Supplementary Materials for Statistically Consistent Saliency Estimation

Shunyan Luo
George Washington University
shine_lsy@gwu.edu

Emre Barut
Amazon Alexa
ebarut@amazon.com

Fang Jin
George Washington University
fangjin@email.gwu.edu

A. Alternative Formulation

Our linear program can also be recast by a change of variables and setting $\alpha = Dg$. In this case, the elements of α correspond to differences between adjoint pixels. This program can be written as:

$$\begin{aligned} &\min \|\alpha\|_1 \\ \text{s.t.} &\left\| D^+ \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{f}\left(\tilde{x}_i\right) \tilde{x}_i - \Sigma D^+ \alpha \right) \right\|_\infty \leq L, \\ &U_2^T \alpha = 0, \end{aligned}$$

where D^+ is the pseudo-inverse of D and U_2 is related to the left singular vectors of D. More precisely, letting $D = U\Theta V^T$ denote the singular value decomposition of D, U_2 is the submatrix that corresponds to the columns of U for which Θ_j is zero. The linearity constraint ensures that the differences between the adjoint pixels is proper. Derivation of the alternative formulation follows from Theorem 1 in [3] and is omitted.

This formulation can be expressed in the standard augmented form, i.e. $\min_{Ax=b, x\geq 0} c^T x$, by writing $x = [\alpha_+, \alpha_-, s_+, s_-]^T$,

$$A = \begin{bmatrix} U_2 & -U_2 & 0 & 0 \\ -D^+\Sigma D^+ & D^+\Sigma D^+ & \mathbb{I}_{m\times m} & 0 \\ -D^+\Sigma D^+ & D^+\Sigma D^+ & 0 & -\mathbb{I}_{m\times m} \end{bmatrix},$$

$$b = \begin{bmatrix} 0 \\ L1_m - D^+ y \\ -L1_m - D^+ y \end{bmatrix}, c = \begin{bmatrix} 1_m \\ 1_m \\ 0 \\ 0 \end{bmatrix},$$

where $y = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}\left(\tilde{x}_i\right)\tilde{x}_i$ and $m = 2p_1p_2 - p_1 - p_2$. The γ coefficient in the original formulation can be obtained by setting $\gamma = D^+\left(\alpha_+ - \alpha_-\right)$.

B. Proof of Theorem 1

Our proof depends on the following lemma.

Lemma 1. *For $L \geq \sqrt{2\|D^+\|_1 \log\left(p_1p_2/\epsilon\right)/n}$, γ^* is in the feasibility set with probability $1-\epsilon$, that is*

$$\left\| D^+ \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{f}\left(\tilde{x}_i\right) \tilde{x}_i \right) - D^+ \Sigma \gamma^* \right\|_\infty \leq L.$$

Proof. For ease of notation, let $G = D^+\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \tilde{f}\left(\tilde{x}_i\right)\tilde{x}_i\right]$, and note that $G = D^+\Sigma\gamma^*$. Furthermore, let $z_i = \tilde{f}\left(\tilde{x}_i\right) D^+ \tilde{x}_i$. We also assume that the images have been rescaled so that the maximum value of $\tilde{x}_i$ is 1 (without rescaling, the maximum would be given as the largest intensity, i.e. 255). Since, the function values are also in the range given by [-2,2], we can bound $|z_{i,j}|$, that is

$$|z_{i,j}| = \left|\tilde{f}\left(\tilde{x}_i\right) D_j^+ \tilde{x}_i\right| \leq 2 \left\|D_j^+\right\|_1 \max_i |x_{i,j}| \leq 2 \left\|D_j^+\right\|_1 .$$

The proof follows by applying the McDiarmid's inequality [9] for each row of the difference and then taking the supremum over the terms. By application of McDiarmid's inequality, we have that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i z_{ij} - G_j\right| \geq L\right) \leq 2e^{\frac{-L^2 n}{2\|D^+\|_1}}.$$

Let $L = \sqrt{2\|D^+\|_1 \log\left(p_1 p_2/2\epsilon\right)/n}$. Then, taking a union bound over all variables, we have

$$\mathbb{P}\left(\max_j \left|\frac{1}{n}\sum_i z_{ij} - G_j\right| \geq L\right) \leq \sum_{j=1}^{p} e^{\frac{-L^2 n}{2\|D^+\|_1}} = \epsilon.$$

Now note that that the feasibility set for any $L^{'} \geq L$ contains that of L and thus γ^* is automatically included. □

We now present the proof of the theorem. Note that the technique is based on the Confidence Set approach by [2]. In the proof, we use γ to refer to vec(γ) for ease of presentation.

Proof. First, let the high probability set for which Lemma 2 holds by A. All of the following statements hold true for A. We let $\Delta = D\left(\hat{\gamma} - \gamma^*\right)$. We know that $\|D\hat{\gamma}\|_1 \leq \|D\gamma^*\|_1$ since both are in the feasibility set, as stated in Lemma 2. Let $\alpha^* = D\gamma^*$, $\hat{\alpha} = D\hat{\gamma}$ and define $S = \{j : \alpha_j^* \neq 0\}$, and the complement of S as S^C. By assumption of the Theorem, we have that the cardinality of S is s, i.e. $|S| = s$. Now let Δ_S as the elements of Δ in S. Then, using the above statement, one can show that $\|\Delta_S\|_1 \geq \|\Delta_{S^C}\|_1$. Note,

$$\begin{aligned}
\|\hat{\alpha}\|_1 &= \|\alpha^* + \Delta\|_1 \\
&= \|\alpha^* + \Delta_S\|_1 + \|\Delta_{S^C}\|_1 \\
&\geq \|\alpha^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^C}\|_1 \\
&\geq \|\hat{\alpha}\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^C}\|_1 ,
\end{aligned}$$

and $\|\Delta_S\|_1 \geq \|\Delta_{S^C}\|_1$ follows immediately. Furthermore

$$\left\|\hat{\Delta}\right\|_2 \geq \left\|\hat{\Delta}_S\right\|_2 \geq \left\|\hat{\Delta}_S\right\|_1 /\sqrt{s} \geq \frac{\left\|\hat{\Delta}\right\|_1}{2\sqrt{s}}_1 ,$$

where the last line uses the previous result.

Additionally, note that

$$\begin{aligned}
\Delta^T D^+ \Sigma D^+ \Delta &\leq \|\Delta\|_1 \|D^+ \Sigma D^+ \Delta\|_\infty \\
&\leq 2L\|\Delta\|_1,
\end{aligned}$$

where the first inequality follows by Holder's inequality and the second follows from Lemma 2 and the fact that both $\hat{\gamma}$ and γ^* are in the feasibility set for $L = \sqrt{2\|D^+\|_1 \log\left(p_1 p_2/\epsilon\right)/n}$. We further bound the right hand side of the inequality by using the previous result, which gives

$$\Delta^T D^+ \Sigma D^+ \Delta \leq 4L\sqrt{s}\|\Delta\|_2.$$

Next, we bound $\|\Delta\|_2$ by combining the previous results. Now, by assumption of the Theorem, we have that

$$\begin{aligned}
a \|\Delta\|_2^2 &\leq \Delta^T D^{+T} \Sigma D^+ \Delta \\
&\leq 4L\sqrt{s}\|\Delta\|_2.
\end{aligned}$$

Dividing both sides by $\|\Delta\|_2$, we obtain that

$$\|D\hat{\gamma} - D\gamma^*\|_2 \leq \frac{C_p}{a}\sqrt{\frac{s \log p_1 p_2/\epsilon}{n}}.$$

Finally, we note that

$$\begin{aligned}\|D(\hat{\gamma} - \gamma^*)\|_2^2 &= \|D(m1 + \hat{\gamma} - \gamma^*)\|_2^2 \\ &\geq C_D \|m1 + \hat{\gamma} - \gamma^*\|_2^2 \\ &\quad + \frac{1}{p_1 p_2}\left(p_1 p_2 m + \sum_j \tilde{\gamma}_j - \sum_j \gamma_j^*\right)^2,\end{aligned}$$

where D is the smallest singular value of D that is positive. This follows from the fact that D has only one zero right singular value, whose eigenvector is given by a vector of ones multiplied by $1/\sqrt{p_1 p_2}$. Letting $m = (p_1 p_2)^{-1}\left(\sum_j \gamma_j^* - \sum_j \tilde{\gamma}_j\right)$ concludes the proof.

□

C. Proof of Lemma 1

Proof. Let

$$h(g) = \mathbb{E}_{x \sim F + x_0}\left[\left(f(x) - f(x_0) - \text{vec}(g)^T \text{vec}(x_0 - x)\right)^2\right].$$

Note that $h(g)$ is quadratic and convex in g. Taking the derivative with respect to $\text{vec}(g)$, and setting it to zero we obtain

$$\mathbb{E}_{x \sim F + x_0}\left[-2\,\text{vec}(x_0 - x)\left(f(x) - f(x_0) - \text{vec}(x_0 - x)^T \text{vec}(g^*)\right)\right] = 0,$$

where g^* is the minimizer. After reorganizing the terms and setting $z = x - x_0$, we get

$$\mathbb{E}_{z \sim F}\left[\text{vec}(z)\left(f(x_0 + z) - f(x_0)\right)\right] = \mathbb{E}_{z \sim F}\left[\text{vec}(z)\,\text{vec}(z)^T \text{vec}(g^*)\right] = \Sigma\,\text{vec}(g^*),$$

where we use that $\Sigma = \text{Cov}(\text{vec}(z))$ in the last equation. The result follows trivially. □

D. Equivalency of LEG-TV with Empirical LEG if $L = 0$

Lemma 2. *For the LEG-TV estimate with $L = 0$, if the one vector is an eigenvector of Σ, i.e. $\Sigma 1_{p_1 p_2} = \lambda 1_{p_1 p_2}$, then the solution is equal to the empirical LEG estimate up to a location shift. That is, $\tilde{\gamma} = \hat{\gamma} + a 1_{p_1 p_2}$, for some $a \in \mathbb{R}$.*

Before the proof, we note that the eigenvector condition on Σ can satisfied either with independent noise or our suggested scheme in Section 4.2.

Proof. Note that, if $L = 0$, then we have that

$$D^{+T}\left(\frac{1}{n}\sum_{i=1}^n \text{vec}\left(\tilde{y}_i z_i\right)\right) = D^{+T}\Sigma g.$$

As the only right singular vector of D^{+T} with zero singular value is the one vector, the above statement is true iff

$$\frac{1}{n}\sum_{i=1}^n \text{vec}\left(\tilde{y}_i z_i\right) = \Sigma g + c 1_{p_1 p_2},$$

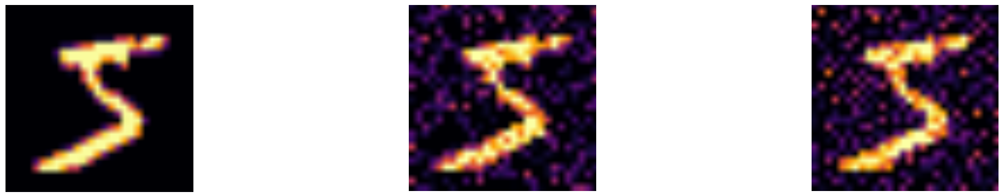
for some $c \in \mathbb{R}$. Solving for g, we obtain,

$$\begin{aligned} g =& \Sigma^{-1}\left(\frac{1}{n}\sum_{i=1}^n \text{vec}\left(\tilde{y}_i z_i\right) - c 1_{p_1 p_2}\right) \\ =& \Sigma^{-1}\frac{1}{n}\sum_{i=1}^n \text{vec}\left(\tilde{y}_i z_i\right) - c\Sigma^{-1} 1_{p_1 p_2} \\ =& \hat{\gamma} - \frac{c}{\lambda} 1_{p_1 p_2},\end{aligned}$$

where we use the fact that the one vector is an eigenvector of Σ^{-1} with eigenvalue λ^{-1}. Setting $a = -\frac{c}{\lambda}$ concludes the proof.

$\square$

E. Examples with different perturbation scheme



(a) Origin (x_0) (b) Independent perturbations (c) New perturbation scheme

Figure 1: Demonstration of the new perturbation scheme on an example from the MNIST dataset. Noise sample of the new scheme have a checkerboard pattern and the perturbation is uniformly distributed across the image.

F. Examples of explanations on MNIST

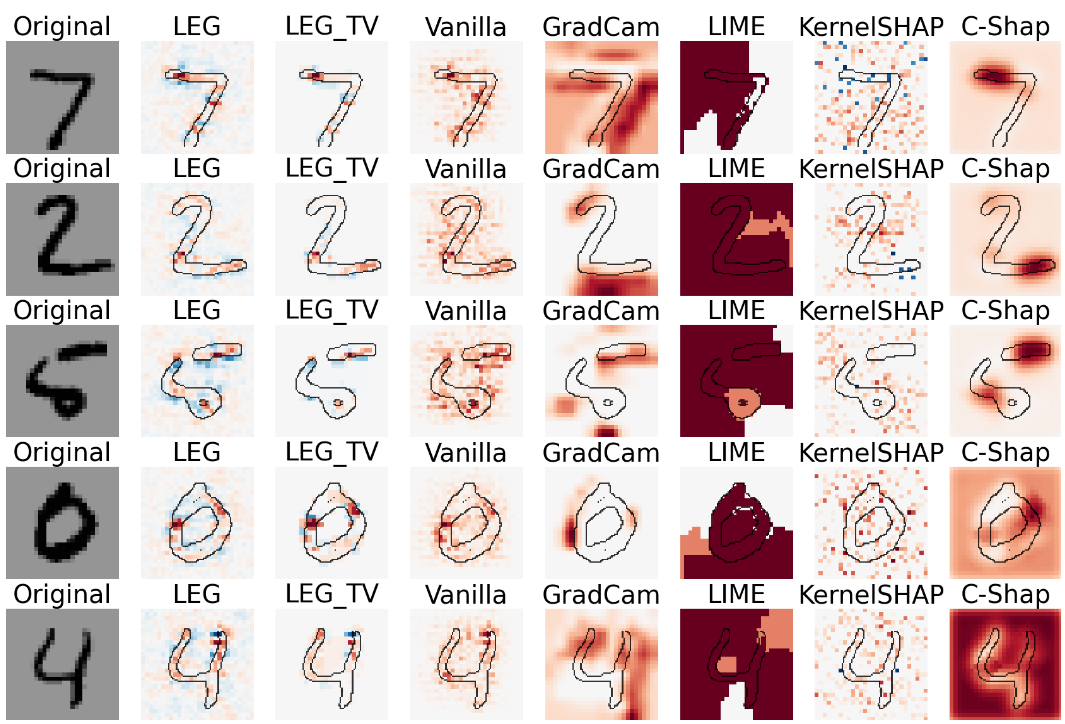


Figure 2: Examples of LEG, LEG-TV[1], Vanilla Gradient[8], GradCam[7], LIME[6], KernelSHAP[5] and C-Shapely[1] explanations for LeNet-5 on MNIST dataset[4]. Among them, Vanilla Gradient, GradCam are model-specific while the others are model-agnostic. Red pixel represents positive saliency while blue pixel represents negative saliency. LEG, LEG-TV have similar pattern as the model-specific method Vanilla Gradient. KernelSHAP takes each pixel as single feature in this case.

---

[1] Implementation details of LEG-TV: sample size is 20k, noise level equals 0.01 and $L = 0.2$, range of pixel intensity is reduced by 20% to satisfy the normality assumption of the perturbations.

G. Sensitivity Plots on MNIST dataset

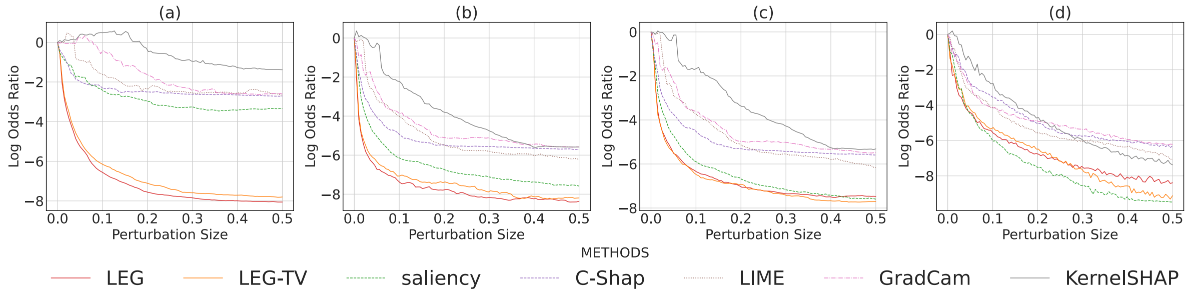


Figure 3: Sensitivity results of LEG, LEG-TV, Vanilla Gradient(saliency), C-Shapley, LIME, GradCam and KernelSHAP with different masking schemes on first 100 test images of MNIST dataset. (a),(b),(c),(d) stand for Distance-100, Distance-255, Mean vector and Noise masking respectively. KernelSHAP performs worse with high-dimensional input space. LEG and LEG-TV outperform within three out of four schemes and also achieve excellent performance on Noise Masking.

H. Sanity check of LEG-TV on VGG-19 model

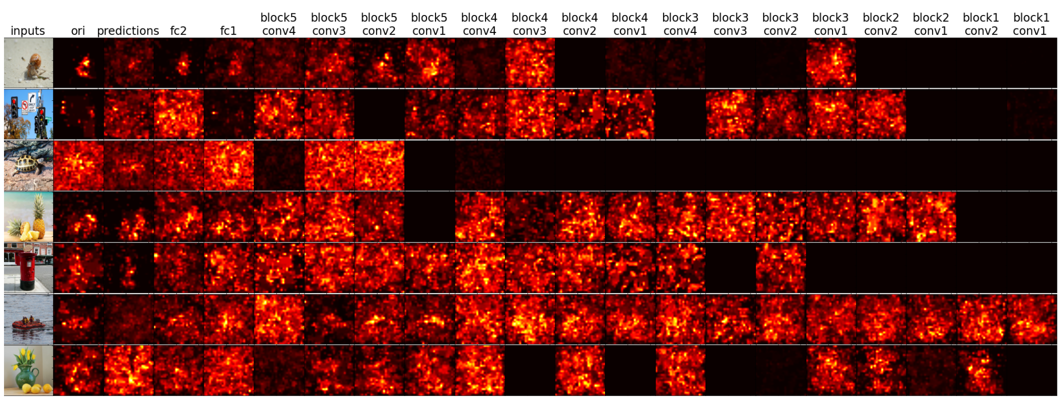


Figure 4: LEG-TV estimates with whole path of cascading randomization on VGG-19.

I. Sanity check of LEG on VGG-19 model

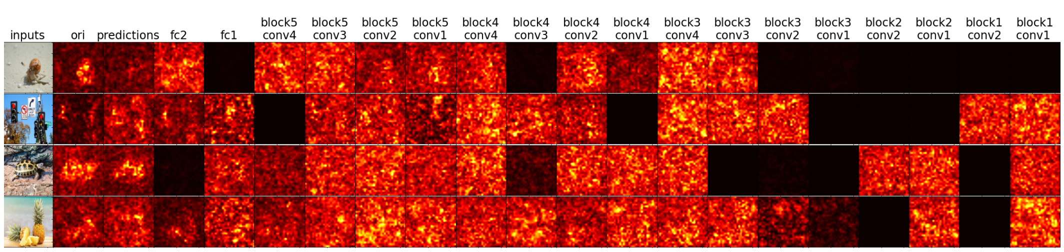


Figure 5: LEG estimates with whole path of cascading randomization on VGG-19. The corresponding estimates after cascading randomization are either noisy or nearly zero after two or three perturbations and show that LEG without regularization is sensitive to the underlying model as well.

J. More Examples on Sensitivity Analysis

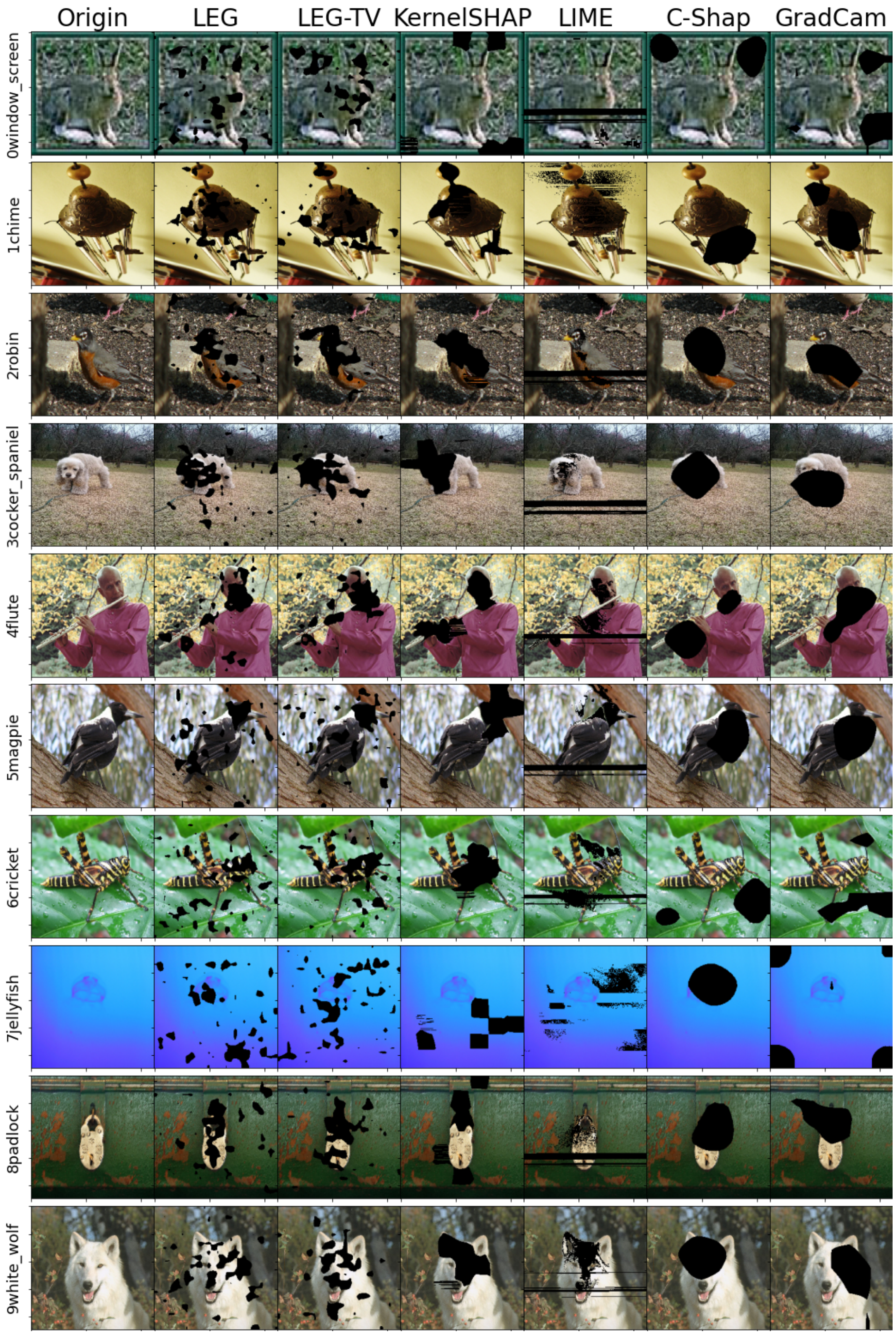


Figure 6: More examples of sensitivity analysis on ImageNet shown by masking 10% of images based on different saliency methods discussed in Section 6.2

References

[1] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. In *International Conference on Learning Representations*, 2019. 5

[2] Jianqing Fan. Features of big data and sparsest solution in high confidence set. *Past, present, and future of statistical science*, pages 507–523, 2013. 2

[3] Brian R Gaines, Juhyun Kim, and Hua Zhou. Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871, 2018. 1

[4] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 5

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017. 5

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. KDD*, pages 1135–1144. ACM, 2016. 5

[7] Ramprasaath R Selvaraju, Cogswell, and et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 5

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 5

[9] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018. 2