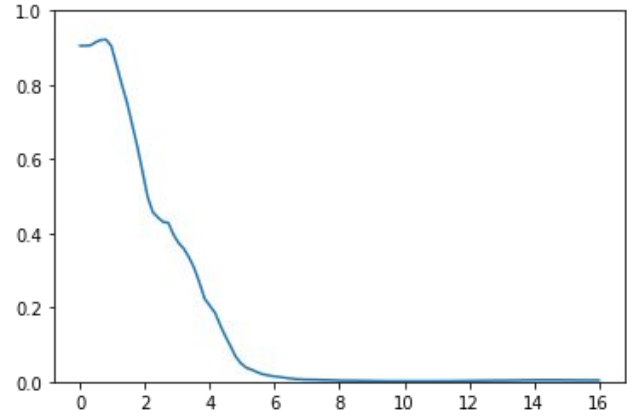
Effect of crop and zoom

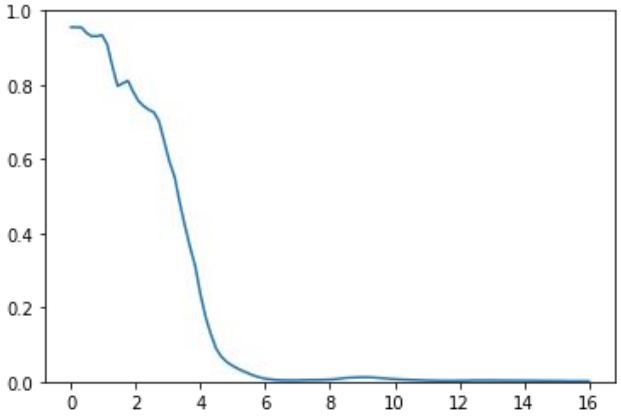


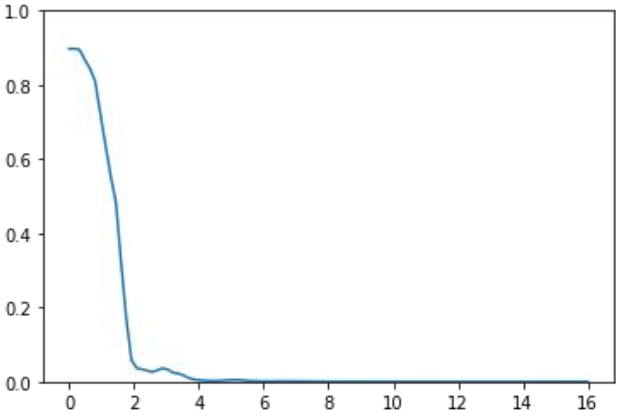












Links

Google Research

Paper: https://arxiv.org/abs/2004.03383

Code: https://github.com/PAIR-code/saliency

Video:
https://www.youtube.com/watch?v=0iof_BMe1Q0

{jinhuaxu@,vsubhashini@,mukunds@}google.com

* Background

* Scientific Discovery by Generating Counterfactuals using Image Translation

Narayanaswamy*, Venugopalan*, et. al. MICCAI 2020

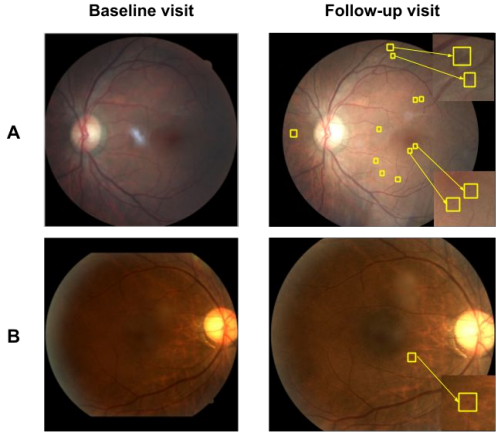
* Attribution in Scale and Space

Xu, Venugopalan, Sundararajan CVPR 2020

* **Predicting risk of developing diabetic retinopathy using deep learning**

Bora et. al. Lancet Digital Health 2021

Predicting risk of DR (6m - 2yrs) before symptoms show



Predicting risk of developing diabetic retinopathy using deep learning

Ashish Bora, Siva Balasubramanian, Boris Babenko, Sunny Virmani, Subhashini Venugopalan, Akinori Mitani, Guilherme de Oliveira Marinho, Jorge Cuadros, Paisan Ruamviboonsuk, Greg S Corrado, Lily Peng, Dale R Webster, Avinash V Varadarajan, Naama Hammel, Yun Liu, Pinal Bavishi

Lancet Digital Health 2021

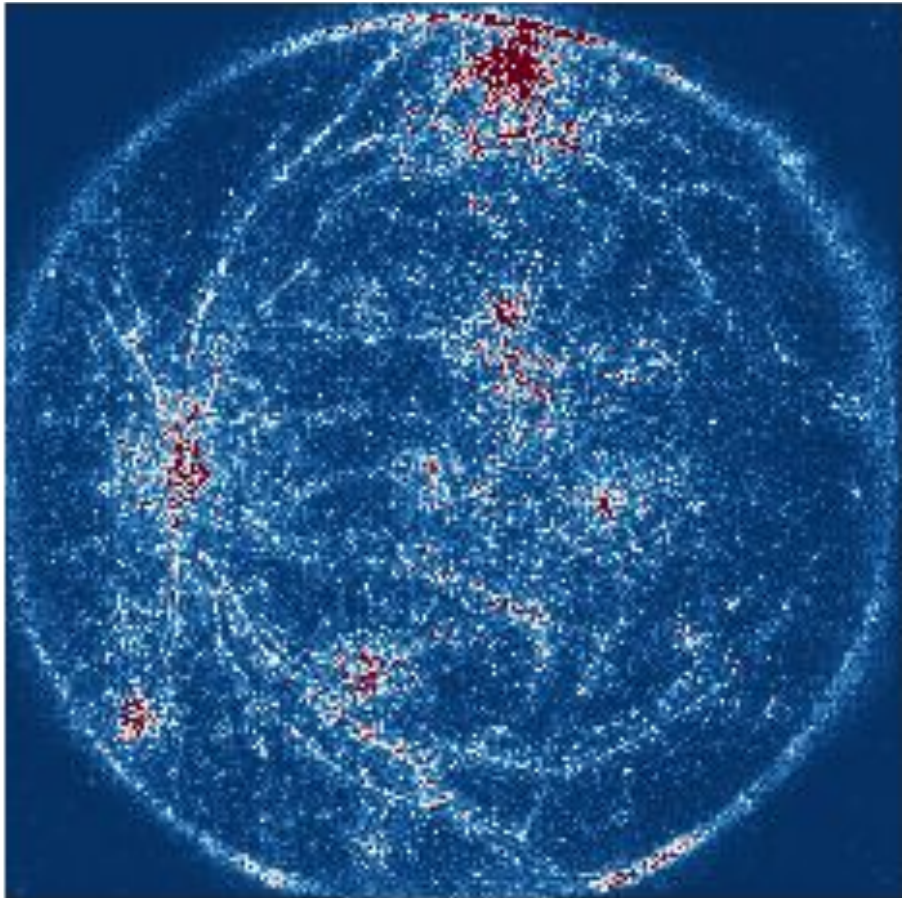
Baseline visit - NO DR

At baseline:

Doctor and DR Model : NO DR

Risk of DR (DR progression model): 0.72

BlurIG - pixel attribution

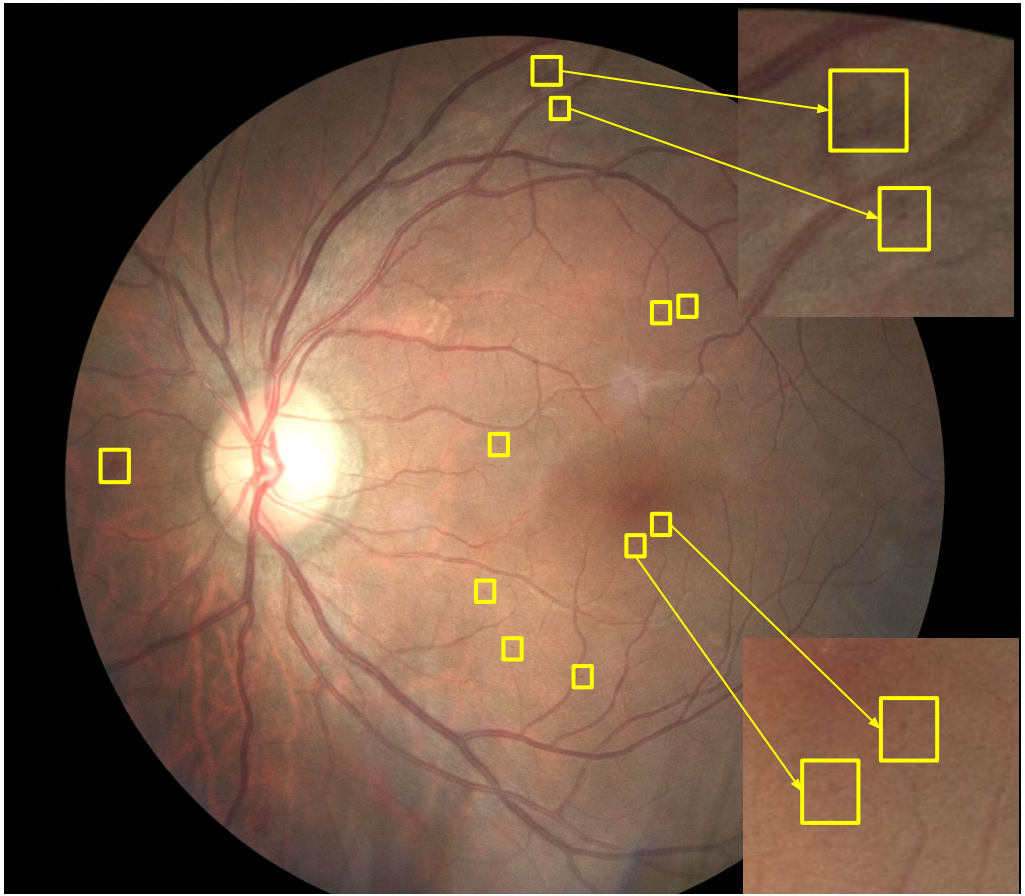


Pixel attribution on image from baseline visit.
Red indicates regions that the model is looking at.

Follow-up visit +1.5yrs: Attributed regions develop symptoms

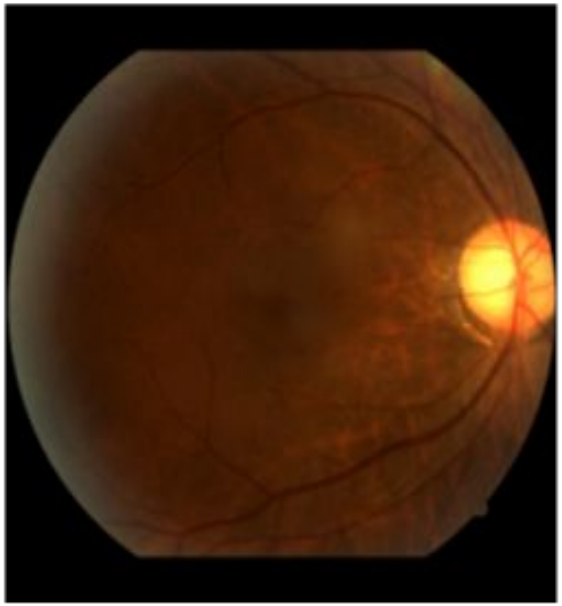
DR Model: Mild+ DR

Yellow boxes highlight microaneurysms identified by retina specialist.

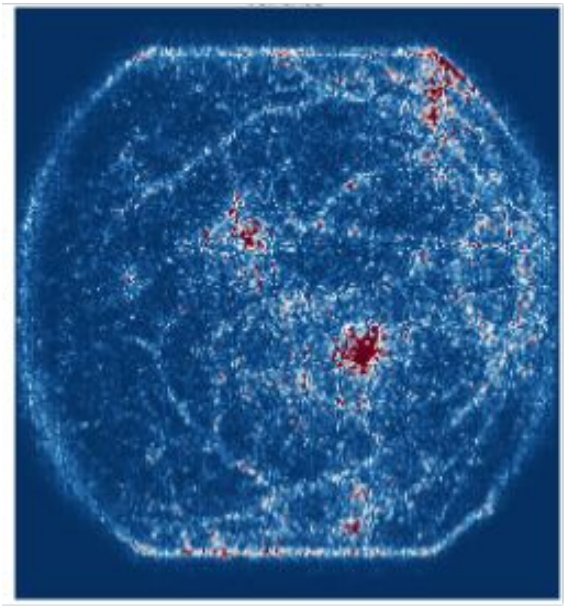


Attribution highlights where symptoms show-up in future

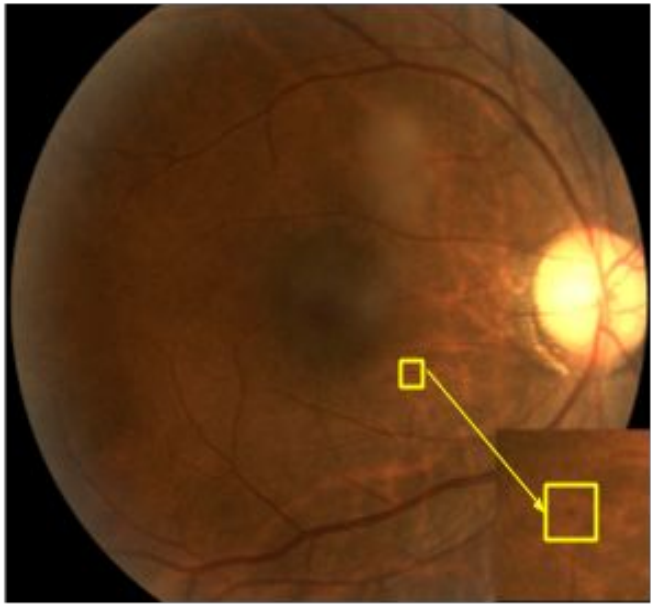
Baseline visit



BlurIG



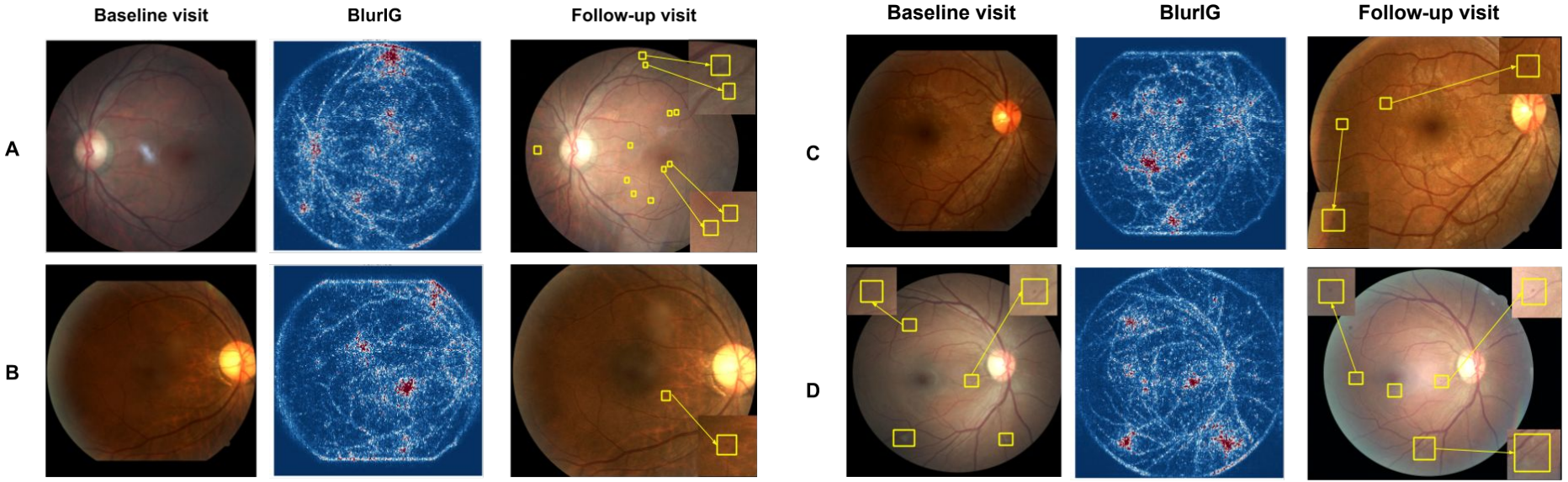
Follow-up visit



Predicting risk of DR (6m - 2yrs) before symptoms show

Interpretability techniques (BlurIG and IG) localize on regions where micro-aneurysms develop in the future.

(At baseline visit patient has NO DR)



Bora et. al. Lancet Digital Health '21