Supplementary Material for Conditional Noise-Contrastive Estimation of Unnormalised Models

Ciwan Ceylan
UMIC, RWTH Aachen University*
Aachen, Germany
ceylan@vision.rwth-aachen.de

Michael Gutmann
School of Informatics, University of Edinburgh
Edinburgh, United Kingdom
michael.gutmann@ed.ac.uk

A Proof of nonparametric estimation theorem

We here prove the consistency theorem for nonparametric estimation. Additionally, an extension to the theorem where the condition $\mathbb{X} = \mathbb{Y}$ is relaxed to $\mathbb{X} \subseteq \mathbb{Y}$ is presented and proved. For this extended theorem, the following definition is required:

$$p_d^{ext}(\mathbf{u}) = \begin{cases} p_d(\mathbf{u}) & \text{if } \mathbf{u} \in \mathbb{X} \\ 0 & \text{if } \mathbf{u} \in \mathbb{Y} \setminus \mathbb{X}. \end{cases} \tag{1}$$

In order to simplify the notation for the proof, the following definitions are introduced,

$$r(\mathbf{u}_1, \mathbf{u}_2) = \frac{p_c(\mathbf{u}_2|\mathbf{u}_1)}{p_c(\mathbf{u}_1|\mathbf{u}_2)} = \frac{1}{r(\mathbf{u}_2, \mathbf{u}_1)}, \tag{2}$$

and

$$\Omega = \{(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{X} \times \mathbb{X} \mid p_d(\mathbf{u}_1) > 0 \wedge p_c(\mathbf{u}_1|\mathbf{u}_2) > 0\}. \tag{3}$$

Furthermore, the following Taylor expansion will be used in the the proof,

$$\begin{aligned} \log(1 + \exp(-(G + \varepsilon q))) = &\log(1 + \exp(-G)) \\ &- \varepsilon q \frac{\exp(-G)}{1 + \exp(-G)} \\ &+ \frac{\varepsilon^2 q^2}{2} \frac{\exp(-G)}{(1 + \exp(-G))^2} \\ &+ \mathcal{O}(\varepsilon^3). \end{aligned} \tag{4}$$

Using these new definitions, the extended theorem reads:

Theorem (**Nonparametric estimation ext.**). *Let $G : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ be a function of the form*

$$G(\mathbf{u}_1, \mathbf{u}_2) = f(\mathbf{u}_1) - f(\mathbf{u}_2) + \log r(\mathbf{u}_1, \mathbf{u}_2), \tag{5}$$

where f is a function from $\mathbb{U}$ to $\mathbb{R}$. Under the assumption $\mathbb{X} \subseteq \mathbb{Y}$, $\tilde{\mathcal{J}}$ attains a unique minimum at

$$G^*(\mathbf{u}_1, \mathbf{u}_2) = \log \frac{p_d^{ext}(\mathbf{u}_1) p_c(\mathbf{u}_2|\mathbf{u}_1)}{p_d^{ext}(\mathbf{u}_2) p_c(\mathbf{u}_1|\mathbf{u}_2)} \tag{6}$$

for $(\mathbf{u}_1, \mathbf{u}_2) \in \Omega$.

*Affiliated with KTH Royal Institute of Technology and University of Edinburgh during project timespan.

First, the proof of the theorem of the main article is presented, followed by the extra steps required to prove the extended theorem.

Proof of ***Nonparametric estimation***. The proof is divided into two parts. First, G^* is proved to be a critical point of $\tilde{\mathcal{J}}$ by showing that the linear term of the Taylor expansion for $\tilde{\mathcal{J}}$ with respect to G is zero for G^*. In the second part, we prove that G^* is a minimum and the only extremum by showing that the quadratic part of the Taylor expansion is strictly positive on the set Ω.

The functional $\tilde{\mathcal{J}}[G]$ is expressed as the integral

$$\tilde{\mathcal{J}}[G] = \mathbb{E}_{\mathbf{xy}} \log\left(1 + \exp(-G(\mathbf{x}, \mathbf{y}))\right) \tag{7}$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \log\left(1 + \exp(-G(\mathbf{x}, \mathbf{y}))\right) p_d(\mathbf{x}) p_c(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{8}$$

Inserting Equation (5), we obtain the functional

$$\begin{aligned}\tilde{\mathcal{J}}_f[f] &= \mathbb{E}_{\mathbf{xy}} \log\left(1 + \exp(f(\mathbf{y}) - f(\mathbf{x}) + \log r(\mathbf{x}, \mathbf{y}))\right) \\ &= \int_{\mathbb{X}\times\mathbb{Y}} \log\left(1 + \exp(f(\mathbf{y}) - f(\mathbf{x}) + \log r(\mathbf{x}, \mathbf{y}))\right) p_d(\mathbf{x}) p_c(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.\end{aligned} \tag{9}$$

Now consider an arbitrary perturbation $\psi : \mathbb{U} \to \mathbb{R}$ of f

$$\tilde{\mathcal{J}}_f[f + \varepsilon\psi] = \mathbb{E}_{\mathbf{xy}} \log\left(1 + \exp\left[f(\mathbf{y}) + \varepsilon\psi(\mathbf{y}) - f(\mathbf{x}) - \varepsilon\psi(\mathbf{x}) + \log r(\mathbf{x}, \mathbf{y})\right]\right) \tag{10}$$

$$= \int_{\mathbb{X}\times\mathbb{Y}} \log\left(1 + \exp\left[-(G(\mathbf{x}, \mathbf{y}) + \varepsilon(\psi(\mathbf{x}) - \psi(\mathbf{y})))\right]\right) p_d(\mathbf{x}) p_c(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{11}$$

The perturbation of $\tilde{\mathcal{J}}_f[f]$ corresponds to the following perturbation of $\tilde{\mathcal{J}}[G]$,

$$\tilde{\mathcal{J}}[G + \varepsilon(\psi(\mathbf{x}) - \psi(\mathbf{y}))] = \mathbb{E}_{\mathbf{xy}} \log\left(1 + \exp[-(G(\mathbf{x}, \mathbf{y}) + \varepsilon(\psi(\mathbf{x}) - \psi(\mathbf{y})))]\right). \tag{12}$$

Using the Taylor expansion from Equation (4) gives

$$\begin{aligned}\tilde{\mathcal{J}}[G + \varepsilon(\psi(\mathbf{x}) - \psi(\mathbf{y}))] = \mathbb{E}_{\mathbf{xy}} &\log(1 + \exp(-G(\mathbf{x}, \mathbf{y}))) \\ &- \varepsilon \mathbb{E}_{\mathbf{xy}}(\psi(\mathbf{x}) - \psi(\mathbf{y})) \frac{\exp(-G(\mathbf{x}, \mathbf{y}))}{1 + \exp(-G(\mathbf{x}, \mathbf{y}))} \\ &+ \frac{\varepsilon^2}{2} \mathbb{E}_{\mathbf{xy}}(\psi(\mathbf{x}) - \psi(\mathbf{y}))^2 \frac{\exp(-G(\mathbf{x}, \mathbf{y}))}{(1 + \exp(-G(\mathbf{x}, \mathbf{y})))^2} \\ &+ \mathcal{O}(\varepsilon^3).\end{aligned} \tag{13}$$

Equating the 1st order term with 0 lets us find a necessary condition for the optimal G,

$$0 = \mathbb{E}_{\mathbf{xy}}(\psi(\mathbf{x}) - \psi(\mathbf{y})) \frac{\exp(-G(\mathbf{x}, \mathbf{y}))}{1 + \exp(-G(\mathbf{x}, \mathbf{y}))} \tag{14}$$

$$\begin{aligned}&= \int_{\mathbb{X}\times\mathbb{Y}} \psi(\mathbf{x}) \frac{\exp(-G(\mathbf{x}, \mathbf{y}))}{1 + \exp(-G(\mathbf{x}, \mathbf{y}))} p_d(\mathbf{x}) p_c(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &\quad - \int_{\mathbb{X}\times\mathbb{Y}} \psi(\mathbf{y}) \frac{\exp(-G(\mathbf{x}, \mathbf{y}))}{1 + \exp(-G(\mathbf{x}, \mathbf{y}))} p_d(\mathbf{x}) p_c(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\end{aligned} \tag{15}$$

We now make a change of variables. For the first term in Equation (15) we write $\mathbf{u}$ for $\mathbf{x}$ and $\mathbf{v}$ for $\mathbf{y}$ while for the second term we use the transform

$$T_2 : \begin{pmatrix}\mathbf{u} \\ \mathbf{v}\end{pmatrix} = \begin{pmatrix}\mathbf{0} & \mathbb{1} \\ \mathbb{1} & \mathbf{0}\end{pmatrix} \begin{pmatrix}\mathbf{x} \\ \mathbf{y}\end{pmatrix} \tag{16}$$

$$\det \begin{pmatrix}\mathbf{0} & \mathbb{1} \\ \mathbb{1} & \mathbf{0}\end{pmatrix} = -1 \tag{17}$$

$$T_2(\mathbb{X} \times \mathbb{Y}) = \mathbb{Y} \times \mathbb{X}. \tag{18}$$

In the resulting equation, the integrals for the two terms are taken over different domains,

$$\begin{aligned} 0 = &\int_{\mathbb{X}\times\mathbb{Y}} \psi(\mathbf{u}) \frac{\exp(-G(\mathbf{u},\mathbf{v}))}{1+\exp(-G(\mathbf{u},\mathbf{v}))} p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &- \int_{\mathbb{Y}\times\mathbb{X}} \psi(\mathbf{u}) \frac{\exp(-G(\mathbf{v},\mathbf{u}))}{1+\exp(-G(\mathbf{v},\mathbf{u}))} p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}. \end{aligned} \tag{19}$$

For the first theorem we assume $\mathbb{Y} = \mathbb{X}$ so that

$$\begin{aligned} 0 = &\int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u}) \frac{\exp(-G(\mathbf{u},\mathbf{v}))}{1+\exp(-G(\mathbf{u},\mathbf{v}))} p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &- \int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u}) \frac{\exp(-G(\mathbf{v},\mathbf{u}))}{1+\exp(-G(\mathbf{v},\mathbf{u}))} p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \end{aligned} \tag{20}$$

$$\begin{aligned} = &\int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u}) \left(\frac{\exp(-G(\mathbf{u},\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{1+\exp(-G(\mathbf{u},\mathbf{v}))} \right. \\ &\left. - \frac{\exp(-G(\mathbf{v},\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{1+\exp(-G(\mathbf{v},\mathbf{u}))} \right) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \end{aligned} \tag{21}$$

Since Equation (21) should hold for any ψ on $\mathbb{X}\times\mathbb{X}$, the factor in the parenthesis must equal 0. The factor can be expanded by inserting the assumed form of G, see Equation (5),

$$\frac{\exp(-G(\mathbf{u},\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{1+\exp(-G(\mathbf{u},\mathbf{v}))} = \frac{\exp(-G(\mathbf{v},\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{1+\exp(-G(\mathbf{v},\mathbf{u}))} \tag{22}$$

$$\frac{p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{\exp(G(\mathbf{u},\mathbf{v}))+1} = \frac{p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{\exp(G(\mathbf{v},\mathbf{u}))+1} \tag{23}$$

$$\frac{p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{\exp(f(\mathbf{u})-f(\mathbf{v}))r(\mathbf{u},\mathbf{v})+1} = \frac{p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{\exp(f(\mathbf{v})-f(\mathbf{u}))r(\mathbf{v},\mathbf{u})+1} \tag{24}$$

$$\frac{\exp(f(\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{\exp(f(\mathbf{u}))r(\mathbf{u},\mathbf{v})+\exp(f(\mathbf{v}))} = \frac{\exp(f(\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{\exp(f(\mathbf{v}))r(\mathbf{v},\mathbf{u})+\exp(f(\mathbf{u}))} \tag{25}$$

Using $r(\mathbf{v},\mathbf{u}) = 1/r(\mathbf{u},\mathbf{v})$ from Equation (2), a factor can be taken out of the denominator of the r.h.s,

$$\frac{\exp(f(\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u})}{\exp(f(\mathbf{u}))r(\mathbf{u},\mathbf{v})+\exp(f(\mathbf{v}))} = \frac{1}{r(\mathbf{v},\mathbf{u})} \frac{\exp(f(\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v})}{\exp(f(\mathbf{v}))+\exp(f(\mathbf{u}))r(\mathbf{u},\mathbf{v})} \tag{26}$$

$$\exp(f(\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u}) = \frac{1}{r(\mathbf{v},\mathbf{u})} \exp(f(\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v}) \tag{27}$$

$$\exp(f(\mathbf{v})) p_d(\mathbf{u}) p_c(\mathbf{v}|\mathbf{u}) = \frac{p_c(\mathbf{v}|\mathbf{u})}{p_c(\mathbf{u}|\mathbf{v})} \exp(f(\mathbf{u})) p_d(\mathbf{v}) p_c(\mathbf{u}|\mathbf{v}). \tag{28}$$

Now consider only the set Ω where both sides in the above equation are not trivially zero,

$$\exp(f(\mathbf{v})) p_d(\mathbf{u}) = \exp(f(\mathbf{u})) p_d(\mathbf{v}) \tag{29}$$

$$\frac{p_d(\mathbf{u})}{\exp(f(\mathbf{u}))} = \frac{p_d(\mathbf{v})}{\exp(f(\mathbf{v}))} = Z \tag{30}$$

$$f^*(\mathbf{u}) = \log p_d(\mathbf{u}) - \log Z \tag{31}$$

$$G^*(\mathbf{u}_1,\mathbf{u}_2) = \log p_d(\mathbf{u}_1) - \log p_d(\mathbf{u}_2) + \log r(\mathbf{u}_1,\mathbf{u}_2). \tag{32}$$

The first part of the proof is now completed as G^* in Equation (32) is a critical point of $\tilde{\mathcal{J}}$.

It is straightforward to show that G^* is minimising $\tilde{\mathcal{J}}$ and is the only extreme point. By considering the second order term of the Taylor expansion in Equation (13),

$$\mathbb{E}_{\mathbf{xy}} (\psi(\mathbf{x}) - \psi(\mathbf{y}))^2 \frac{\exp(-G(\mathbf{x},\mathbf{y}))}{(1+\exp(-G(\mathbf{x},\mathbf{y})))^2}, \tag{33}$$

we observe that it is positive for all non-constant perturbations ψ. Since constant perturbations of f does not change G, it can be concluded that Equation (32) describes a minimum and the only extreme point on the set Ω. ■

*Proof of **Nonparametric estimation ext.***. We can follow the previous proof until Equation (19), just after the change of variables. We now observe the following

$$\begin{aligned}\mathbb{X}\times\mathbb{Y} = (\mathbb{X}\cap\mathbb{Y})\times(\mathbb{X}\cap\mathbb{Y}) \cup (\mathbb{X}\setminus\mathbb{Y})\times(\mathbb{X}\cap\mathbb{Y}) \\ \cup (\mathbb{X}\cap\mathbb{Y})\times(\mathbb{Y}\setminus\mathbb{X}) \cup (\mathbb{X}\setminus\mathbb{Y})\times(\mathbb{Y}\setminus\mathbb{X})\end{aligned} \tag{34}$$

The assumption $\mathbb{X}\subseteq\mathbb{Y}$ implies $(\mathbb{X}\setminus\mathbb{Y})=\emptyset$ and $(\mathbb{X}\cap\mathbb{Y})=\mathbb{X}$. Therefore,

$$\mathbb{X}\times\mathbb{Y} = \Big((\mathbb{X}\cap\mathbb{Y})\times(\mathbb{X}\cap\mathbb{Y})\Big) \cup \Big((\mathbb{X}\cap\mathbb{Y})\times(\mathbb{Y}\setminus\mathbb{X})\Big) \tag{35}$$

$$= \Big(\mathbb{X}\times\mathbb{X}\Big) \cup \Big(\mathbb{X}\times(\mathbb{Y}\setminus\mathbb{X})\Big), \tag{36}$$

and similarly

$$\mathbb{Y}\times\mathbb{X} = \Big(\mathbb{X}\times\mathbb{X}\Big) \cup \Big((\mathbb{Y}\setminus\mathbb{X})\times\mathbb{X}\Big). \tag{37}$$

It is now possible to reevaluate Equation (19),

$$\begin{aligned}0 = &\int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{u},\mathbf{v}))}{1+\exp(-G(\mathbf{u},\mathbf{v}))}p_d(\mathbf{u})p_c(\mathbf{v}|\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &+\int_{\mathbb{X}\times(\mathbb{Y}\setminus\mathbb{X})} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{u},\mathbf{v}))}{1+\exp(-G(\mathbf{u},\mathbf{v}))}p_d(\mathbf{u})p_c(\mathbf{v}|\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &-\int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{v},\mathbf{u}))}{1+\exp(-G(\mathbf{v},\mathbf{u}))}p_d(\mathbf{v})p_c(\mathbf{u}|\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &-\int_{(\mathbb{Y}\setminus\mathbb{X})\times\mathbb{X}} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{v},\mathbf{u}))}{1+\exp(-G(\mathbf{v},\mathbf{u}))}p_d(\mathbf{v})p_c(\mathbf{u}|\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}\end{aligned} \tag{38}$$

$$\begin{aligned}0 = &\int_{\mathbb{X}\times\mathbb{X}} \psi(\mathbf{u})\left(\frac{\exp(-G(\mathbf{u},\mathbf{v}))p_d(\mathbf{u})p_c(\mathbf{v}|\mathbf{u})}{1+\exp(-G(\mathbf{u},\mathbf{v}))}\right. \\ &\left.-\frac{\exp(-G(\mathbf{v},\mathbf{u}))p_d(\mathbf{v})p_c(\mathbf{u}|\mathbf{v})}{1+\exp(-G(\mathbf{v},\mathbf{u}))}\right)\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &+\int_{\mathbb{X}\times(\mathbb{Y}\setminus\mathbb{X})} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{u},\mathbf{v}))}{1+\exp(-G(\mathbf{u},\mathbf{v}))}p_d(\mathbf{u})p_c(\mathbf{v}|\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v} \\ &-\int_{(\mathbb{Y}\setminus\mathbb{X})\times\mathbb{X}} \psi(\mathbf{u})\frac{\exp(-G(\mathbf{v},\mathbf{u}))}{1+\exp(-G(\mathbf{v},\mathbf{u}))}p_d(\mathbf{v})p_c(\mathbf{u}|\mathbf{v})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}\end{aligned} \tag{39}$$

Following the previous proof,

$$G(\mathbf{u}_1,\mathbf{u}_2) = \log p_d(\mathbf{u}_1) - \log p_d(\mathbf{u}_2) + \log r(\mathbf{u}_1,\mathbf{u}_2) \tag{40}$$

will set the first term of Equation (39) to 0. By using the expanded data distribution p_d^{ext} from Equation (1) in place of p_d, we find

$$G^*(\mathbf{u}_1,\mathbf{u}_2) = \log p_d^{ext}(\mathbf{u}_1) - \log p_d^{ext}(\mathbf{u}_2) + \log r(\mathbf{u}_1,\mathbf{u}_2). \tag{41}$$

Since G^* becomes arbitrarily large on $\mathbb{X}\times(\mathbb{Y}\setminus\mathbb{X})$, the second and third terms of Equation (39) are 0. Again, the second order term is positive for all non-constant perturbations ψ on the set Ω. ■

B Proof of connection to score matching

*Proof of **Connection to score matching***. We here assume that $\mathbf{y} = \mathbf{x} + \varepsilon\boldsymbol{\xi}$ where $\boldsymbol{\xi}$ is a vector of uncorrelated random variables of mean zero and variance one that are independent from $\mathbf{x}$ and have a symmetric density.

Since $\boldsymbol{\xi}$ has a symmetric density, p_c is symmetric and cancels in the definition of $G(\mathbf{u}_1,\mathbf{u}_2;\boldsymbol{\theta})$,

$$G(\mathbf{u}_1,\mathbf{u}_2;\boldsymbol{\theta}) = \log\frac{\phi(\mathbf{u}_1;\boldsymbol{\theta})p_c(\mathbf{u}_2|\mathbf{u}_1)}{\phi(\mathbf{u}_2;\boldsymbol{\theta})p_c(\mathbf{u}_1|\mathbf{u}_2)} = \log\phi(\mathbf{u}_1;\boldsymbol{\theta}) - \log\phi(\mathbf{u}_2;\boldsymbol{\theta}). \tag{42}$$

The loss function thus is

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) &= 2\mathbb{E}_{\mathbf{xy}} \log\left[1+\exp(-G(\mathbf{x},\mathbf{y};\boldsymbol{\theta}))\right] &(43)\\ &= 2\mathbb{E}_{\mathbf{xy}} \log\left[1+\exp\left(-\log\phi(\mathbf{x};\boldsymbol{\theta})+\log\phi(\mathbf{y};\boldsymbol{\theta})\right)\right] \\ &= 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}} \log\left[1+\exp\left(-\log\phi(\mathbf{x};\boldsymbol{\theta})+\log\phi(\mathbf{x}+\varepsilon\boldsymbol{\xi};\boldsymbol{\theta})\right)\right] &(44)\end{aligned}$$

Let us denote the log unnormalised model $\log\phi(\cdot;\boldsymbol{\theta})$ by $f_{\boldsymbol{\theta}}(\cdot)$ so that

$$\mathcal{J}(\boldsymbol{\theta}) = 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}} \log\left[1+\exp\left(-f_{\boldsymbol{\theta}}(\mathbf{x})+f_{\boldsymbol{\theta}}(\mathbf{x}+\varepsilon\boldsymbol{\xi})\right)\right] \quad (45)$$

By assumption ε is small so that for any fixed value of $\boldsymbol{\xi}$, we have

$$f_{\boldsymbol{\theta}}(\mathbf{x}+\varepsilon\boldsymbol{\xi}) = f_{\boldsymbol{\theta}}(\mathbf{x}) + \varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{2}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + O(\varepsilon^3) \quad (46)$$

where $\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})$ is the Hessian with elements $\partial_{x_i}\partial_{x_j} f_{\boldsymbol{\theta}}(\mathbf{x})$. We thus obtain

$$\mathcal{J}(\boldsymbol{\theta}) = 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}} \log\left[1+\exp\left(\varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{2}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + O(\varepsilon^3)\right)\right] \quad (47)$$

The function $\log(1+\exp(v))$ has the following Taylor expansion around $v=0$,

$$\log(1+\exp(v)) = \log(2) + \frac{1}{2}v + \frac{1}{8}v^2 + O(v^3), \quad (48)$$

so that

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) = 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}}\left[\log(2) + \frac{1}{2}\varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{4}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + O(\varepsilon^3)\right] + \\ 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}}\left[\frac{1}{8}\left(\varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{2}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + O(\varepsilon^3)\right)^2\right]. \quad (49)\end{aligned}$$

Squaring the term $\left(\varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{2}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + O(\varepsilon^3)\right)^2$ gives $\varepsilon^2(\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi})^2 + O(\varepsilon^3)$ so that

$$\mathcal{J}(\boldsymbol{\theta}) = 2\mathbb{E}_{\mathbf{x}\boldsymbol{\xi}}\left[\log(2) + \frac{1}{2}\varepsilon\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} + \frac{\varepsilon^2}{4}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + \frac{1}{8}\varepsilon^2(\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi})^2\right] + O(\varepsilon^3) \quad (50)$$

By assumption, $\mathbf{x}$ and $\boldsymbol{\xi}$ are independent, and $\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi} = 0$, so that we have

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) = 2\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\boldsymbol{\xi}}\left[\log(2) + \frac{\varepsilon^2}{4}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} + \frac{1}{8}\varepsilon^2(\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi})^2\right] \\ + O(\varepsilon^3) \quad (51)\end{aligned}$$

Furthermore, with $\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}\boldsymbol{\xi}^T$ being equal to the identity matrix $\mathbb{1}$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}^T\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi} &= \mathbb{E}_{\boldsymbol{\xi}}\operatorname{tr}\left[\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\boldsymbol{\xi}\boldsymbol{\xi}^T\right] &(52)\\ &= \operatorname{tr}\left[\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}\boldsymbol{\xi}^T\right] &(53)\\ &= \operatorname{tr}\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x}). &(54)\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}(\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi})^2 &= \mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}^T\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi} &(55)\\ &= \mathbb{E}_{\boldsymbol{\xi}}\operatorname{tr}\left[\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\boldsymbol{\xi}\boldsymbol{\xi}^T\right] &(56)\\ &= \operatorname{tr}\left[\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\mathbb{E}_{\boldsymbol{\xi}}\boldsymbol{\xi}\boldsymbol{\xi}^T\right] &(57)\\ &= \operatorname{tr}\left[\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})^T\right] &(58)\\ &= ||\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})||_2^2. &(59)\end{aligned}$$

With both identities plugged into (51), we can write $\mathcal{J}(\boldsymbol{\theta})$ as

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{\varepsilon^2}{2}\mathbb{E}_{\mathbf{x}}\left[\operatorname{tr}\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x}) + \frac{1}{2}||\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})||_2^2\right] + 2\log(2) + O(\varepsilon^3). \tag{60}$$

Since $\operatorname{tr}\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})$ equals the sum of the second derivatives of $f_{\boldsymbol{\theta}}(\mathbf{x}) = \log\phi(\mathbf{x};\boldsymbol{\theta})$,

$$\operatorname{tr}\mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_i \frac{\partial^2 f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_i^2} \tag{61}$$

the term in the brackets is the loss function that is minimised in score matching, which completes the proof. ■

C Empirical validation on non-negative and discrete data

CNCE was also verified on a heavy-tailed distribution of positive data (log-normal) and a discrete distribution (Bernoulli).

The log-normal distribution is a univariate continuous heavy-tailed distribution that is defined to have its samples normal distributed in the log domain. Consequently, it is only defined on the positive real axis $\mathbb{X} = \mathbb{R}^+$. For this reason, the log-normal distribution is suitable to illustrate the fact that only $\mathbb{X} \subseteq \mathbb{Y}$ is required for CNCE given that the conditional noise distribution p_c defined in the main paper generates noise samples in $\mathbb{Y} = \mathbb{R}$. We used the following unnormalised log-normal model defined over the whole real axis

$$\log\phi(u;\theta,C) = \begin{cases} -\frac{\theta}{2}(\log u)^2 - \log u & \text{if } u > 0 \\ C & \text{if } u \leq 0 \end{cases} \tag{62}$$

where $\theta, C \in \mathbb{R}$. On the positive axis, the model is proportional to a log-normal distribution with mean zero in the log domain and with precision θ. On the negative axis, the model assumes the constant value C. In theory, the optimal value of C would be $-\infty$. Since this can never be reached in practice, we only measured the estimation error for θ as the absolute error between true and estimated parameter.

The Bernoulli model defines a simple probability mass function for a binary random variable taking values on $\mathbb{X} = \{0, 1\}$. In the normalised version, the Bernoulli model only has one free parameter. Here, an unnormalised version with two free parameters $\theta_1, \theta_2 \in \mathbb{R}^+$ is used,

$$\log\phi(u;\theta_1,\theta_2) = \begin{cases} \log\theta_1 & \text{if } u = 0 \\ \log\theta_2 & \text{if } u = 1. \end{cases} \tag{63}$$

The use of two free parameters means that there exist an infinite set of equivalent model parameters which only differs from (θ_1, θ_2) by a scaling factor. Consequently, to measure the error for a parameter estimate of the Bernoulli model, i.e. $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$, we normalised the parameters before computing the estimation error as $\|(\hat{\theta}_1 + \hat{\theta}_2)^{-1}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, where $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)$ denotes the true parameter values (which are related by $\theta_2^* = 1 - \theta_1^*$).

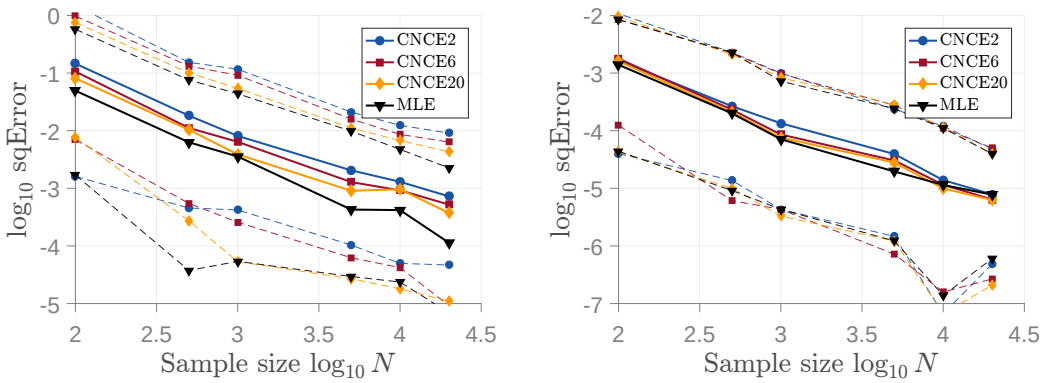
The discrete conditional noise distribution defined by Equation (64) was used for the Bernoulli model. Again ε controls the similarity between data and noise, but with the added restriction $\varepsilon \in [0, 1]$.

$$p_c^{Ber}(y|x;\varepsilon) = \begin{cases} 1-\varepsilon & \text{if } y = x \\ \varepsilon & \text{if } y \neq x, \end{cases} \tag{64}$$

D Supplemental feature visualisations

D.1 Neural network layer sizes

The sizes of the four layers are provided in Table 1. Note that the dimensionality of the data was reduced by four using PCA as part of the gain control between the 2nd and 3rd layers.



(a) Log-normal model.

(b) Bernoulli model.

Figure 1: Empirical validation of consistency for CNCE. The x-axis shows the sample size in the $\log_{10}$ domain, the y-axis the squared estimation error in the $\log_{10}$ domain. The solid lines show the median result across 100 different simulations, and the dashed lines are the 0.1 and 0.9 quantiles. For each of the 100 simulations, a new random set of parameters were used to generate the data. The different coloured and marked lines correspond to different values of κ for CNCE and the black line to the MLE results.

Table 1: Neural network input and output dimensions.

Layer	Input $D^{(L)}$	Output $K^{(L)}$
1	600	600
2	600	200
Intermediate gain control		
3	196	60
4	60	30

D.2 CNCE and NCE 1st layer features comparison

In addition to the quantitative comparisons between CNCE and NCE, which were presented in Section 3 of the main paper, it is desirable to evaluate qualitative differences between the methods. To this end we compared the easily interpretable 1st layer features at different stages of training with the aim to determine qualitative differences between learned features and if learning is faster for one method or the other.

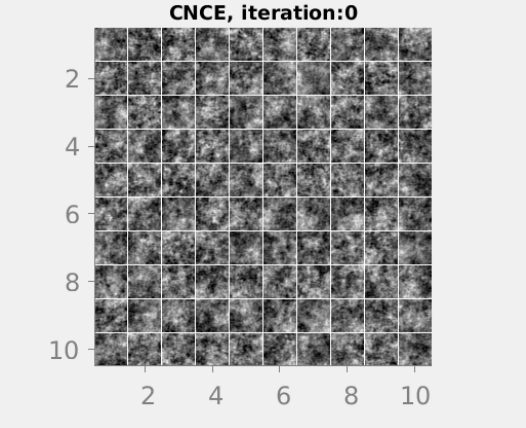
Figure 2 shows the common initialisation and Figures 3 to 13 one hundred 1st layer features at the end of the first eleven meta-iterations. Each meta-iteration consists of ten gradient steps after which new noise samples are generated. The methods seem to learn similar features and for this model, while we do not want to claim superior performance given the qualitative nature of the comparison, CNCE does appear to learn slightly faster than NCE.

D.3 3rd layer features

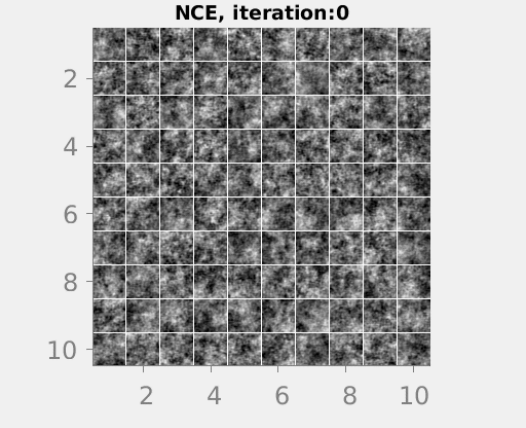
All 60 3rd layer space-orientation receptive fields and maximal response patches are shown in Figures 14 to 17.

D.4 4th layer features

All 30 4th units are visualised in Figures 18 to 47 in the same manner as in the main text.

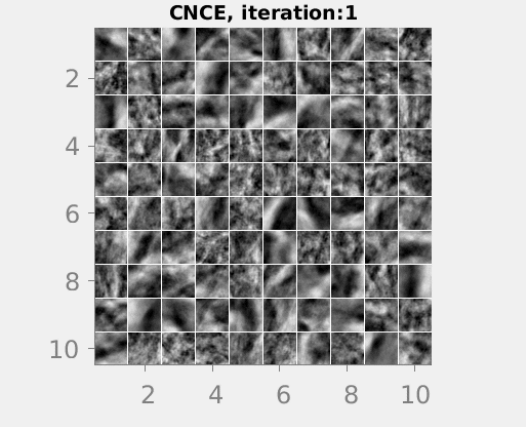


(a) CNCE

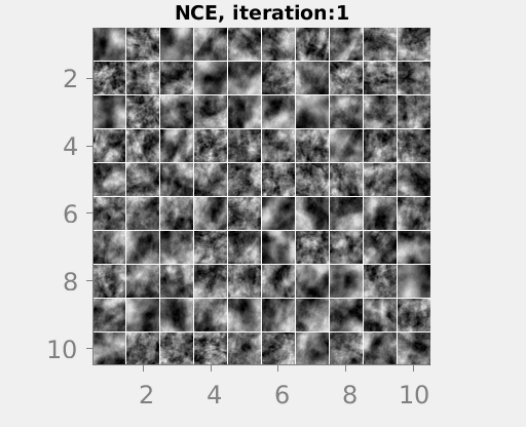


(b) NCE

Figure 2: The common initialisation for the 100 1st later features used for the qualitative comparison between CNCE and NCE.

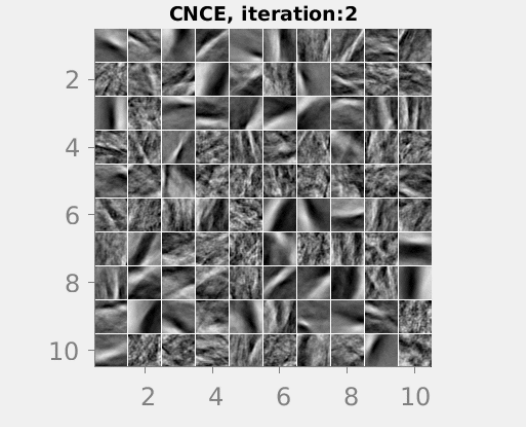


(a) CNCE

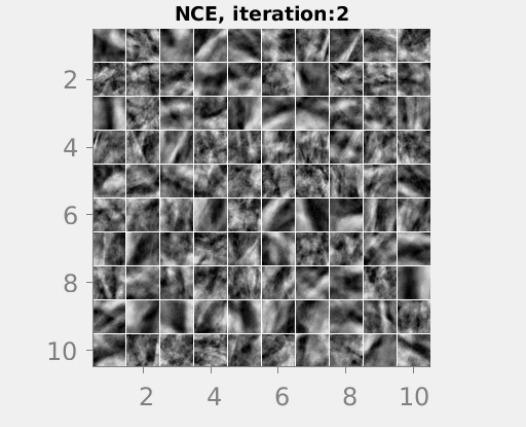


(b) NCE

Figure 3: A sample of the 1st layer features after 1 meta-iteration.

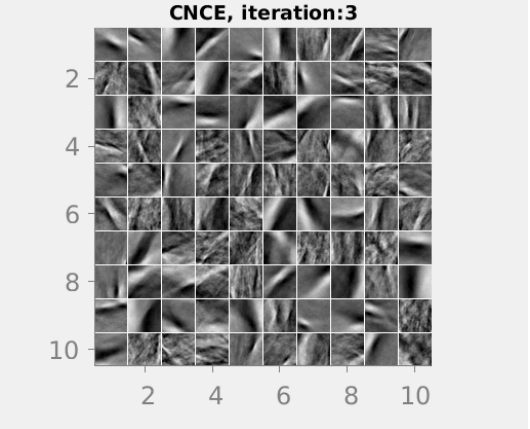


(a) CNCE

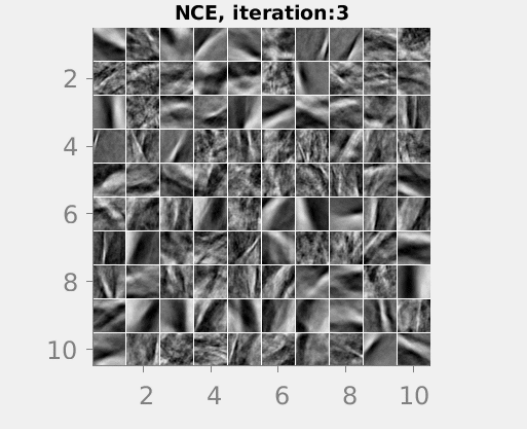


(b) NCE

Figure 4: A sample of the 1st layer features after 2 meta-iterations.

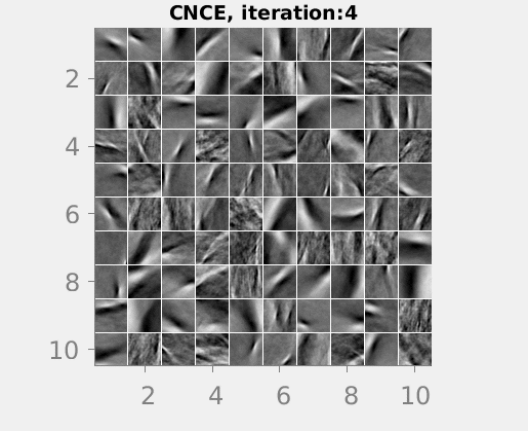


(a) CNCE

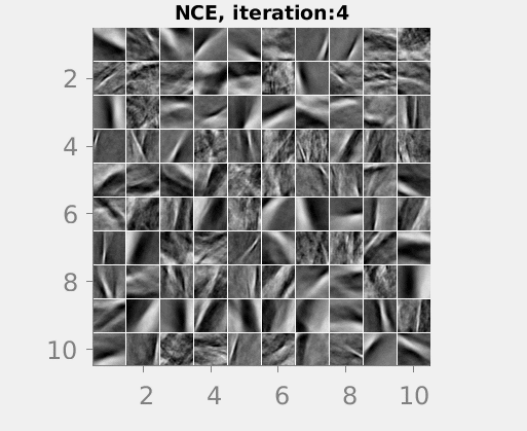


(b) NCE

Figure 5: A sample of the 1st layer features after 3 meta-iterations.

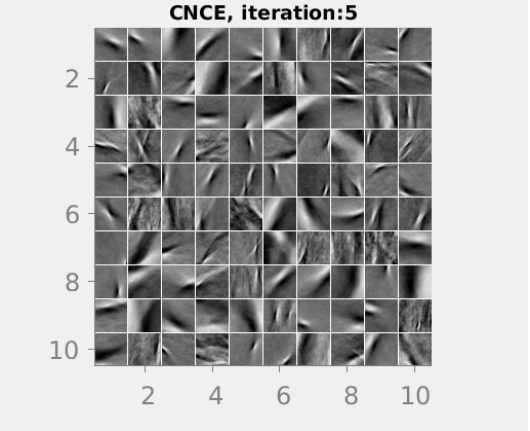


(a) CNCE

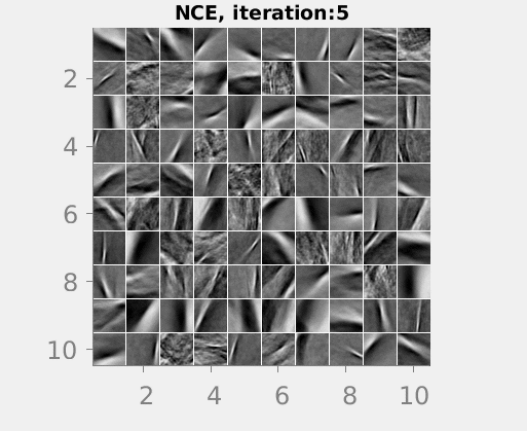


(b) NCE

Figure 6: A sample of the 1st layer features after 4 meta-iterations.

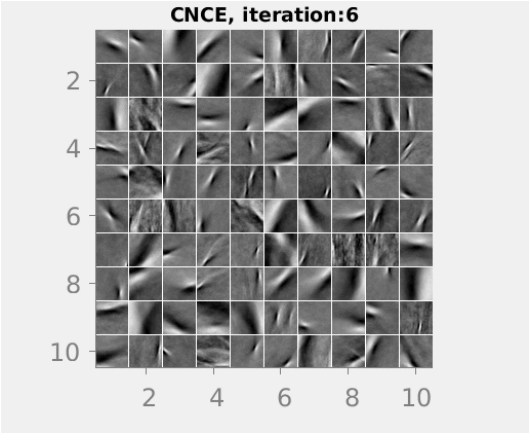


(a) CNCE

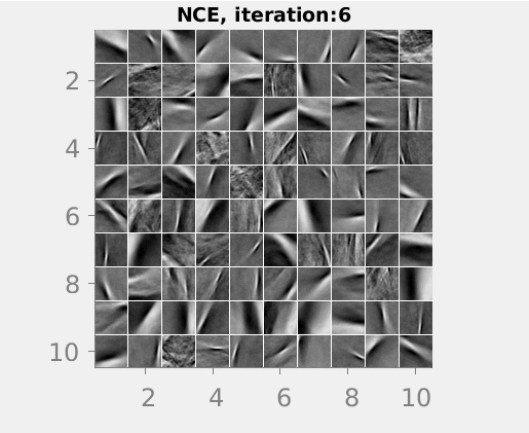


(b) NCE

Figure 7: A sample of the 1st layer features after 5 meta-iterations.

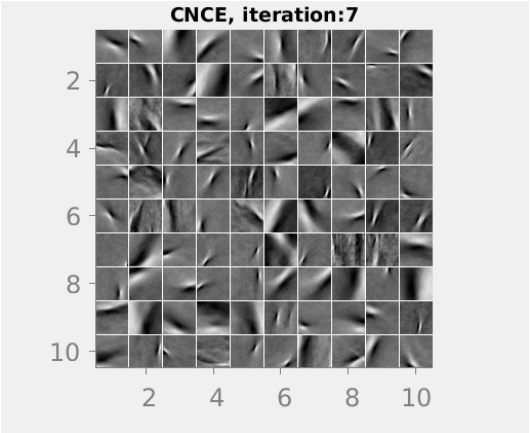


(a) CNCE

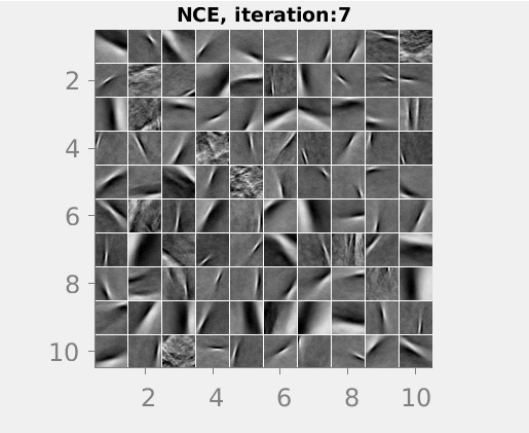


(b) NCE

Figure 8: A sample of the 1st layer features after 6 meta-iterations.

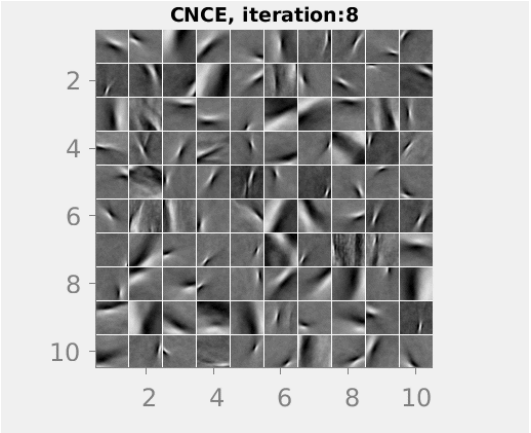


(a) CNCE

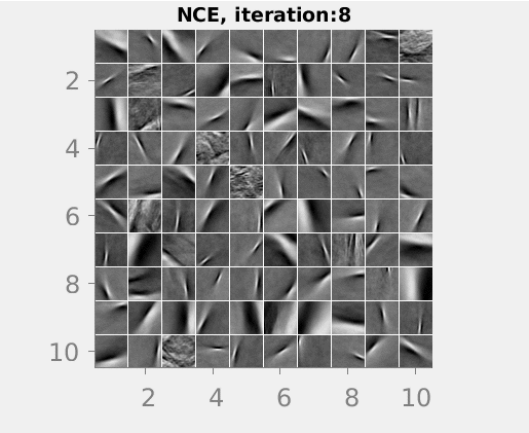


(b) NCE

Figure 9: A sample of the 1st layer features after 7 meta-iterations.

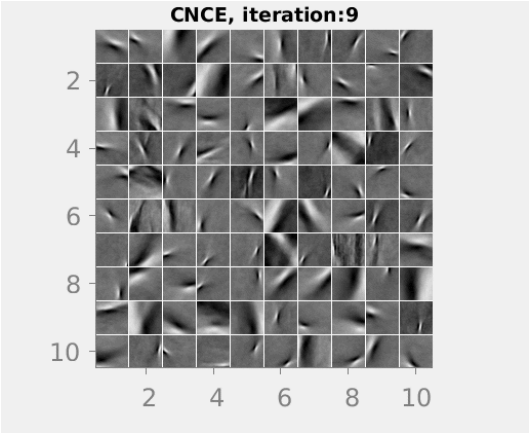


(a) CNCE

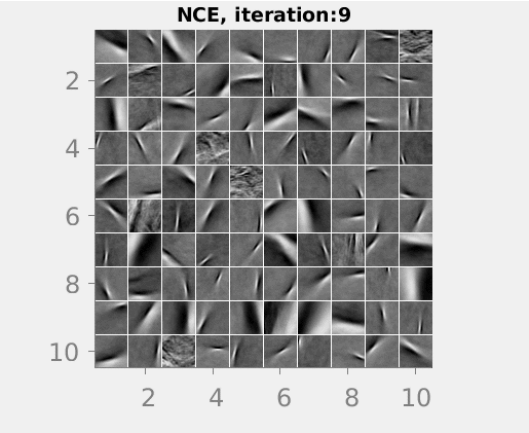


(b) NCE

Figure 10: A sample of the 1st layer features after 8 meta-iterations.

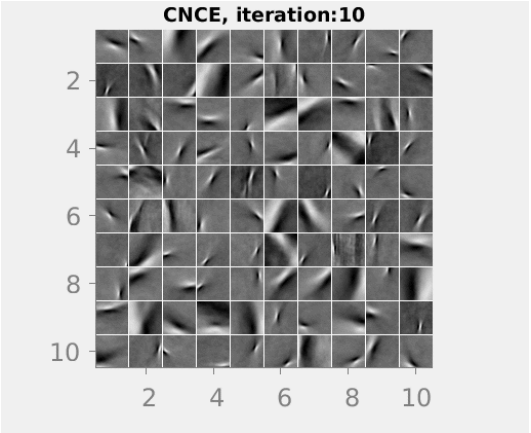


(a) CNCE

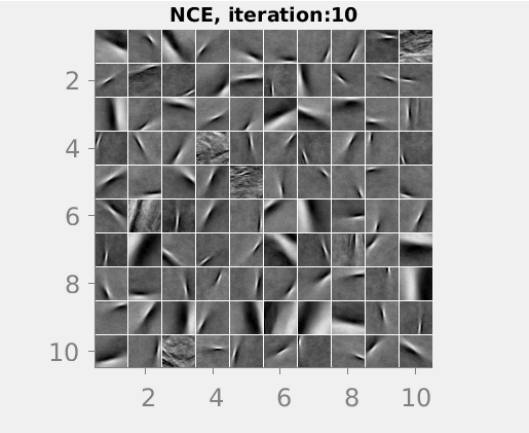


(b) NCE

Figure 11: A sample of the 1st layer features after 9 meta-iterations.

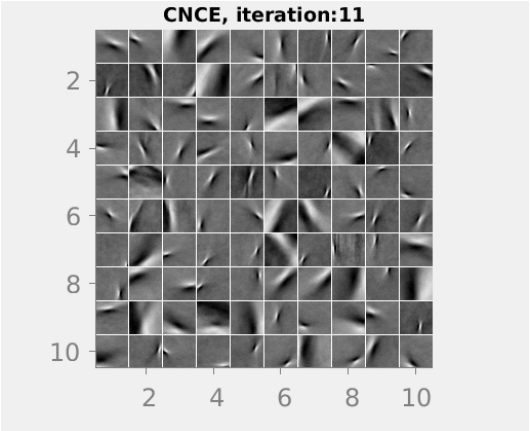


(a) CNCE

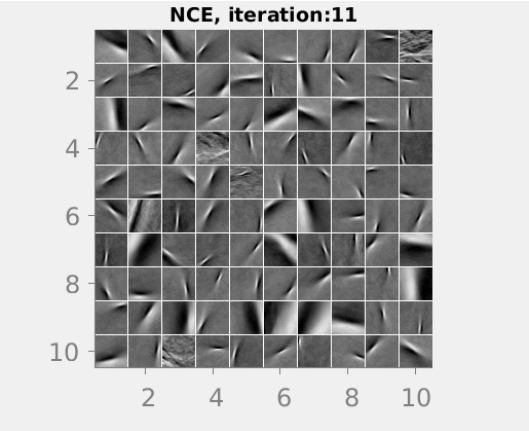


(b) NCE

Figure 12: A sample of the 1st layer features after 10 meta-iterations.

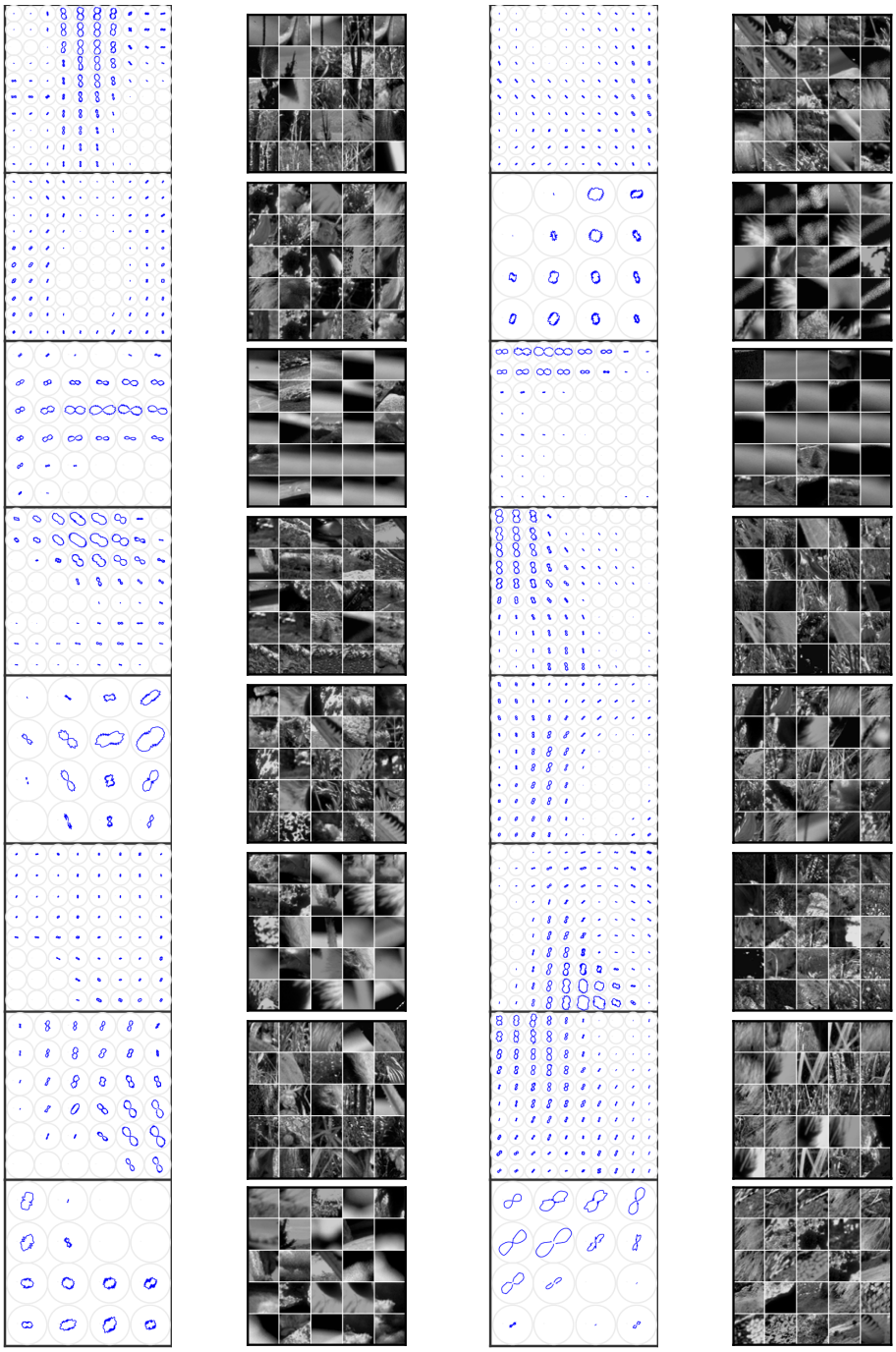


(a) CNCE



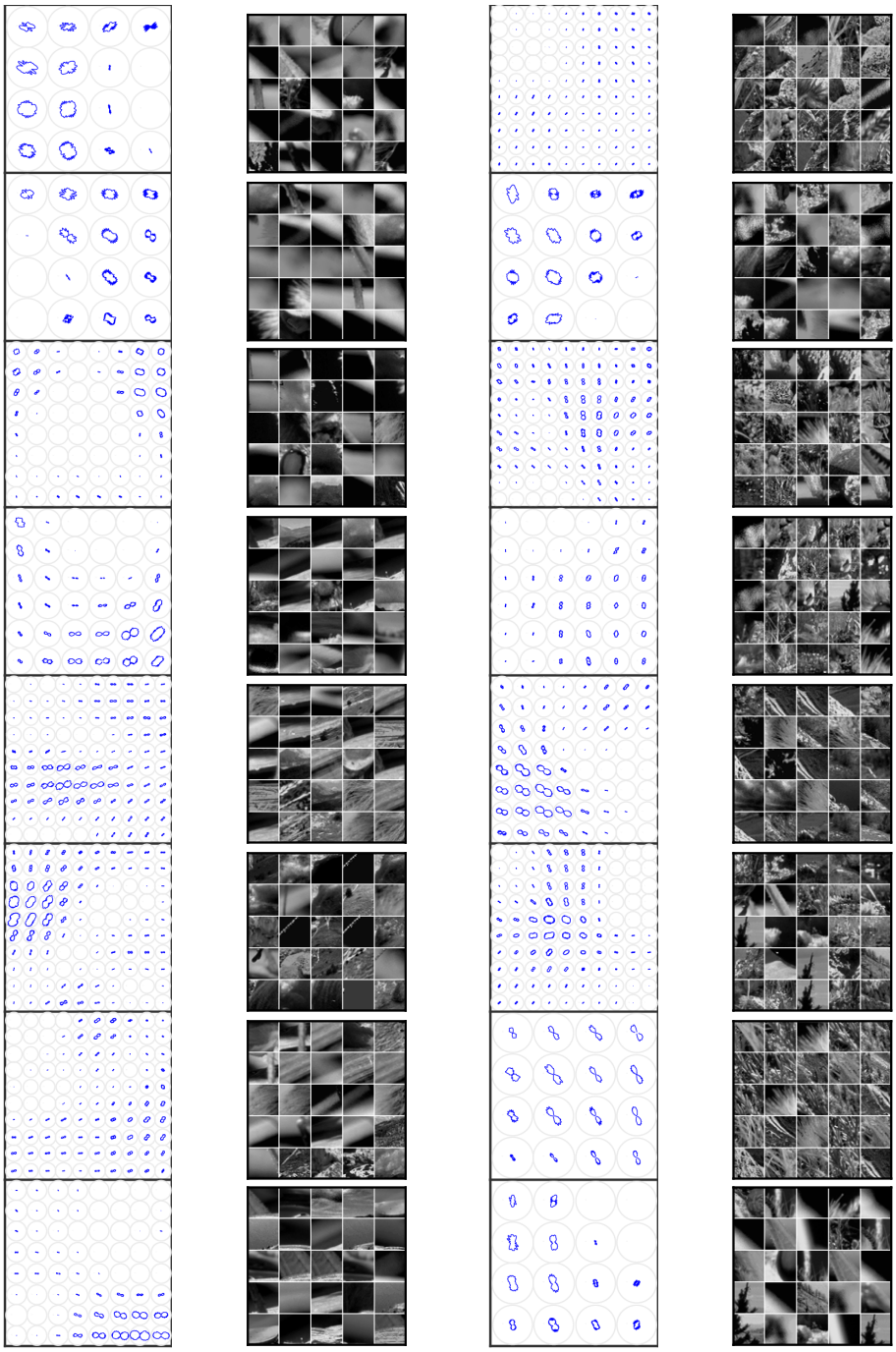
(b) NCE

Figure 13: A sample of the 1st layer features after 11 meta-iterations.



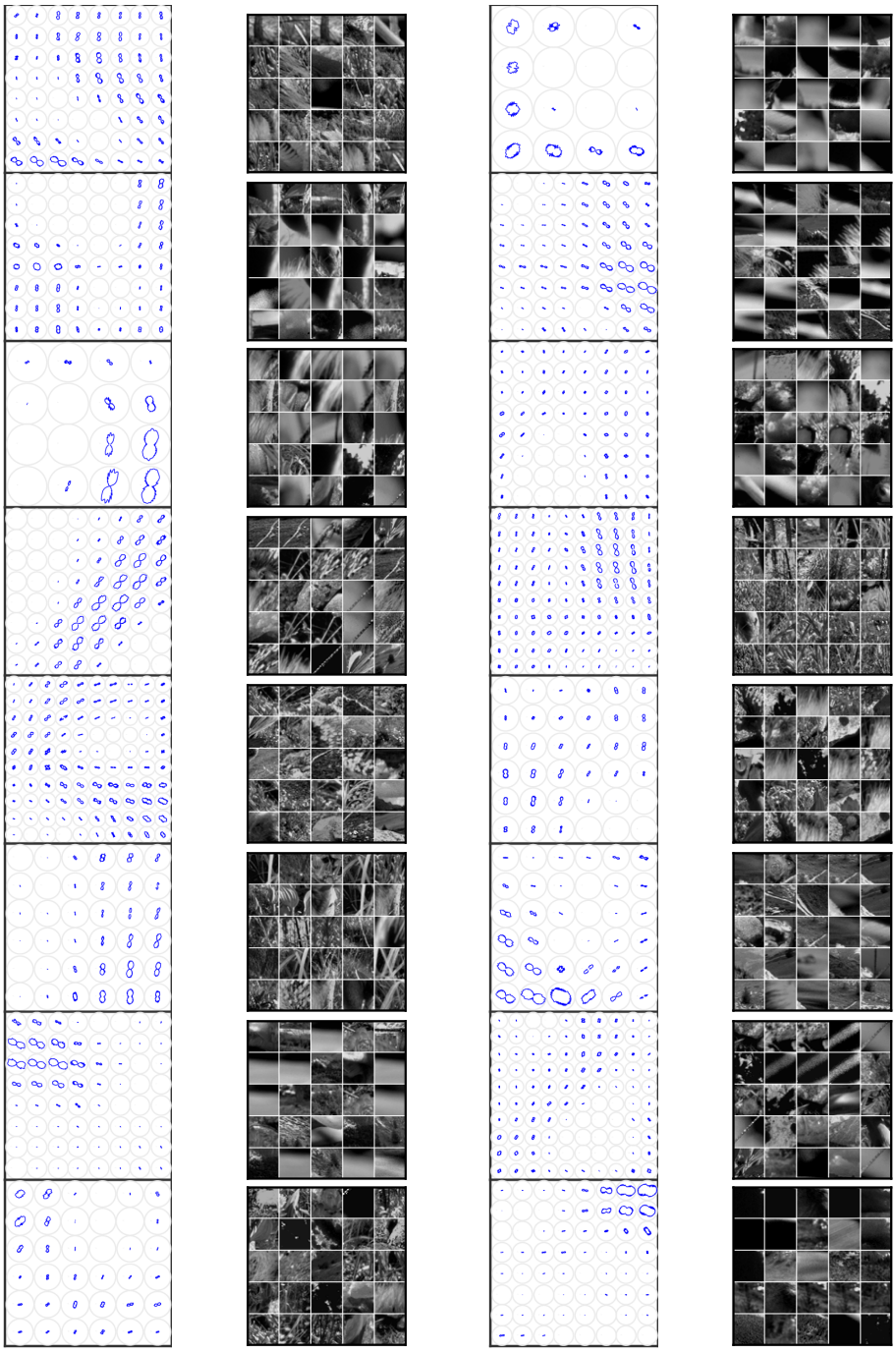
(a) (b) (c) (d)

Figure 14: Each row pair of one receptive field and one icon represent a 3rd layer unit. (a) and (b) show units 1 to 8, and (c) and (d) 9 to 16.



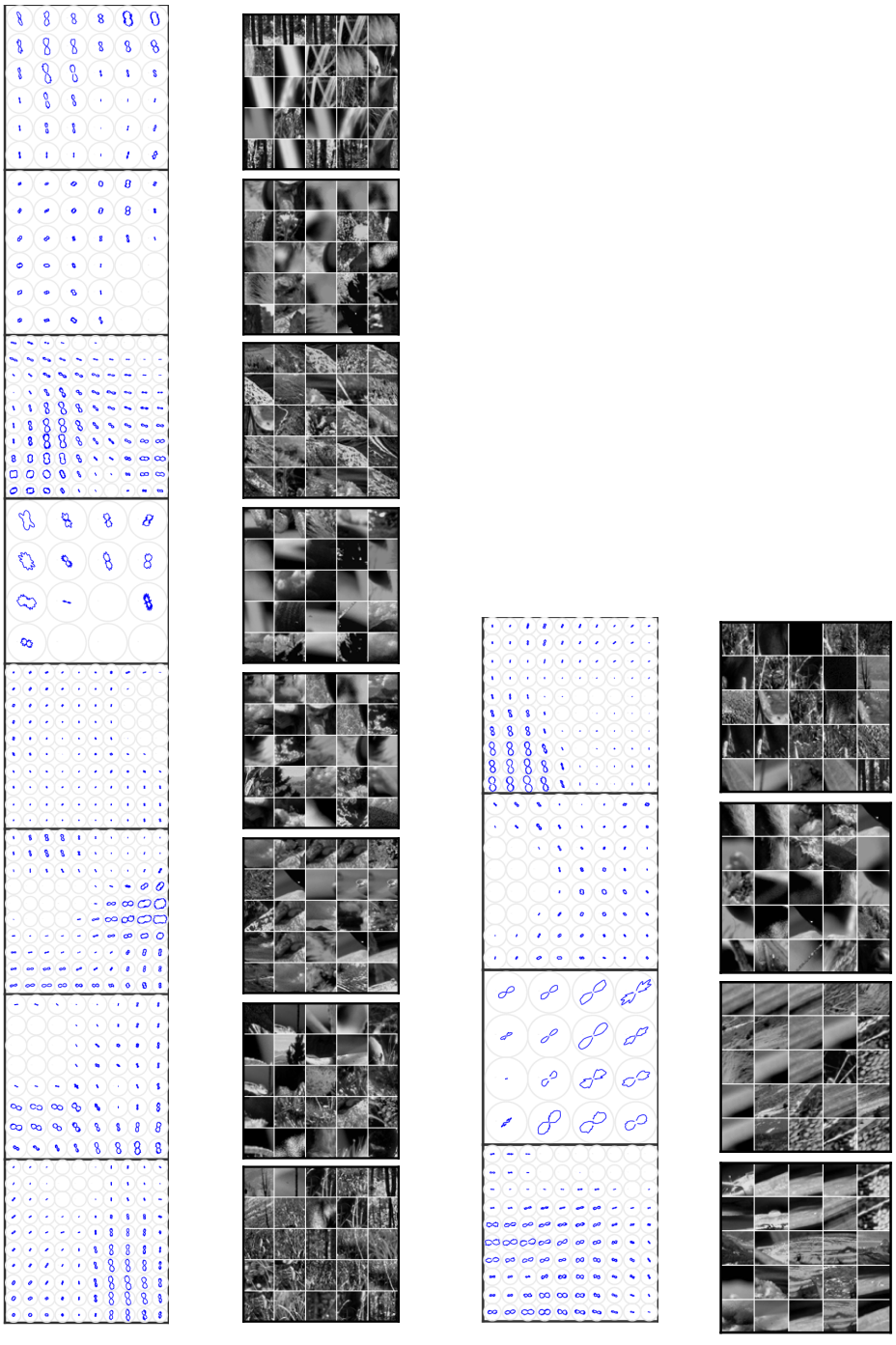
(a) (b) (c) (d)

Figure 15: Each row pair of one receptive field and one icon represent a 3rd layer unit. (a) and (b) show units 17 to 24, and (c) and (d) 25 to 32.



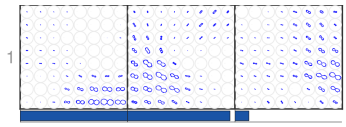
(a) (b) (c) (d)

Figure 16: Each row pair of one receptive field and one icon represent a 3rd layer unit. (a) and (b) show units 33 to 40, and (c) and (d) 41 to 48.

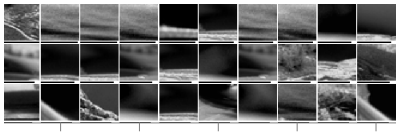


(a) (b) (c) (d)

Figure 17: Each row pair of one receptive field and one icon represent a 3rd layer unit. (a) and (b) show units 49 to 56, and (c) and (d) 57 to 60.

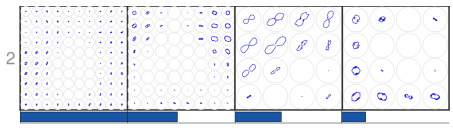


(a) Pooling and space-orientation receptive fields.

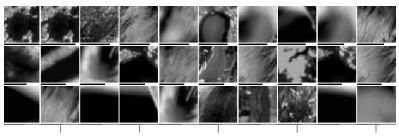


(b) Maximal response inputs.

Figure 18: The estimation for unit 1 in the 4th layer. (a) shows the learned pooling of the 3rd layer units and (b) shows 30 image patches that produced maximal responses for a batch of 10000 inputs. For the space-orientation receptive fields, the bar beneath each icon indicates the relative size of the 4th weight, i.e. $q^{(4)}_{1,k} / \max_k q^{(4)}_{1,k}$. The receptive fields shown account for 90% of the sum of the weight vector. The thin bars beneath each image patch indicate the response strength relative to the patch with the maximal response.

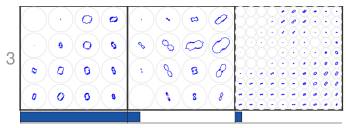


(a) Pooling and space-orientation receptive fields.

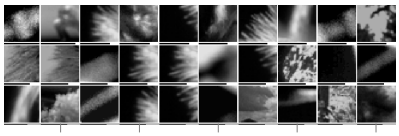


(b) Maximal response inputs.

Figure 19: Unit 2 in the 4th layer, visualised as in Figure 18.

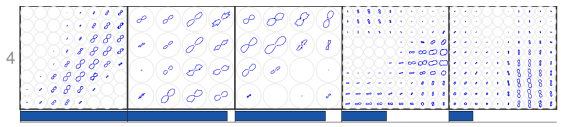


(a) Pooling and space-orientation receptive fields.

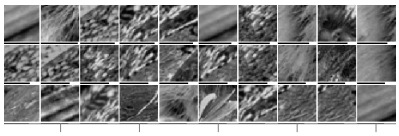


(b) Maximal response inputs.

Figure 20: Unit 3 in the 4th layer, visualised as in Figure 18.

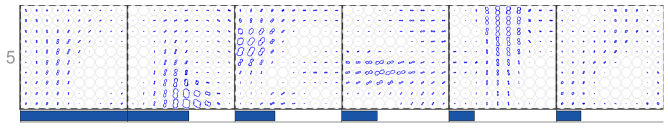


(a) Pooling and space-orientation receptive fields.

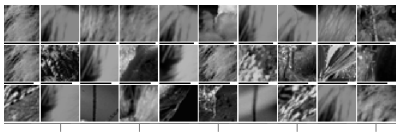


(b) Maximal response inputs.

Figure 21: Unit 4 in the 4th layer, visualised as in Figure 18.

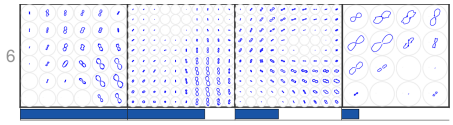


(a) Pooling and space-orientation receptive fields.

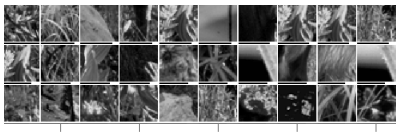


(b) Maximal response inputs.

Figure 22: Unit 5 in the 4th layer, visualised as in Figure 18.

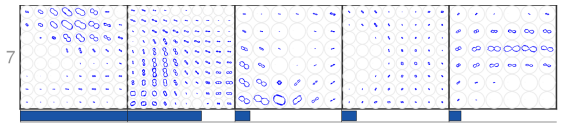


(a) Pooling and space-orientation receptive fields.

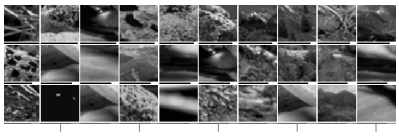


(b) Maximal response inputs.

Figure 23: Unit 6 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

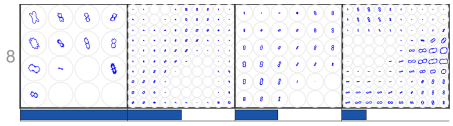


(a) Pooling and space-orientation receptive fields.

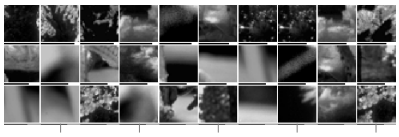


(b) Maximal response inputs.

Figure 24: Unit 7 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

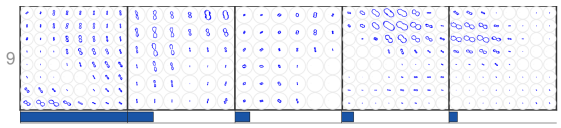


(a) Pooling and space-orientation receptive fields.

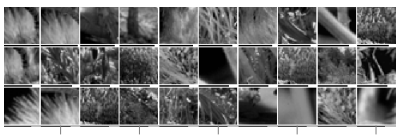


(b) Maximal response inputs.

Figure 25: Unit 8 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

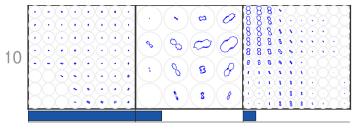


(a) Pooling and space-orientation receptive fields.

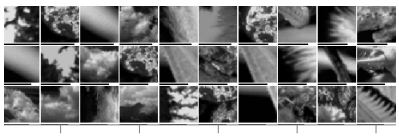


(b) Maximal response inputs.

Figure 26: Unit 9 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

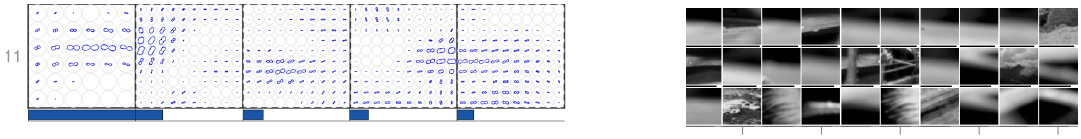


(a) Pooling and space-orientation receptive fields.



(b) Maximal response inputs.

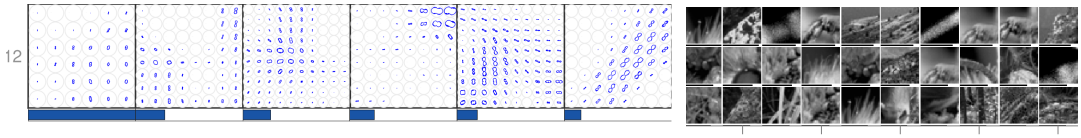
Figure 27: Unit 10 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

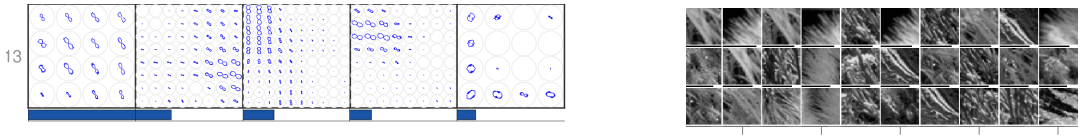
Figure 28: Unit 11 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

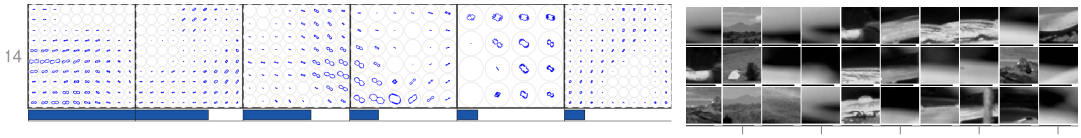
Figure 29: Unit 12 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

Figure 30: Unit 13 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

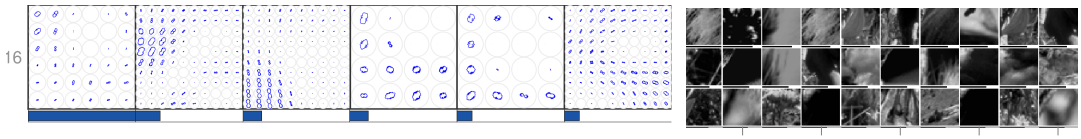
(b) Maximal response inputs.

Figure 31: Unit 14 in the 4th layer, visualised as in Figure 18.

(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

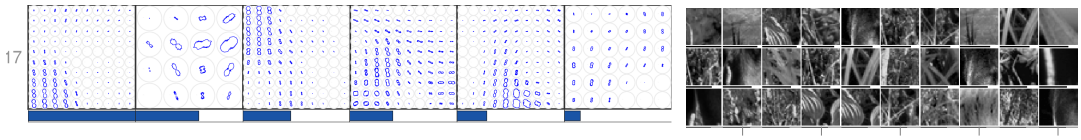
Figure 32: Unit 15 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

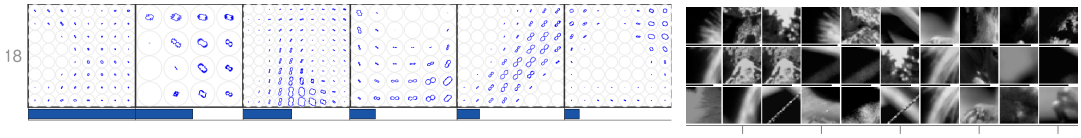
Figure 33: Unit 16 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

Figure 34: Unit 17 in the 4th layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

Figure 35: Unit 18 in the 4th layer, visualised as in Figure 18.

(a) Pooling and space-orientation receptive fields.

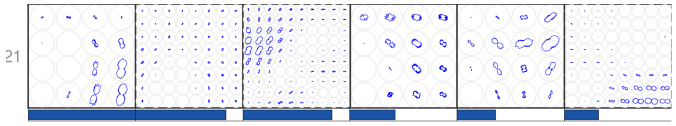
(b) Maximal response inputs.

Figure 36: Unit 19 in the 4th layer, visualised as in Figure 18.

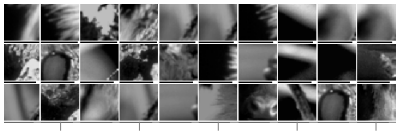
(a) Pooling and space-orientation receptive fields.

(b) Maximal response inputs.

Figure 37: Unit 20 in the 4th layer, visualised as in Figure 18.

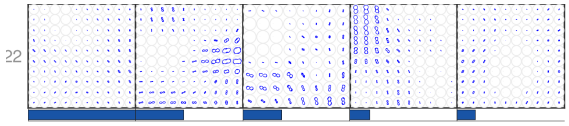


(a) Pooling and space-orientation receptive fields.

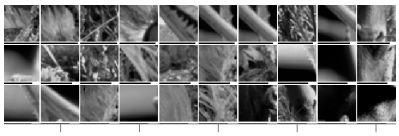


(b) Maximal response inputs.

Figure 38: Unit 21 in the 4th layer, visualised as in Figure 18.

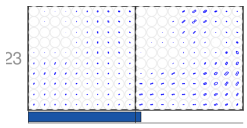


(a) Pooling and space-orientation receptive fields.

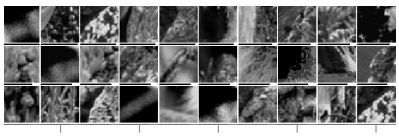


(b) Maximal response inputs.

Figure 39: Unit 22 in the 4th layer, visualised as in Figure 18.

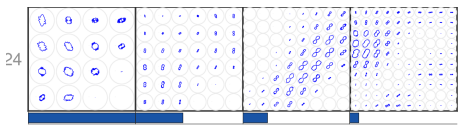


(a) Pooling and space-orientation receptive fields.

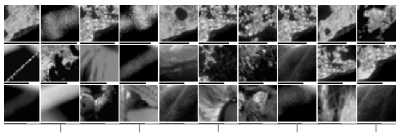


(b) Maximal response inputs.

Figure 40: Unit 23 in the 4th layer, visualised as in Figure 18.

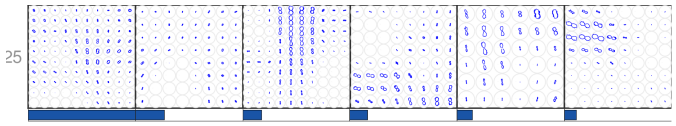


(a) Pooling and space-orientation receptive fields.

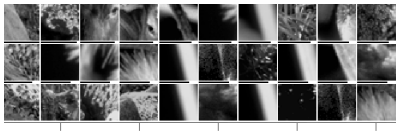


(b) Maximal response inputs.

Figure 41: Unit 24 in the 4th layer, visualised as in Figure 18.

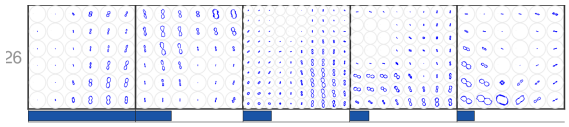


(a) Pooling and space-orientation receptive fields.

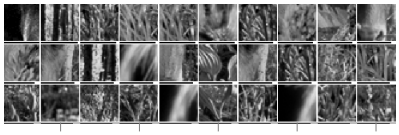


(b) Maximal response inputs.

Figure 42: Unit 25 in the 4th layer, visualised as in Figure 18.

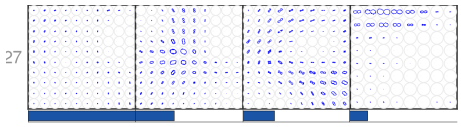


(a) Pooling and space-orientation receptive fields.

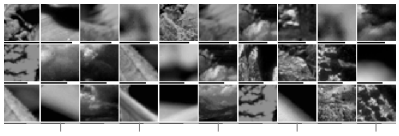


(b) Maximal response inputs.

Figure 43: Unit 26 in the 4th layer, visualised as in Figure 18.

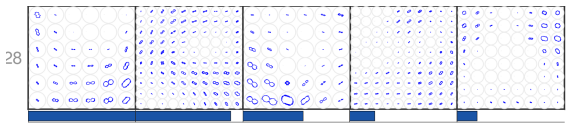


(a) Pooling and space-orientation receptive fields.

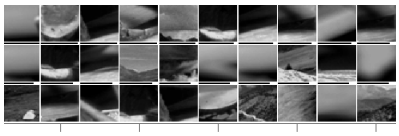


(b) Maximal response inputs.

Figure 44: Unit 27 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

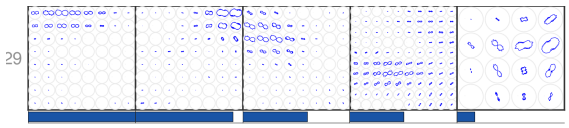


(a) Pooling and space-orientation receptive fields.

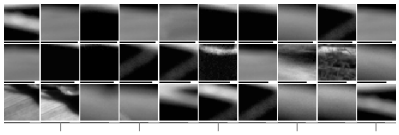


(b) Maximal response inputs.

Figure 45: Unit 28 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.

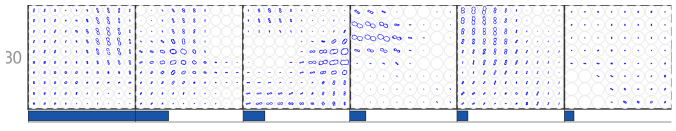


(a) Pooling and space-orientation receptive fields.

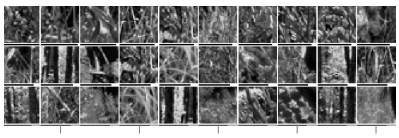


(b) Maximal response inputs.

Figure 46: Unit 29 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.



(a) Pooling and space-orientation receptive fields.



(b) Maximal response inputs.

Figure 47: Unit 30 in the 4$^{\text{th}}$ layer, visualised as in Figure 18.