

## ABSTRACT

Methods, systems, and computer program products are provided for cleaning noisy data from unlabeled datasets using autoencoders. A method includes receiving training data including noisy samples and other samples. An autoencoder network is trained based on the training data to increase a first metric based on the noisy samples and to reduce a second metric based on the other samples. Unlabeled data including unlabeled samples is received. A plurality of third outputs is generated by the autoencoder network based on the plurality of unlabeled samples. For each respective unlabeled sample, a respective third metric is determined based on the respective unlabeled sample and a respective third output, and whether to label the respective unlabeled sample as noisy or clean is determined based on the respective third metric and a threshold. Each respective unlabeled sample determined to be labeled as noisy is cleaned.

<Figure 1>