دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

Trustworthy AI

تمرین شماره ۲

طراحان: پیمان باقرشاهی، پرهام زیلوچیان مقدم

زمان تحویل: **۱۴۰۲/۰۲/۱۷**

بهار ۱۴۰۲

فهرست

پرسش ۱ – SHAP

استفاده از مقادیر Shapley یکی از روش های در نظر گرفتن تاثیر ویژگی های مختلف در خروجی حاصل از همبستگی (coalition) ویژگی های دیگر است. در همین راستا، مقادیر SHAP (SHapley Additive exPlanations) یک روش کارا برای توضیح عملکرد مدل ها است.

الف: ابتدا در رابطه با مقاله SHAP به سوالات زیر پاسخ دهید:

۱. با تعریف یک روش additive feature attribution سه ویژگی منحصر به فرد local accuracy, missingness و consistency روش SHAP را به صورت خلاصه معرفی کنید.

۲. برای مقابله با پیچیدگی بالای محاسباتی مقادیر SHAP، روش model-agnostic به نام Kernel SHAP معرفی شده است. نحوه عملکرد این روش را در مقایسه با محاسبه دقیق مقادیر SHAP بیان کنید.

۳. در کنار روش های model-agnostic این مقاله روش model-specific با عنوان Deep SHAP را برای مدل شبکه های عصبی معرفی میکند. تفاوت این روش با Kernel SHAP را بررسی کنید.

ب: در این بخش هدف استفاده از Deep SHAP و Kernel SHAP برای توضیح عملکرد یک مدل رگرسیون خطی است. ابتدا دادگان Life Expectancy را دریافت کنید. یک مدل رگرسیون ساده با معماری دلخواه برای پیش بینی سن امید به زندگی پیاده کنید که عملکرد نسبتا مطلوبی داشته باشد. 10 درصد دادها را برای تست و باقی را برای آموزش بگیرید (در هر قاره حتماً از 3 کشور حداقل یک نمونه در دادگان تست داشته باشید).

با استفاده از پکیج آماده shap و با دو روش Kernel SHAP و Deep SHAP مقادیر SHAP را با تابع summary_plot برای تمام نمونه‌های تست و تمام ویژگی های مدل بدست آورید تا تاثیر هر کدام در خروجی مشخص شود. نتایج خود را بررسی کنید.

حال از دو کشور در یک قاره، یک سمپل به دلخواه انتخاب کنید و نمودار force_plot را برای آن رسم کنید. نتیجه رو نمایش دهید توضیحی مختصری روی آن بدهید (تنها تحلیل مقادیر اهمیت دارد، تحلیل‌های دیگر اختیاری هستند).

پرسش ۲ – Knowledge Distillation

مقاله زیر را مطالعه کنید:

Distilling a Neural Network Into a Soft Decision Tree

۱. خلاصه ای از مزایای مدل معرفی شده در این مقاله نسبت به شبکه های عصبی را بیان کنید و علت آن بیان کنید.

۲. چگونه این مدل به جای یک سلسله از ویژگی ها (hierarchy of features)، با یک سلسله از تصمیم ها (hierarchy of decisions) کار میکند؟

۳. در رابطه با تابع هزینه مدل بحث کنید و تفاوت آن ها را با تابع هزینه cross-entropy مقایسه کنید.

۴. علت اضافه کردن ترم regularization در این مدل چیست؟

پرسش ۳ – D-RISE

در این سوال قصد داریم تا به بررسی Object Detectorها با استفاده از Saliency Mapها بپردازیم. بدین منظور ما مقاله D-RISE را انتخاب کرده‌ایم.

از طریق لینک زیر می‌توانید به این مقاله دسترسی پیدا کنید:

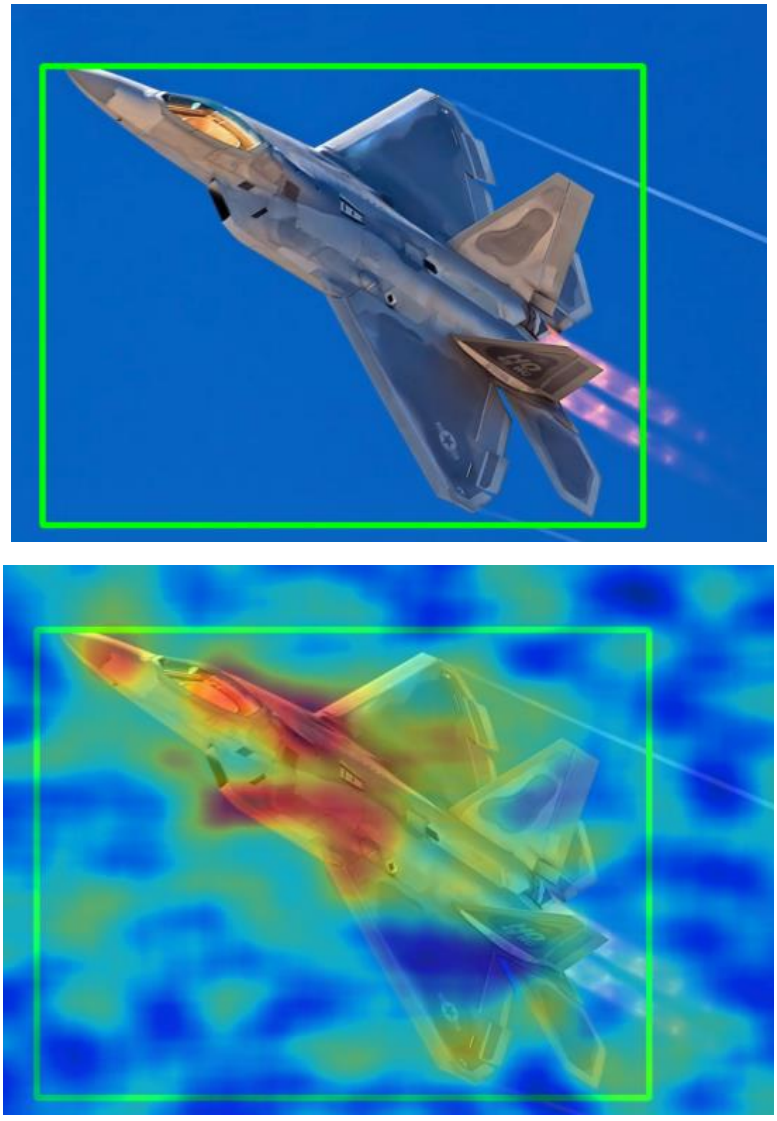
Black-box Explanation of Object Detectors via Saliency Maps

a) ابتدا مقاله را مطالعه کرده و یک خلاصه‌ای از ایده کلی و متمایز کننده آن نسبت به روش‌های دیگر ارائه دهید. همچنین بیان کنید که علت دست یافتن به این روش چه بوده است و روش‌های مشابه مبتنی بر شبکه‌های عصبی چه مشکلاتی داشته‌اند.
b) الگوریتم استفاده شده در این مقاله برای تولید Mask را بیان کنید و توضیح دهید.
c) معیار Similarity استفاده شده در این مقاله را توضیح دهید و علت انتخاب این روش را به اختصار توضیح دهید.
d) (اختیاری) چنانچه پیشنهادی برای بهبود روش ارائه شده (مخصوصا از نظر سرعت، عملکرد و کارایی) در این مقاله در طول مطالعه آن به ذهن شما رسیده است بیان کنید. همچنین در صورتی در طول مطالعه مقاله به مواردی جهت نقد مقاله و روش ارائه شده در آن برخورد کرده‌اید، می‌توانید در اینجا به آن‌ها اشاره کنید.

e) از طریق این لینک می‌توانید به نوتبوک مربوط به نسخه آزمایشی این مقاله دسترسی پیدا کنید. این نوتبوک را روی سیستم خود دریافت کرده یا آن را در گوگل کولب باز کنید. سپس از بین دسته‌های زیر سه دسته را انتخاب کرده و تصاویری مرتبط با آن را در Google Images یا هر پلتفرم مشابهی جستجو کنید و یک تصویر را از هر دسته به انتخاب خود دریافت کنید و به عنوان ورودی به برنامه داده و خروجی آن را مشاهده و دریافت کنید و آن را در گزارش خود قرار دهید و در صورت نیاز آن را تحلیل کنید. (این تحلیل باید شامل این باشد که به عنوان مثال چنانچه برنامه در تشخیص ناحیه‌های مربوطه اشتباه کرده است علتی که موجب آن شده است به نظر شما چه بوده است و همچنین این که دقت مدل در انجام این کار و کارایی مدل تا چه مقدار راضی کننده هست)

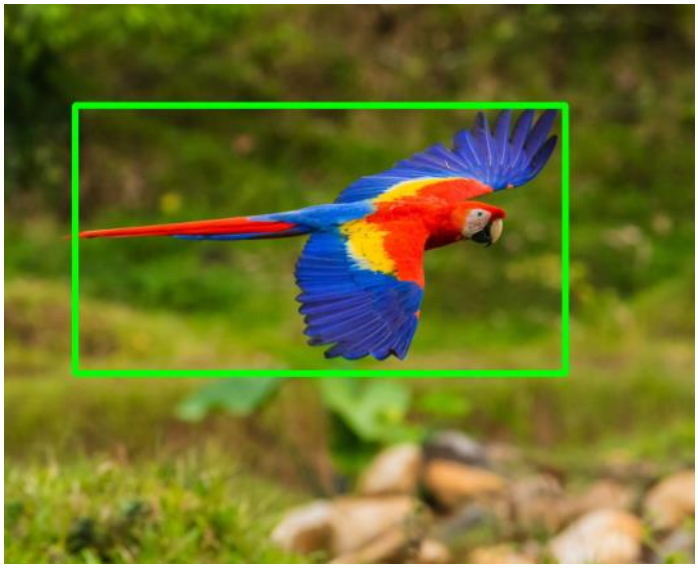
نمونه‌های از خروجی‌های این برنامه را می‌توانید در ادامه مشاهده کنید.

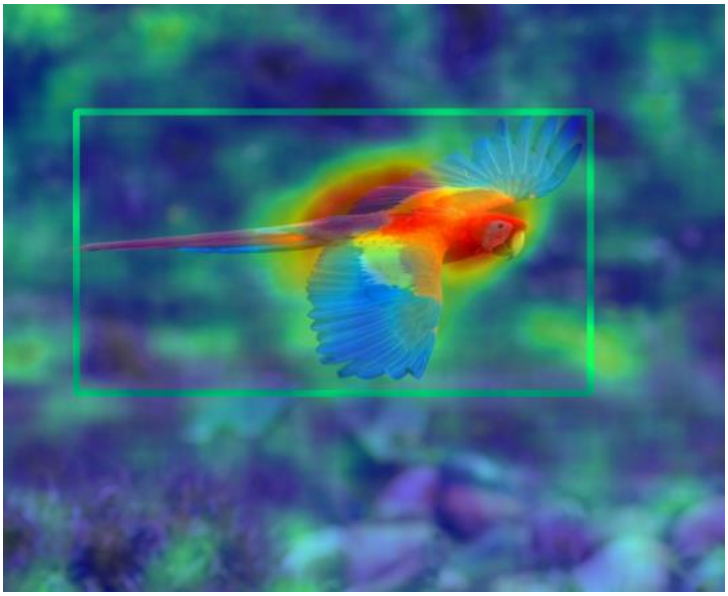
نمونه اول:



شکل ۱: تشخیص **Saliency**

نمونه دوم:





شکل2 : تشخیص **Saliency**

Label Name Choices: {'person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet', 'tv', 'laptop',

'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush'}

پرسش ۴ – LIME

در این پرسش قصد داریم تا با ساز و کار و نحوه عملکرد LIME[۱] آشنا شویم. همان‌طور که از اسم آن مشخص است این روش Model-agnostic هست و مدل را یک موجودیت black-box در نظر می‌گیرد. بنابراین از این روش می‌توان برای تفسیر هر مدل یادگیری ماشینی بهره برد.

ایده خلاصه روش Lime این است که ما در ناحیه محلی هر پیش‌بینی جداگانه زوم می‌کنیم؛ بنابراین می‌توانیم یک توضیحی که در مورد آن ناحیه محلی صدق می‌کند را ارائه دهیم. با بهره‌گیری از این روش، ما دیگر نگرانی‌ای از بابت بقیه مدل نداریم و همزمان یک توضیحی معتبری در مورد این که چرا این پیش‌بینی توسط مدل انجام شده است، در اختیار داریم. جهت توضیحات بیشتر می‌توانید مقاله این روش را از طریق این لینک مطالعه کنید.

بدین منظور پیشنهاد می‌کنیم که از کتابخانه lime در پایتون استفاده کنید. البته می‌توانید خودتان نیز پیاده‌سازی آن را انجام دهید. همچنین چون در این سوال با تصویر سر و کار داریم پس باید از ماژول lime_image این کتابخانه استفاده نمایید. همچنین مدلی که ما به منظور بررسی انتخاب کرده‌ایم مدل MobileNet V2 هست. برای مشاهده و تست عملکرد این مدل می‌توانید از طریق این لینک اقدام کنید.

اکنون مراحل زیر را باید برای این پرسش تکمیل نمایید.

a) مدل آموزش داده شده MobileNet V2 را در فریمورکی که انتخاب کرده‌اید دریافت کنید. به عنوان مثال در PyTorch و در TensorFlow را می‌توانید در لینک‌های مشخص شده مشاهده کنید. این مدل دریافت شده بر روی دیتاست ImageNet آموزش دیده است. کلاس‌های موجود در این دیتاست را می‌توانید از طریق این لینک در اختیار داشته باشید.

b) اکنون با توجه به دسته‌های موجود در این دیتاست یک دسته را انتخاب کرده و در اینترنت برای تصاویر آن دسته جستجو کنید و یک تصویر را از میان آن‌ها انتخاب کرده و عملکرد مدل خود را بر روی آن بسنجید و در خروجی مدل خود ۵ دسته با بالاترین احتمال و مشابهت را نشان دهید.

c) پس از اطمینان از صحت کارکرد مدل بارگیری شده، اکنون زمان استفاده از پکیج lime هست. ماژول مربوط به کار با تصویر این کتابخانه را برای مدل تعریف کنید.

[۱] Local Interpretable Model-agnostic Explanations

d) اکنون با استفاده از پکیج skimage و خروجی‌های بدست آمده از lime_image می‌توانید boundaries را بر روی تصویر رسم نمایید.

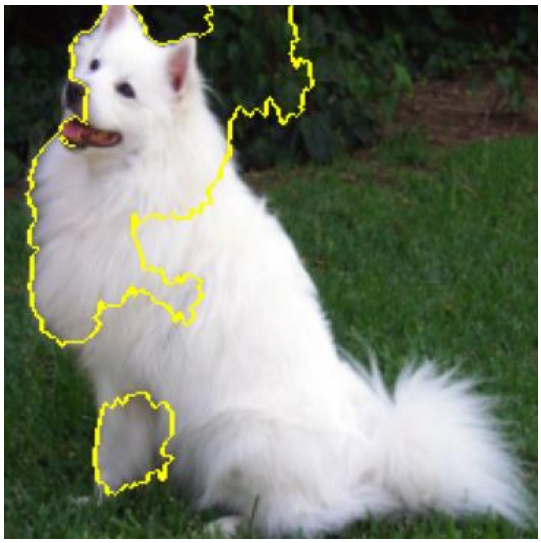
e) اکنون نواحی Pros and cons را بر روی تصویر و boundary های تشخیص داده شده اضافه کنید و در نظر بگیرید.

f) در انتها نیز نمودار Heatmap مربوط به تصویر را به همراه وزن‌های مربوطه را رسم کنید. بدین صورت می‌توانید اهمیت هر یک نواحی و میزان اهمیت آن‌ها را مشاهده کنید.

g) این کار را برای دو تصویر دیگر نیز تکرار کنید. سعی کنید برای یک تصویر که شامل مواردی از چند کلاس هست نیز این کار را انجام دهید.

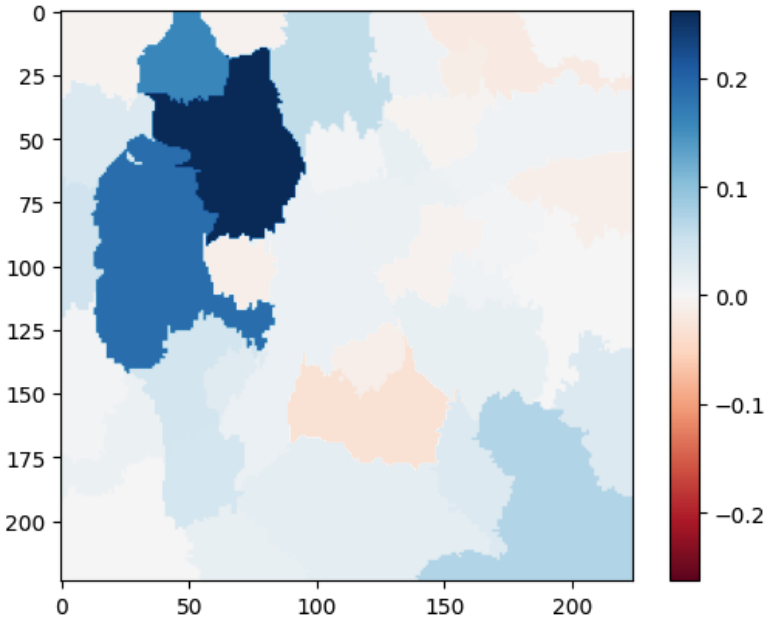
h) با توجه به نتایج به دست آمده خروجی‌های به دست آمده در هر یک از مراحل و برای هر یک تصاویر را تحلیل کنید.

در ادامه می‌توانید نمونه‌ای از خروجی مدل را بر روی یک تصویر مشاهده کنید:



شکل ۳: **boundary** رسم شده بر روی نواحی **Superpixel**

شکل ۴: مشخص کردن **Pros and cons**



شکل ۵: رسم **Heatmap** برای نمونه مربوطه

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز "یکشنبه ۱۷ اردیبهشت ماه" خواهد بود.
- این زمان به هیچ وجه قابل تمدید نیست و درصورت نیاز میتوانید از grace time استفاده کنید.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد.
- انجام این تمرین به صورت یک نفره می‌باشد.
- در صورت مشاهده هر گونه تشابه در گزارش کار یا کدهای پیاده‌سازی، این امر به منزله تقلب برای طرفین در نظر گرفته خواهد شد.
- استفاده از کدهای آماده بدون ذکر منبع و بدون تغییر به منزله تقلب خواهد بود و نمره تمرین شما صفر در نظر گرفته می‌شود
- در صورت رعایت نکردن فرمت گزارش کار نمره گزارش به شما تعلق نخواهد گرفت.
- تمامی تصاویر و جداول مورد استفاده در گزارش کار باید دارای توضیح (caption) و شماره باشند.
- بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است. خواهشا از هر گونه اطناب در گزارش کار پرهیز کرده و به موارد خواسته شده به صورت کامل پاسخ دهید.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW2_[Lastname]_[StudentNumber].zip

به طور مثال:

HW2_Zilouchian_12345678.zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع HW2_TAI با دستیاران آموزشی در ارتباط باشید:

- پرسش اول و دوم

p.baghershahi@ut.ac.ir یا تلگرام namyeP101@

- پرسش سوم و چهارم

p.zilouchian@ut.ac.ir یا تلگرام parham_zm@

شاد و سلامت باشید.