Data standardization and augmentation Prior to feeding the data to the neural network for training, some preprocessing is usually done. Many beginners fail to obtain reasonable results not because of the architectures or methods or lack of regularization, but instead because they simply did not normalize and visually inspect their data. Two most important forms of pre-processing are data standardization and dataset augmentation. There are a few data standardization techniques common in imaging. ● Mean subtraction. During mean subtraction, the mean of every channel is computed over the training dataset, and these means are subtracted channelwise from both the training and the testing data. ● Scaling. Scaling amounts to computing channelwise standard deviations across the training dataset, and dividing the input data channelwise by these values so as to obtain a distribution with standard deviation equal to 1 in each channel. In place of division by standard deviation one can divide, e.g., by 95-percentile of the absolute value of a channel. ● Specialized methods. In addition to these generic methods, there are also some specialized standardization methods for medical imaging tasks, e.g., in chest X-ray one has to work with images coming from different vendors, furthermore, X-ray tubes might be deteriorating. In [17] local energy-based normalization was investigated for chest X-ray images, and it was shown that this normalization technique improves model performance on supervised computer-aided detection tasks. For another example, when working with hematoxylin and eosin (H&E) stained histological slides, one can observe variations in color and intensity in samples coming from different laboratories and performed on different days of the week. These variations can potentially reduce the effectiveness of quantitative image analysis. A normalization algorithm specifically designed to tackle this problem was suggested in [18], where it was also shown that it improves the performance for a few computer-aided detection tasks on these slide images. Finally, in certain scenarios (e.g., working directly with raw sinogram data for CT or Digital Breast Tomosynthesis [19]) it is reasonable to take log-transform of the input data as an extra preprocessing step. Neural networks are known to benefit from large amounts of training data, and it is a common practice to artificially enlarge an existing dataset by adding data to it in a process called "augmentation". We distinguish between train-time augmentation and test time augmentation, and concentrate on the first for now (which is also more common). In case of train-time augmentation, the goal is to provide a larger training dataset to the algorithm. In a supervised learning scenario, we are given a dataset D consisting of pairs (x_j, y_j) of a training sample $x_j \in R^d$ and the corresponding label y_j. Given the dataset D, one should design transformations $T_1, T_2, ..., T_n : R^d \to R^d$ which are label-preserving in a sense that for every sample $(x_j, y_j) \in D$ and every transformation T_i the resulting vector $T_i x_j$ still looks like a sample from D with label y_j. Multiple transformations can be additionally stacked, resulting in greater number of new samples. The resulting new samples with labels assigned to them in this way are added to the training dataset and optimization as usual is performed. In case of the test-time augmentation the goal is to improve test-time performance of the model as follows. For a predictive model f, given a test sample $x \in R^d$, one computes the model predictions $f(x), f(T_1 x), . . . , f(T_n x)$ for different augmenting transformations and aggregates these predictions in a certain way (e.g., by averaging softmax-output from classification layer [6]). In general, choice of the augmenting transformation depends on the dataset, but there are a few common strategies for data augmentation in imaging tasks: ● Flipping. Image x is mirrored in one or two dimensions, yielding one or two additional

samples. Flipping in horizontal dimension is commonly done, e.g., on the ImageNet dataset [6], while on medical imaging datasets flipping in both dimensions is sometimes used. ● Random cropping and scaling. Image x of dimensions $W \times H$ is cropped to a random region $[x1, x2] \times [y1, y2] \subseteq [0,W] \times [0,H]$, and the result is interpolated to obtain original pixel dimensions if necessary. The size of the cropped region should still be large enough to preserve enough global context for correct label assignment. ● Random rotation. An image x is rotated by some random angle ϕ (often limited to the set $\phi \in [\pi/2, \pi, 3\pi/2]$). This transformation is useful, e.g., in pathology, where rotation invariance of samples is observed; however, it is not widely used on datasets like ImageNet. ● Gamma transform. A grayscale image x is mapped to image x^γ for $\gamma > 0$, where $\gamma = 1$ corresponds to identity mapping. This transformation in effect adjusts the contrast of an image. ● Color augmentations. Individual color channels of the image are altered in order to capture certain invariance of classification with respect to variation in factors such as intensity of illumination or its color. This can be done, e.g., by adding small random offsets to individual channel values; an alternative scheme based on PCA can be found in [6].