DEEP LEARNING (SS2022)
SEMINAR 5

Assignment 1 (Receptive fields)**.** Consider a convolutional network consisting of convolution layers and max-pooling layers. Each of them is characterised by a kernel size k_ℓ and a stride s_ℓ. The *receptive field* of a neuron in layer ℓ is the bounding box of all nodes in the input layer that can influence its output. Similarly, the *stride of the receptive field* is defined as the translation between receptive fields of two neighbouring neurons in the same layer. Knowing the receptive field size and stride of neurons in layer $l-1$ and the type and attributes of layer l, find the receptive field size and stride of neurons in layer l.
Note: This relation will be needed for the lab on CNN visualisation & adversarial patterns.

Assignment 2 (Trust Region Problem with Box Constraints, FGSM)**.**
Let us consider a loss function $L(\theta)$ and denote its gradient at θ^t by $g^t = \nabla_\theta L(\theta^t)$.

a) Solve the following trust region problem:

$$\begin{aligned} &\arg\min_\theta \big[L(\theta^t) + \langle g, \theta - \theta^t \rangle\big], \\ &\text{s.t. } |\theta_i - \theta_i^t| \leq \varepsilon \; \forall i \end{aligned}$$

by using the technique of Lagrange multipliers.
Hint: Make a substitution of variables $\Delta\theta = \theta - \theta^t$. Square the constraints to make their derivative simpler. Note that the Lagrange multipliers for inequality constraints must be non-negative.

b) Show that the fast gradient sign attack described in the lecture solves a similar constrained optimization problem (formulate this problem).

Assignment 3 (Proximal Problem with Regularization).
Consider the problem of minimizing the training loss of a neural network with a weight regularization $\frac{\lambda}{2}\|\theta\|^2$. The normal SGD makes steps by solving the proximal problem:

$$\theta_{t+1} := \arg\min_{\theta}\Big[\langle g_t + \lambda\theta_t, \theta - \theta_t\rangle + \frac{1}{\varepsilon}\|\theta - \theta_t\|^2\Big], \tag{1}$$

where θ_t is the current parameter vector, g_t is the stochastic gradient of the training loss at θ_t, λ is the regularization strength and ε is the learning rate.

a) Verify that the solution of the proximal step problem indeed leads to a common SGD with a weight decay.

b) Since the weight regularization is known in closed form, we do not need to approximate it linearly by computing its derivative. Derive an "improved" SGD algorithm that makes steps by solving the composite proximal step problem:

$$\theta_{t+1} := \arg\min_{\theta}\Big[\langle g_t, \theta - \theta_t\rangle + \lambda\|\theta\|^2 + \frac{1}{\varepsilon}\|\theta - \theta_t\|^2\Big]. \tag{2}$$

c) Is the SGD in b) equivalent to SGD in a) with a possibly different choice of λ, ε?

Assignment 4 (BN with Weight Decay).
We discussed in the lecture that combining Batch Normalisation with weight decay regularisation leads to an ill posed optimisation problem. Let us consider this in a simplified scenario for a single neuron. Its output is given by $y = \frac{w^\intercal x}{\|w\|}$, where x is the input. The regularized loss function is given by $\tilde{L}(w) = L(y(w)) + R(w)$, where $R(w) = \frac{\lambda}{2}\|w\|^2$ and $\lambda > 0$.

a) Compute the gradient of $R(w)$ and show that a step towards decreasing it can be interpreted as "weight decay". Suppose that w_0 is optimal for the non-regularized loss L. What will the gradient descent on $\tilde{L}$ do if started at w_0?

b) Consider a point w_0 on the unit sphere for which the gradient $g = \nabla_w L(y(w_0))$ is non-zero. Show that g is orthogonal to w_0 and hence also to $\nabla_w R(w_0)$. Draw these vectors and the sphere $\|w\| = 1$.

c) Let $\|g\| = a$ and $\|\nabla_w R(w_0)\| = \lambda$ at w_0 with $\|w_0\| = 1$. Consider a single step of the gradient descent for $\tilde{L}$ with step length $\alpha > 0$. Give a condition on α that ensures a decreasing norm $\|w\|$.