به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس هوش مصنوعی قابل اعتماد

مدرس: دکتر مصطفی توسلی‌پور

تمرین شماره ۲

اردیبهشت ماه ۱۴۰۳

فهرست

در این تمرین قصد داریم تفسیرپذیری مدل شبکه عصبی را با چند روش مختلف با استفاده از داده‌های جدولی[1] بررسی کنیم. دیتاست مورد استفاده اطلاعات سلامت افراد و مبتلا بودن به دیابت را نشان می‌دهد. ما با کمک یک شبکه عصبی چند لایه[2] مدلی برای تخمین احتمال ابتلا به دیابت می‌سازیم، سپس مدل را با دو روش مختلف تفسیر می‌کنیم. در ادامه یک روش تفسیرپذیری دیگر به نام NAM آشنا می‌شویم.

بخش اول: بارگذاری داده و آموزش مدل

بارگذاری داده

۱) فایل دیتاست diabetes.csv پیوست شده به همراه تمرین را بارگذاری کنید، سپس به انجام تحلیل کاوشگرانه داده‌ها[3] بپردازید. در ادامه موارد خواسته شده را بدست آورید: **(۵ نمره)**

الف) وابستگی میان داده‌های موجود را با استفاده از ماتریس وابستگی[4] و نمودار pairplot نمایش دهید و تحلیل کنید. به غیر از ستون outcome کدام دو جفت ویژگی با یکدیگر وابستگی دارند؟

ب) توزیع افراد سالم و مبتلا به دیتابت را رسم کنید. به نظر شما توازن داده موجود به چه صورت است؟

ج) بررسی کنید که داده پرت در کدام ستون‌های دیتاست مشاهده می‌شود و همچنین پراکندگی داده‌ها را نمایش دهید. آیا این داده‌ها در دقت و تحلیل مدل مشکل ایجاد می‌کنند؟

۲) برای ستون‌های مختلف داده پیش‌پردازش‌های لازم مانند نرمال کردن داده را انجام دهید (برای این کار می‌توانید از ماژول sklearn.preprocessing کتابخانه scikit-learn استفاده کنید.). پس از اتمام پیش‌پردازش داده‌ها را با توزیع یکسان به سه دسته آموزش[5]، اعتبارسنجی[6] و تست[7] به نسبت ۷۰، ۱۰و ۲۰ تقسیم کنید. **(۱ نمره)**

---

[1] Tabular data
[2] MLP
[3] Exploratory data analysis
[4] Correlation matrix
[5] Train
[6] Validation
[7] Test

آموزش و ارزیابی مدل

۳) برای دیتاست ذکر شده مدل شبکه عصبی موجود در جدول ۱ را طراحی کرده و آن را آموزش دهید. در مرحله آموزش مدل، نمودار خطا و دقت را در هر ایپاک رسم کنید. سپس با استفاده از داده‌های تست مقادیر دقت[۱]، بازیابی[۲]، f1-score و ماتریس درهم‌ریختگی[۳] را گزارش کنید. **(۳ نمره)**

جدول 1. معماری مدل

Layer	Config
Linear	input_dim=8, output_dim=100, activation: Relu
Batch Norm	size=100
Linear	input_dim=50, output_dim=50, activation: Relu
Dropout	p=0.2
Linear	input_dim=50, output_dim=50, activation: Relu
Linear	input_dim=50, output_dim=20, activation: Relu
Linear	input_dim=10, output_dim=1

بخش دوم: تفسیر مدل

مراحل تفسیرپذیری زیر را برای داده‌های تست اجرا کنید.

روش LIME

۴) مدل آموزش دیده را با استفاده از دیتای تست توسط روش LimeTabularExplainer تحلیل کنید. برای این کار از کتابخانه LIME [۴] استفاده کنید. برای سه نمونه تصادفی از داده‌های تست با استفاده از تابع explain_instance (در کتابخانه LIME) اهمیت ویژگی‌های[۵] متفاوت را بررسی کنید. **(۷ نمره)**

[1] Precision
[2] Recall
[3] Confusion matrix
[4] https://github.com/marcotcr/lime
[5] feature importance

روش SHAP

۵) مدل آموزش دیده را با استفاده از دیتای تست توسط روش KernelExplainer تحلیل کنید. برای روش SHAP از کتابخانه shap [1] استفاده کنید. برای سه نمونه انتخاب شده در قسمت ۴، نمودار force_plot مربوط به آنها را با استفاده از کتابخانه shap رسم کرده و نتایج آنها را مانند سوال قبل (بررسی اهمیت ویژگی‌های متفاوت) تحلیل کنید. **(۷ نمره)**

۶) در این قسمت قصد داریم نتایج بدست آمده از مرحله ۴ و ۵ را با یکدیگر مقایسه کنیم: **(۵ نمره)**

الف) در میان نتایج بدست آمده از این دو روش چه شباهت و تفاوت‌هایی مشاهده می‌کنید؟

ب) به نظر شما کدام روش با دقت بیشتری اهمیت ویژگی‌ها را برای نمونه انتخابی پیش بینی کرده است؟ برای بررسی اهمیت ویژگی‌ها می‌توانید به صورت دستی مقدار آن ویژگی را با یک مقدار از همان ستون دیتاست که کلاس متفاوتی دارد، تعویض کرده و تاثیر آن را بررسی کنید.

ج) ارتباط میان خروجی دو روش LIME وSHAP (مراحل ۴و ۵) و ماتریس همبستگی [2] بدست آمده از مرحله ۱، را بررسی کرده و تحلیل خود را گزارش کنید.

۷) برای هر دو روش LIME و SHAP، مقادیر اهمیت ویژگی‌ها را برای داده تست محاسبه کنید. این مقادیر تا چه اندازه نزدیک به مقادیر نمونه‌های بدست آمده از مراحل ۴ و ۵ می‌باشند؟ **(۷ نمره)**

بخش سوم: مدل NAM

مدل NAM

۸) در این قسمت با استفاده از مدل NAM که در مقاله[1] معرفی شده، می‌خواهیم یک طبقه‌بند تفسیرپذیر طراحی کنیم:

الف) ابتدا مقاله معرفی شده را مطالعه کنید و تفاوت‌های مدل NAM نسبت به یک مدل black-box هوش مصنوعی را توضیح دهید. استفاده از این مدل چه مزایا یا معایبی از نظر عملکرد و تفسیرپذیری مدل خواهد داشت؟ **(۴ نمره)**

[1] https://github.com/shap/shap

[2] correlation matrix

ب) یک مدل NAMClassifier که در مقاله معرفی شده، برای طبقه‌بندی دیتاست طراحی کنید (برای ساخت مدل از کتابخانه NAM[1]استفاده کنید). **(۳ نمره)**

ج) بررسی کنید آیا این روش توانسته تفسیرپذیری مدل را بهبود بخشد؟ و نتایج آن تا چه اندازه با روش‌های تفسیرپذیر مانند LIME و SHAP متفاوت می‌باشد. **(۸ نمره)**

بخش امتیازی

تفسیرپذیری روش GRACE

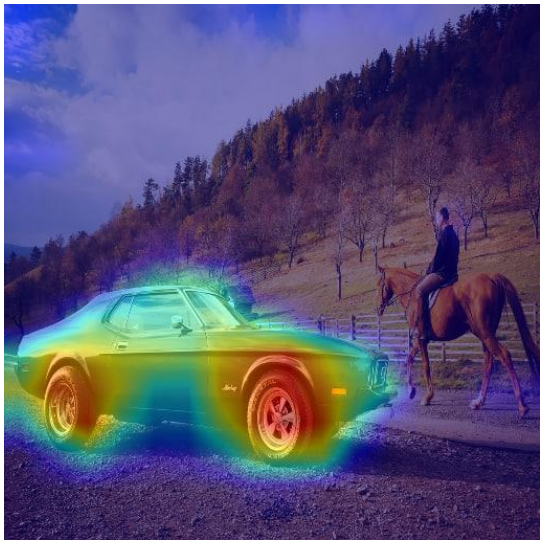
امتیازی) در روند تولید نمونه توسط روش GRACE، برای پیدا کردن نمونه جدید بررسی کنید که روند تغییرات احتمال پیش‌بینی[2] مدل برای هر تغییر به چه صورت است. بیشترین تغییر احتمال حاصل شده از تغییر ویژگی‌ها را پیدا کرده و دلیل احتمالی تغییر دقت را بررسی کنید (تحلیل این تغییرات را با استفاده از روش SHAP انجام دهید). برای آشنایی با روش GRACE مقاله[2] مطالعه کنید. از کتابخانه GRACE[3] برای پیاده سازی مراحل خواسته شده، استفاده نمایید. **(۵ نمره)**

[1] https://github.com/lemeln/nam
[2] Probability prediction
[3] https://github.com/lethaiq/GRACE_KDD20

سوال دوم – تفسیرپذیری در حوزه تصویر

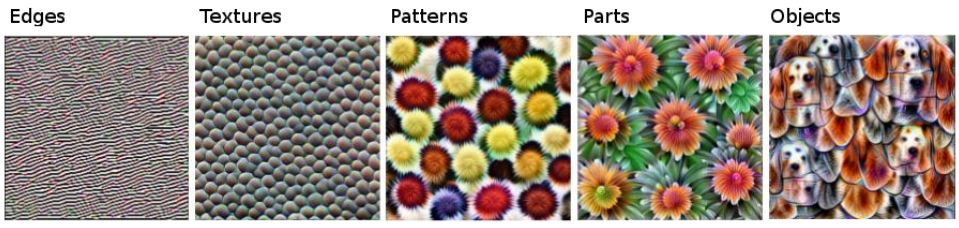
مقدمه

در حوزه یادگیری عمیق در بینایی ماشین روش‌های تفسیرپذیری مبتنی بر pixel attribution از جمله روش‌هایی هستند که در آن می توانند اهمیت هر pixel عکس را در تعیین کلاس یک داده در مسئله دسته‌بندی توسط شبکه‌های عمیق مشخص کنند. این تعیین اهمیت معمولا با تشکیل یک heatmap (که به آن Saliency map گفته می شود) از عکس مورد نظر ساخته می شود (مانند عکس زیر برای تشخیص یک ماشین)



شکل 1- نمونه ای از Saliency map

روش تفسیر پذیری Feature visualization از جمله روش هایی است که می توان یک تفسیری در مورد اینکه یک بخش از شبکه و یا بخش هایی از شبکه چه ویژگی از داده ورودی یاد گرفته اند به کار می رود. این روش ها عموما بر روش شبکه های عصبی کانولوشنی به کار می رود که به وسیله آن می توان فهمید که یک نورون خاص، یک فیلتر خاص و یک لایه خاص در شبکه عصبی کانولوشنی چه ویژگی هایی از تصاویر ورودی از داده ها را یاد گرفته است.



شکل 2- نمونه ای از Feature visualization در شبکه عصبی کانولوشنی

در مورد انجام این سوال از تمرین به نکات گفته شده زیر توجه داشته باشید:

1- از مدل پیش آموزش دیده شده VGG16 روی مجموعه داده ImageNet برای این تمرین استفاده کنید.
2- روش‌هایی که در ادامه گفته می شود را روی ۶ عکس(از کلاس های متفاوت) به انتخاب خودتان اجرا کنید و saliency mapهای هر عکس را گزارش کنید. فقط توجه داشته باشید که عکس‌های انتخابی توسط دسته بند به درستی کلاس عکس را دسته بندی کند.
3- گزارش شما برای این سوال از اهمیت زیادی دارد، لذا گزارش خود را به صورت دقیق و بدون ابهام بنویسید همچنین باید بخش مهم از کد که مرتبط پیاده سازی هر کدام از روش‌ها هست توضیح داده شود.
4- خروجی‌های کد در فایل Jupyter Notebook خود را پاک نکنید و باید منطبق با آن چیزی باشد که در گزارش خود می آورید.
5- عکس‌های انتخابی خود را به همراه کد قرار دهید، چون کدهای زده شده بررسی و اجرا خواهد شد. و ممکن است با عکس‌هایی غیر از عکس‌های که نتایج آن گزارش دادید کد چک شود. لذا از اجرا کامل کد و بدون خطا آن اطمینان حاصل کنید.

GRADIENT BASED PIXEL ATTRIBUTION

در روش‌های pixel attribution مبتنی بر گرادیان، از مقدار گرادیان و مثبت یا منفی بودن آن به عنوان معیاری برای تعیین اهمیت ویژگی‌ها (مثلا پیکسل‌های تصویر) برای مسئله دسته بندی (عموما برای مسئله دسته بندی به کار می رود ولی لزوما محدود به این نوع مسئله نیست) به کار می رود.

GRAD-CAM-1

مقاله Grad-CAM [3] را مطالعه کنید و به سوال‌های زیر پاسخ دهید.

1-1- ایده روش Grad-CAM برای ایجاد saliency را به صورت دقیق و با ذکر جزئیات روابط آن بیان کنید. **(3 نمره)**

1-2- روش Grad-CAM را روی مدل مورد نظر خود پیاده سازی کنید و بر اساس عکس‌ها و مدل saliency mapها را تولید و گزارش کنید. **(5 نمره)**

GUIDED GRAD-CAM-2

مقاله Grad-CAM [3] را مطالعه کنید و به سوال‌های زیر پاسخ دهید.

2-1- در مورد روش Guided backpropagation که در مقاله بیان شده است تحقیق کنید و مزیت روش Guided backpropagation نسبت به روش Backpropagation معمولی برای ایجاد saliency mapها بیان کنید. **(۳ نمره)**

2-2- روش Guided-backpropagation را روی مدل مورد نظر خود پیاده سازی کنید و بر اساس عکس‌های saliency mapها را تولید و گزارش کنید. **(۵ نمره)**

2-3- حال ایده روش Guided-CAM را بیان کنید و هدف اینکه این روش Grad-CAM با روش Guided-backpropagation ترکیب شده است را بیان کنید. **(۲ نمره)**

2-4- روش Guided Grad-CAM را روی مدل مورد نظر خود پیاده سازی کنید و بر اساس عکس‌ها و مدل saliency mapها را تولید و گزارش کنید. **(۵ نمره)**

3- SMOOTHGRAD

3-1- مقاله SmoothGrad [4] را مطالعه کنید. ایده بیان شده در این مقاله را بیان کنید و بیان کنید این مقاله سعی داشته است چه چیزی را بهبود دهد. **(۵ نمره)**

3-2- روش Guided-backpropagation را با ایده گفته شده در مقاله SmoothGrad ترکیب کنید و آن را روی مدل خود پیاده سازی کنید و بر اساس عکس‌های انتخابی saliency mapها را تولید و گزارش کنید. در مورد تعداد نمونه و مقدار اضافه کردن نویز در روش SmoothGrad از مقادیری که در مقاله عنوان شده می توانید برای مقادیر اولیه این پارامترها برای شروع استفاده کنید و همچنین می توانید به صورت دلخواه برای رسیدن به نتیجه بهتر این مقادیر را تغییر دهید. **(۵ نمره)**

3-3- حال روش Smoothgrad + Guided-backpropagation که در بخش قبل پیاده سازی کردید با روش Grad-CAM ادغام کنید. این کار را به همان صورتی روش Guided-backpropagation با روش Grad-CAM به منظور ایجاد Guided Grad-CAM ادغام شد، انجام دهید. بر اساس عکس‌ها و مدل saliency mapها را تولید و گزارش کنید. **(۵ نمره)**

4- ADVERSARIAL PETREBUTION AND PIXEL ATTRIBUTION

4-1- یکی از عکس‌های sample images به صورت دلخواه انتخاب کنید و با روش حمله دلخواه (مانند FGSM، PGD و ...) عکسی تولید کنید که به صورتی که کلاس پیش بینی شده آن عکس مورد نظر

عوض شود سپس Saliency map آن مربوط به کلاس اصلی عکس بدست بیاورید و با Saliency map در حالتی که عکس فاقد Adversarial perturbation است مقایسه کنید. برای بدست آوردن Saliency mapها از روش بخش قبل استفاده کنید. **(2 نمره)**

FEATURE VISUALIZATION

در این بخش قصد داریم برای کلاس مرغ (Hen) در مدل VGG16 از پیش آموزش دیده شده ImageNet داریم در فضای داده، تصویری را پیدا کنیم که به صورتی که کلاس پیش بینی شده توسط مدل برای کلاس مرغ بیشینه شود (روش Activation maximization).

$$\hat{img}^* = \arg\max_{img} h_{n,x,y,z}(img)$$

۱- در ابتدا روی logits (قبل Softmax) کلاس Hen، بهینه سازی روی داده ورودی به صورت که انجام دهید که مقدار این logits را بیشینه کند و تصویر تولیدی را گزارش کنید. **(3 نمره)**

۲- توضیح دهید به چه دلیل تصویر تولیدی، تصویر با معنایی نشده است **(3 نمره)**

۳- حال برای اینکه تصویر با معنایی تولید شود علاوه بر بهینه سازی گفته شده در بالا موارد زیر را به آن اضافه کنید:

1- Total variance regularization
2- Random shift در هر forward pass (پیکسل ها را جا به جا کنید به صورت رندوم نسبت به محور x و y)

در مورد روش Total variance regularization برای داده تصویر مختصرا توضیح دهید و همچنین بیان کنید به چه دلیل دو روش بیان شده بر اینکه بتوانیم تصاویر با معنایی تولید کنیم کمک کننده است. **(4 نمره)**

۴- حال با موارد گفته شده ۵ تصویر تولید کنید و آن ها را گزارش دهید. **(5 نمره)**

توجه:

در مورد روش های گفته شده در مورد ۳ برای بدست آوردن تصویر می توانید از روش های دلخواه خود استفاده کنید ولی به موارد زیر توجه داشته باشید:

۱- در مورد روش دلخواه توضیح کافی در مورد آن بدهید و علت موثر بودن آن را بیان کنید.
۲- در مورد روش های گفته شده در بخش ۳، سوای استفاده از روش دلخواه هر دو روش گفته شده در قسمت ۳ در مورد موثر بودن آن توضیح داده شود.

مراجع

[1] R. Agarwal *et al.*, “Neural Additive Models: Interpretable Machine Learning with Neural Nets,” *arXiv.org*, Apr. 29, 2020. https://arxiv.org/abs/2004.13912v2

[2] T. Le, S. Wang, and D. Lee, “GRACE: Generating concise and informative contrastive sample to explain neural network model’s prediction,” *arXiv.org*, Nov. 05, 2019. https://arxiv.org/abs/1911.02042

[3] - Selvaraju, R. R., A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. "Grad-cam: Why did you say that? Visual explanations from deep networks via gradient-based localization. arXiv." *arXiv preprint arXiv:1610.02391* (2016).

[4] - Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. "Smoothgrad: removing noise by adding noise." arXiv preprint arXiv:1706.03825 (2017).

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز "جمعه ۲۱ اردیبهشت ماه" خواهد بود.
- این زمان قابل تمدید نیست و درصورت نیاز می‌توانید از grace time استفاده کنید.
- در نظر داشته باشید که حداکثر مهلت آپلود تمرین در سامانه تا ۷ روز پس مهلت تحویل است و پس از آن سامانه بسته خواهد شد.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد و کدهای شما باید قابل اجرا بوده و به همراه گزارش آپلود شوند.
- 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.
- تحویل تمرین به صورت **دستنویس** قابل پذیرش نیست.
- 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.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید.

- HW2 _[Lastname]_[StudentNumber].zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع TAI_HW2 با دستیاران آموزشی در ارتباط باشید:
  - سوال اول

    a.jalalifar@ut.ac.ir یا تلگرام @g_1_9_98

  - سوال دوم

    smousavichashmi@ut.ac.ir (ممنون خواهم شد از طریق ایمیل اقدام شود)

با آرزوی سلامتی و موفقیت روزافزون