DEEP LEARNING (SS2022)
EXAM T1 (90 MIN / 26P)

Assignment 1 (5p)**.** Let the neural network consist of first layer, $f(x, w)$ with input x and parameters w and the second layer $g(f, v)$ with parameters v. Assume that all inputs and functions f and g are vector-valued and $\mathcal{L}$ is scalar-valued.

a) Draw the computation graph. Express gradients of the loss $\mathcal{L}(g)$ in parameters w, v and input x using Jacobians of all given functions: $J_x^f = \frac{\partial f(x,w)}{\partial x}$, $J_w^f = \frac{\partial f(x,w)}{\partial w}$, $J_f^g = \frac{\partial g(f,v)}{\partial f}$, $J_v^g = \frac{\partial g(f,v)}{\partial v}$, $J_g^{\mathcal{L}} = \frac{\partial \mathcal{L}(g)}{\partial g}$.

b) In which order to organize the computation and which intermediate results can be reused for better efficiency?

Assignment 2 (5p)**.** Let $\mathcal{L}(x)$ be the loss function which we want to minimize in x. Let $\tilde{g}(x)$ be its stochastic gradient estimate at a point x. SGD with momentum is given by:

$$v_{t+1} = \mu v_t + \tilde{g}(x_t); \tag{1a}$$

$$x_{t+1} = x_t - \varepsilon v_{t+1}. \tag{1b}$$

a) Express x_{t+1} through x_t, x_{t-1} and $g(x_t)$ only, eliminating the sequence v_t.

b) Show that if the sequence x_t converges ($\|x_{t+1} - x_t\| \to 0$), then necessarily the sequence of gradients $g(x_t)$ converges to zero $\|g(x_t)\| \to 0$.

Assignment 3 (3p)**.** Give a recursive formula for computing the size and stride of the receptive fields of neurons in a convolutional network. *I.e.* express the size R_k and stride S_k of the receptive fields of neurons in layer k in terms of the respective values for neurons in layer $k-1$ and the kernel size r_k and stride s_k of the convolutional layer k.

Assignment 4. (5p) We want to learn a deep feature representation $f(x;\theta)$ of images x such that images of the same class are close together in the feature space and images of different classes are far away. The training set is given in the form of all triplets (x_a, x_p, x_n) such that x_p is *positive*, *i.e.* the same class as x_a, and x_n is *negative*, *i.e.* is of different class than x_a. The mapping f is parametrized by θ and let (f_a, f_p, f_n) be the triplet of feature vectors corresponding to (x_a, x_p, x_n). The triplet loss is:

$$l(f_a, f_p, f_n) = \max(\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2 + \alpha, 0),$$

where $\alpha > 0$ is a hyperparameter.

a) Show that if the loss is zero for all triplets, then for each query x_a the nearest neighbour in the feature space is a positive match for x_a.

b) For a given triplet, compute the gradient of the loss in f_a, f_p, f_n. Draw f_a, f_p, f_n as points in general positions. Draw in which directions they should be moved to provide the steepest decrease of the loss.

c) Express the gradient of the loss for a given triplet in the network parameters θ.

Assignment 5 (4p)**.** The exponential distribution of a random variable $X \in \mathbb{R}_+$ has density $p_\lambda(x) = \lambda e^{-\lambda x}$. Derive the formula for the Kullback Leibler divergence $D_{KL}(p_\lambda \parallel p_\mu)$ of two exponential distributions with parameters λ and μ.

Assignment 6 (4p)**.** Let us consider the expression

$$L(x, w) = \mathbb{E}_{y \sim p_\lambda} \ell(y) = \int_0^\infty dy\, p_\lambda(y) \ell(y),$$

where p_λ is an exponential distribution with parameter $\lambda = f(x, w)$ given by a neural network. Computing the gradient $\nabla_w L(x, w)$ is not feasible, because it would require to compute the integral in the r.h.s. Derive a reparametrisation trick for computing a stochastic estimator of the gradient $\nabla_w L(x, w)$ from a sample of $y \sim p_\lambda(y)$. We assume that the network output is differentiable in its parameters w and the loss $\ell(y)$ is differentiable in y. *Hint:* Recall that the reparametrisation trick replaces sampling from a parametric distribution by sampling from a distribution with fixed parameter combined with some suitably chosen transformation.