Deep neural networks have been recently achieving high accuracy on many important tasks, most notably image classification. However, these models are not robust to slightly perturbed inputs known as adversarial examples. These can severely decrease the accuracy and thus endanger systems where such machine learning models are employed. We present a review of adversarial examples literature. Then we propose new defenses against adversarial examples: a network combining RBF units with convolution, which we evaluate on MNIST and get better accuracy than with an adversarially trained CNN, and input space discretization, which we evaluate on MNIST and ImageNet and obtain promising results. Finally, we explore a way of generating adversarial perturbation without access to the input to be perturbed.