EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Computer Science Department
Neuro-Cognitive Modeling

Summer Term 2022

Recurrent and Generative Neural Networks

Exercise Sheet 02

Release: May 13, 2022 Deadline: June 2, 2022 (23:59 pm)

General remarks:

- Submit your answers as PDF.
- Ideally, the exercises should be completed in teams of two students. Larger teams are not allowed.
- When questions arise start a forum discussion in ILIAS or contact us via email:
  Sebastian Otte (`sebastian.otte@uni-tuebingen.de`)

Exercise 1 – RNNs Recapitulation [14 points]

(a) Even though that LSTMs (or other gated memory units) use smooth, continuous activation functions (particularly for gating), they can learn to count and solve discrete tasks that involve precise and "sharp" calculations. Explain why this is possible and what you expect to happen during training. [2 points]

(b) Consider a single LSTM unit with D memory cells (recall that an LSTM can have multiple CECs per unit). Following the backward pass equations (lecture 02, slide 23) the forget gate gradient can be calculated with

$$\delta_\phi^t = \varphi'_\phi(net_\phi^t) \sum_{c=1}^{D} \zeta_c^t s_c^{t-1}. \tag{1}$$

Derive this equation step by step from

$$\delta_\phi^t = \frac{\partial E}{\partial net_\phi^t}. \tag{2}$$

You can assume the following definition as given

$$\zeta_c^t =_{\text{def}} \frac{\partial E}{\partial s_c^t}. \tag{3}$$

Keep track of all non-trivial intermediate steps. [10 points]

(c) Can bidirectional RNNs be effectively used for online classification (that is generating a classification label for every new input immediately)? Explain briefly. [2 points].

Exercise 2 – Regularization, Deep Nets [12 points]

(a) Name three forms of commonly applied regularization techniques for neural networks and describe their effects. [6 points]

(b) Explain the internal covariate shift and how it has been addressed in the literature. [2 points]

(c) Consider the following statement: *The more parameter a neural network has the better it will generalize.* If this statement true? Explain briefly. [2 points]

(d) What is the effect of *residual blocks* on the gradient flow in deep networks? [2 points].

Exercise 3 – (Variational) Autoencoder and the KL Divergence [16 points]

(a) What is the general purpose of autoencoders (AEs) and what are their main components? Give an illustration. [4 points]

(b) Briefly explain the purpose of the KL divergence. [2 points]

(c) Given two discrete probability distributions Q, P. Show (by means of an example) that $KL(Q \parallel P) = KL(P \parallel Q)$ does **not** hold in general. [2 points]

(d) Briefly describe the *re-parametrization trick* and explain why it is necessary to train a VAE. [4 points]

(e) Name a frequently mentioned drawback of VAEs and give one example from the literature (reference and brief description) that addresses this issue. [2 points]

(f) What are the major differences of the *variational autoencoders* (VAEs) in comparison with standard autoencoders (name an essential advantage)? Relate your answer to the two sub-losses that are optimized during VAE training. [4 points]

(g) **Bonus**: Given two k-dimensional normal distributions $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ being k-dimensional mean vectors and $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ are positive definite, non-singular $k \times k$ covariance matrices, then it holds

$$KL(\mathcal{N}_1 \parallel \mathcal{N}_2) = \frac{1}{2}\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - k + \log\frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1}\right\}. \tag{4}$$

Show that when the reference distribution is standard normal the following simplification applies

$$KL\left(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \parallel \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)\right) = \frac{1}{2}\sum_{i=1}^{k}(\sigma_i^2 + \mu_i^2 - \log(\sigma_i^2) - 1), \tag{5}$$

if $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_k^2)$ ($\boldsymbol{\Sigma}$ is a diagonal matrix). [4 bonus points]

Exercise 4 – Generative Adversarial Networks [8 points]

(a) Illustrate the major structure of GANs and name the essential components. Briefly describe the principle. [4 points]

(b) Elaborate on why the training of GANs is considered as particularly difficult/challenging and how it is addressed. Comment on what happens if the discriminator dominates the generator and vice versa. [4 points]