Learning Deep Classifiers Consistent with Fine-Grained Novelty Detection Supplementary Material

Jiacheng Cheng Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{jicheng,nvasconcelos}@ucsd.edu

A. Toy Example

Recall that Gaussians have densities of the form

$$\mathcal{G}(\mathbf{v};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu})}.$$

As shown below, $\mathcal{G}(\mathbf{v};\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be expressed in the canonical form for exponential families with canonical parameter $\mathbf{w}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and cumulant function $\psi(\mathbf{w})=\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}$.

$$\begin{aligned}\mathcal{G}(\mathbf{v};\boldsymbol{\mu},\boldsymbol{\Sigma}) =& \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} e^{\langle \mathbf{v},\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\rangle - \frac{1}{2}\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \\ =& \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}} e^{\langle \mathbf{v},\mathbf{w}\rangle - \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w}} \\ =& q(\mathbf{v}) e^{\langle \mathbf{v},\mathbf{w}\rangle - \psi(\mathbf{w})}.\end{aligned}$$

Consider two Gaussians:

- $\mathcal{G}_1$ with $\boldsymbol{\mu}_1 = \begin{bmatrix}1.5\\1.5\end{bmatrix}$ and $\boldsymbol{\Sigma}_1 = \begin{bmatrix}1 & .5\\ .5 & 1\end{bmatrix}$,
- $\mathcal{G}_2$ with $\boldsymbol{\mu}_2 = \begin{bmatrix}1\\1\end{bmatrix}$ and $\boldsymbol{\Sigma}_2 = \begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$.

Contours of $\mathcal{G}_1$ and $\mathcal{G}_2$ are plotted in Figure A.1. Note that $\mathcal{G}_1$ and $\mathcal{G}_2$ are two different distributions sharing the same sufficient statistic $\mathbf{v}$ and canonical parameter $\mathbf{w}_1 = \mathbf{w}_2 = \begin{bmatrix}1\\1\end{bmatrix}$. Our argument about the unidentifiability of the class-conditionals in Section 3.2 can be easily verified.

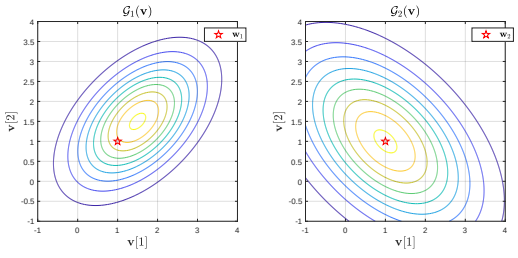


Figure A.1. Contour plots of $\mathcal{G}_1(\mathbf{v})$ and $\mathcal{G}_2(\mathbf{v})$.

B. Proof of Lemma 1

Proof. From (10), it follows that

$$\nabla\psi(\mathbf{w}_y) = \boldsymbol{\mu}_y \qquad \nabla\phi(\boldsymbol{\mu}_y) = \mathbf{w}_y.$$

Hence, (19) holds if and only if

$$\nabla\psi(\mathbf{w}_y) = \boldsymbol{\Sigma}\mathbf{w}_y \qquad \nabla\phi(\boldsymbol{\mu}_y) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y,$$

from which

$$\psi(\mathbf{w}_y) = \frac{1}{2}\mathbf{w}_y^\top\boldsymbol{\Sigma}\mathbf{w}_y + \psi_0 \qquad \phi(\boldsymbol{\mu}_y) = \frac{1}{2}\boldsymbol{\mu}_y^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y + \phi_0$$

for some constants ψ_0, ϕ_0. Using (12), it follows that

$$\begin{aligned}d_\phi(\mathbf{v},\boldsymbol{\mu}_y) &= \frac{1}{2}\mathbf{v}\boldsymbol{\Sigma}^{-1}\mathbf{v} - \frac{1}{2}\boldsymbol{\mu}_y\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \langle\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y, \mathbf{v}-\boldsymbol{\mu}_y\rangle \\ &= \frac{1}{2}\left(\mathbf{v}\boldsymbol{\Sigma}^{-1}\mathbf{v} + \boldsymbol{\mu}_y\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y\right) - \mathbf{v}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y \\ &= \frac{1}{2}(\mathbf{v}-\boldsymbol{\mu}_y)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu}_y)\end{aligned}$$

where $\mathbf{v}(\mathbf{x})$ is shorted for $\mathbf{v}$ and the third equality follows from the symmetry of $\boldsymbol{\Sigma}$. Finally, from (14), (15) holds. □

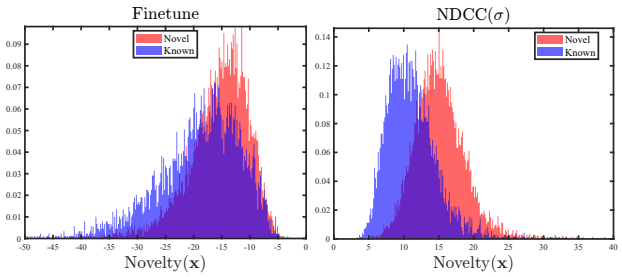
C. Experiments

C.1. Known/novel and Train/test Splits

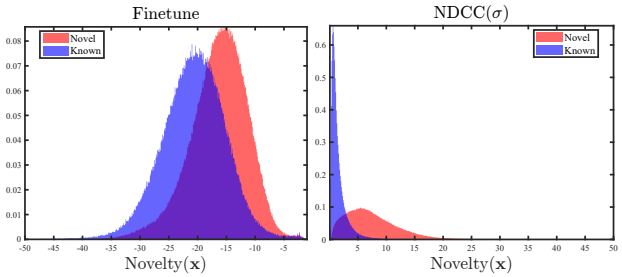
The split protocol in [S1] was followed. For Stanford Dogs, the first 60 breeds were considered as known classes and the remaining 60 breeds were considered as novel classes. For FounderType-200, we picked the first 100 fonts as known classes and the remaining 100 fonts as unseen classes. For CUB-200-2010/Caltech-256, the 200/256 categories were sorted alphabetically and the first 100/128 categories were used as known classes. For all datasets, images from known classes were evenly split between training and test. In other words, half images from known classes were available for training, while the other half images from known classes and all images from novel classes were reserved for test.

	Dogs	FounderType	CUB-200	Caltech-256
input size	(224, 224, 3)	(224, 224, 3)	(336, 336, 3)	(224, 224, 3)
r	16	32	8	16
$\text{lr}(\mathbf{v})$	1e-3	1e-2	1e-2	1e-3
$\text{lr}\,(\mathbf{w}_y, b_y)$	1e-1	1e-1	1e-1	1e-1
$\text{lr}\,(\sigma)$	1e-1	1e-1	1e-1	1e-1
$\text{lr}\,(\delta^{(j)})$	1e-3	1e-3	1e-3	1e-3

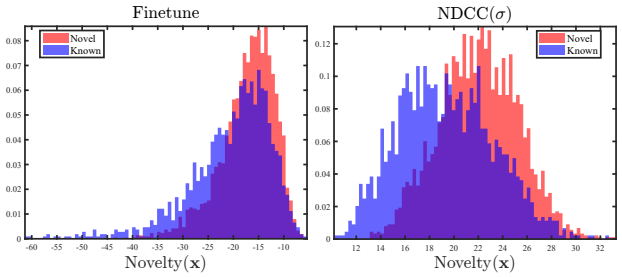
Table C.1. Hyperparameters.



(a) Stanford Dogs



(b) FounderType-200



(c) CUB-200-2010

Figure C.1. Histograms of novelty score for known and novel examples. Evaluations are made with AlexNet as backbone.

C.2. Implementation Details

Denote the initial learning rate for parameters of $\mathbf{v}(\cdot)$, $\{(\mathbf{w}_k, b_k)\}_{k=1}^{C}$, σ, and $\{\delta^{(j)}\}_{j=1}^{d}$ by $\text{lr}(\mathbf{v})$, $\text{lr}\,(\mathbf{w}_y, b_y)$, $\text{lr}\,(\sigma)$, and $\text{lr}\,(\delta^{(j)})$. The learning rate is divided by a factor of 0.1 when the overall loss has stopped decreasing. The learning rate and other hyperparameters we used for different datasets are summarized in Table C.1.

While scanning the literature, we found the results for CUB-200-2010 reported by [S1] to be unrealistically high, given our previous experience with this dataset. Hence, we reevaluated the performance of baseline methods on all datasets using the codes released by the authors. While the results we obtained were similar to those of [S1] for Stanford Dogs, FounderType-200, and Caltech-256, this was not the case for CUB-200-2010, where they were significantly lower. Therefore, we reported the results for CUB-200-2010 produced by our experiments in Table 2.

C.3. Qualitative results

Figure C.1 compares the histograms of the novelty score produced by NDCC(σ) and "Finetune" for known and novel examples in the test set. AlexNet is used as backbone. It is shown that the overlap between scores of novel and known examples is significantly smaller for NDCC(σ) than "Finetune".

References

[S1] Pramuditha Perera and Vishal M Patel. Deep transfer learning for multiple class novelty detection. In *CVPR*, 2019. 1, 2