ENCODER DISCRIMINATOR NETWORKS FOR UNSUPERVISED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning representations of data samples in an unsupervised way is needed whenever computers have to reason about unlabeled data. Applications range from compressing and denoising data to super-resolution, generating new samples from a given sample distribution and much more. In this work, we use information entropy and a little game to motivate a new encoder discriminator architecture in order to learn unsupervised latent representations. Inspired by the game ”Taboo”, we train an encoder network to generate a meaningful representation of one particular sample of a dataset. Using this description, a discriminator network then has to retrieve the same sample from the whole dataset. We show that learning in this manner on many different samples repeatedly minimizes the information entropy given the latent description and, thus, forces the encoder network to make precise descriptions that can be interpreted by the discriminator. We provide first results of this method on the MNIST and the Fashion MNIST dataset.

1 INTRODUCTION AND RELATED WORK

While in supervised machine learning the goal is to find a mapping of samples onto predefined labels, in unsupervised learning algorithms not only have to find a mapping but also — and in particular — to find meaningful representations themselves. The reasons to do this are manifold:

Hinton & Salakhutdinov (2006) show how to generate compressed representations of data using deep autoencoders. Vincent et al. (2008) show how autoencoders can be used to denoise images. This requires finding the hidden structure of the data behind the noise and thus representations that are independent of noise added to the input. Dong et al. (2016) present a way to increase the resolution of lower resolution input images. Most of these encoder-decoder methods optimize their output to match the training data with respect to some metric (usually L_2). These metrics however do not capture the statistics of images well and hence often lead to blurry results in the case of model uncertainty.

Another trend in unsupervised learning are generative adversarial networks (GANs) as introduced by Goodfellow et al. (2014). By training a discriminative loss function to distinguish between samples that originate from the dataset and samples coming from a generator network, new samples can be generated from a latent noise input. These samples capture the statistics of the dataset and often look surprisingly realistic. Radford et al. (2015) showed how the latent noise input can be related to the generator output and was able to perform vector arithmetic for visual concepts in that latent space. However in these generator-discriminator methods, finding a latent representation for a visual concept is difficult — especially in the presence of mode collapse. Furthermore, retrieving a latent representation of a particular sample is impossible since no encoder is given.

In this work, we present a new architecture that combines the benefits of the encoder part of autoencoders with the discriminator part of GANs in order to generate latent sample descriptions that are in good accordance with the statistics of the dataset. We motivate the network using the game ”Taboo” — a word guessing game where words have to be described without using forbidden phrases — and show how this leads to optimal representations by means of minimizing information entropy.

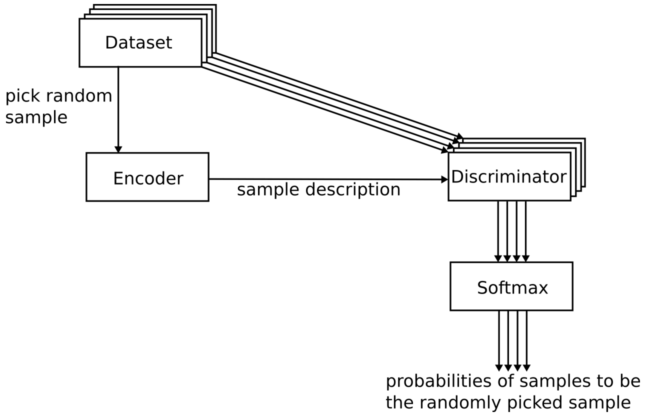

Figure 1: Visualization of the Encoder Discriminator concept

2 Encoder Discriminator Networks

In this section, we introduce encoder discriminator networks as a new way to learn latent representations in an unsupervised way. First, the architecture is motivated by a cooperative game similar to ”Taboo”. Then, we want to reveal how the game can be implemented and our underlying network architecture. In the end of this section, we want to give a more thorough explanation of the theoretical aspects and derive the loss function from an information entropy argument.

2.1 ”Taboo” - a cooperative game

Taboo is a famous game in which players in a team try to find words collaboratively: First, one player picks a random card from a pile and keeps it secret. This card contains a word which the player has to describe as precisely as possible so that it can be guessed by the teammates. To make the game harder, the player is constrained by several words which must not appear in the description. The team which makes the best descriptions and finds the words the fastest wins the game.

The encoder discriminator architecture follows this analogy: First, the encoder picks a random sample from a dataset which contains, for example, an image that needs to be found by the discriminator. Then, the encoder has to make an as precise as possible description of this sample but is constrained by the dimensionality and boundedness of the latent description space. The loss function is minimized when the descriptions of the encoder are precise enough and interpretable by the discriminator network such that the discriminator can retrieve the sample from the whole dataset with high confidence. So in contrast to GANs, where the discriminator is trained adversely, in this setting the discriminator works collaboratively with the encoder and thus the network can be trained end-to-end.

2.2 Encoder Discriminator Architecture

Figure 1 shows the outline of an Encoder Discriminator network. First, a random sample of the dataset is picked and fed into an encoder network. This encoder network generates a latent description of the picked sample and passes it to a discriminator network. The discriminator network takes the sample description and compares it with the original as well as other samples from the dataset. For each sample, the discriminator returns a confidence value of how certain the network is, that the latent description indeed comes from that sample. These confidence values are then passed to a softmax function which computes the predicted probabilities of the samples to be the original randomly picked sample. Finally, the aim of this network is to minimize the cross entropy between the picked samples and the computed probability distribution. A more detailed overview of the architectures used in the encoder and discriminator networks is given in Figure 2.

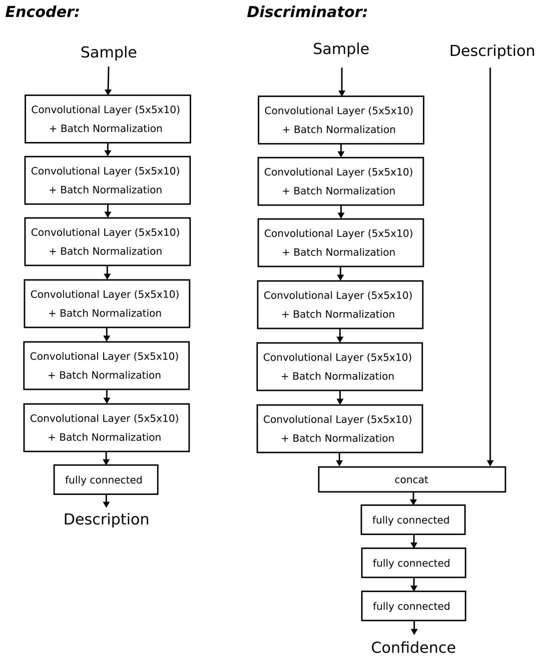

Figure 2: Architecture of the Encoder and Discriminator networks. The convolutional layers use relu activation functions, the fully connected layers use sigmoid activations.

2.3 Discriminator Loss

In this section, we want to motivate our method from an information entropy stand point. So let us first assume, we have an encoder function, latent(x_j), which maps x_j from sample space onto a latent description space. Then we can introduce the probability distribution $p(x_i = x_j|\text{latent}(x_j))$ over x_i which describes the probability that sample x_i is equal to x_j given the latent representation of x_j. The latent description is most precise if latent(x_j) is chosen such that the entropy of $p(x_i = x_j|\text{latent}(x_j))$ is minimized:

$$\begin{aligned} &H(p(x_i = x_j|\text{latent}(x_j))) = \\ &-\sum_{x_i} p(x_i = x_j|\text{latent}(x_j)) \log(p(x_i = x_j|\text{latent}(x_j))) \end{aligned}$$

This entropy value is minimized when there is no x_i other than x_j which is mapped on the same latent(x_j). In other words, this means latent(x_j) provides a unique description of x_j. Now, we can sum the entropy of $p(x_i = x_j|\text{latent}(x_j))$ over all x_j in order to enforce latent(x_j) to provide good descriptions for all samples:

$$\hat{H} = \sum_{x_j} H(p(x_i = x_j|\text{latent}(x_j)))$$

$\hat{H}$ is minimized by latent(x_j), if latent(x_j) provides a unique description for every x_j.

This condition alone, however, is not sufficient to also learn meaningful representations since a simple hashing function which returns a unique hash value for every sample would give perfect results with respect to $\hat{H}$. In the end, we want to make sense of the latent description and this is where the discriminator network $q(latent(x_j), x_i)$ comes in. By training q on p, we can ensure that the descriptions of the encoder are interpretable by the discriminator. By minimizing the cross entropy between p and q, we can train q to become closer and closer to p. Now the loss function becomes:

$$\begin{aligned} L(latent, q) &= \sum_{x_j, x_i} -p(x_i = x_j | \text{latent}(x_j)) \log(q(\text{latent}(x_j), x_i)) \\ &= \sum_{x_j, x_i} -\delta_{i,j} \log(q(\text{latent}(x_j), x_i)) \end{aligned}$$

Here $\delta_{i,j}$ is 1 for i=j and 0 otherwise. This is a valid assumption since the probability that latent(x_j) maps two different samples of the dataset onto exact the same position in the continuous latent space is basically zero. Thus, $p(x_i = x_j | \text{latent}(x_j))$ becomes $\delta_{i,j}$. As q approaches p, the loss function approximates the entropy $\hat{H}$ as defined above. Hence, training on this loss end-to-end not only optimizes the discriminator q with respect to the cross-entropy but also the encoder to make as precise as possible descriptions. Note, however, that q can not be calculated with one separate sample alone but always needs other samples to be properly normalized in the softmax layer.

2.4 LATENT DISTANCE

With the results obtained above, we can introduce a probabilistic measure of distance, which might be helpful to find related samples in a dataset:

$$d(x_i, x_j) = \log \left(\frac{q(\text{latent}(x_i), x_i)}{q(\text{latent}(x_i), x_j)} \right)$$

Note, that this definition of a distance is neither commutative nor positive definite and should not be confused with the L_2 distance in latent space. It rather is a measure of how likely it is that the discriminator confuses x_j with x_i given the latent description of x_i. The smaller the distance, the more likely is a confusion and the more similar are x_i and x_j. If $x_i = x_j$, the distance is 0.

3 EXPERIMENTS

We performed two experiments using the proposed encoder discriminator architecture. The first one was performed on the MNIST (Modified National Institute of Standards and Technology) database (LeCun et al. (2010)). The second was done on the Fashion MNIST dataset (Xiao et al. (2017)).

3.1 TRAINING DETAILS

The network was trained on batches of 100 samples and the latent space was chosen to be two dimensional and bounded between zero and one. We furthermore added Gaussian noise and random affine transformations to the samples before they were passed to the encoder and discriminator networks. This greatly improves the performance since it enforces the encoder to learn more robust representations.

3.2 MNIST

The MNIST dataset contains 60.000 training and 10.000 test images of handwritten digits. Figure 3 shows how the encoder network maps samples of the 10 different digit classes onto the latent space. The encoder clearly separates the different digit classes and thus learns useful abstraction features. However there are still problems in particular in between the digits classes 8 and 5 as well as 4 and 9. This matches our intuition that these classes should be the hardest to discriminate. Another observation is that the encoder fills the available latent space "densely", because the encoder has to use the available latent space as efficiently as possible. This is in contrast to other common clustering algorithms that usually try to cover semantic differences in the form of larger L_2 distances in latent space. Figure 4 shows some MNIST images plotted in the latent space. As one can clearly see in the class of ones, the encoder also learns additional features (like the digit angle) that help the discriminator to also distinguish between samples of the same class.

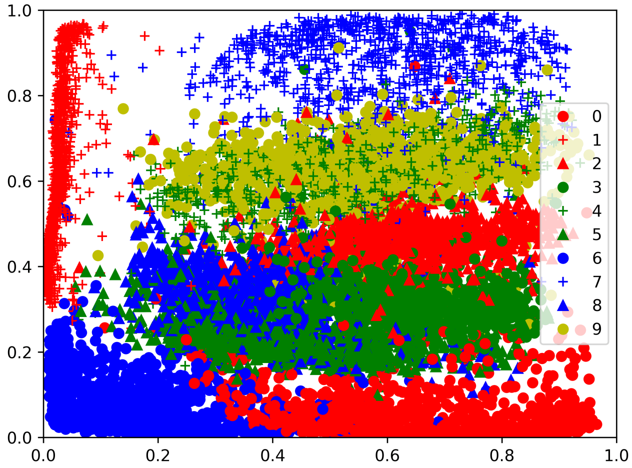

Figure 3: Test labels of the MNIST set plotted in the two dimensional latent space. A plot of the samples themselves in latent space can be found in Figure 4

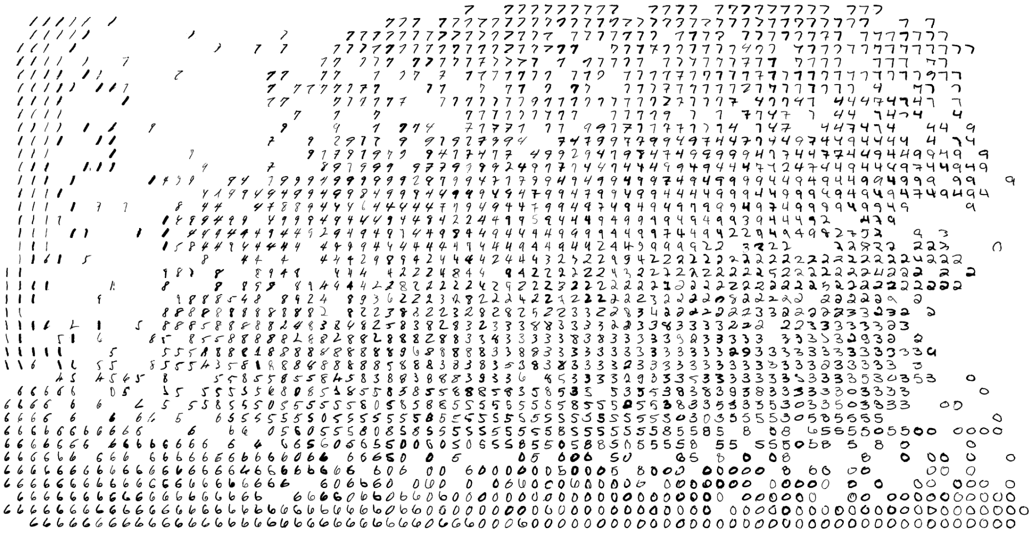

Figure 4: Test samples of the MNIST set plotted in the two dimensional latent space. Best viewed in electronic form

3.3 Fashion MNIST

The Fashion MNIST dataset is a dataset containing 60.000 training and 10.000 test images of 10 different product categories available at Zalando. These categories comprise T-shirts/tops, trousers, pullovers, dresses, coats, sandals and more. The idea behind this dataset is that different classes of product categories contain samples with higher variability than different digits of the original MNIST dataset and thus are harder to cluster. Figure 5 shows how the encoder network maps samples of the different product categories onto the latent space. As for the MNIST dataset, the encoder fills the latent space densely and is capable of separating several categories. Problems arise in particular in between the categories ”Pullover”, ”Coat”, ”Shirt” — categories that are sometimes even for the human eye hard to differentiate. Figure 6 shows some Fashion MNIST images plotted in the latent space. In this plot one can clearly observe that on top of sorting for categories, the encoder

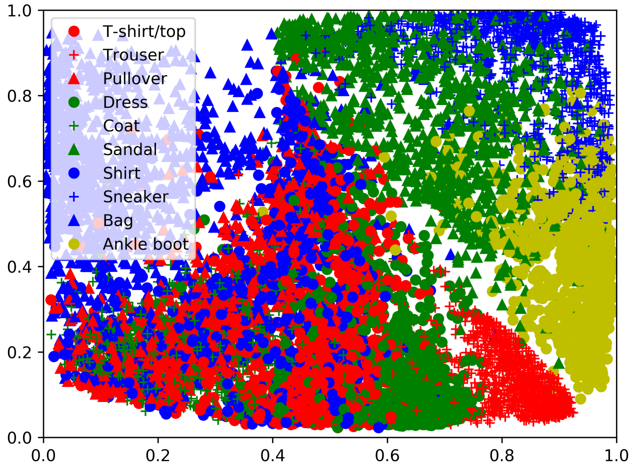

Figure 5: Test labels of the Fashion MNIST set plotted in the two dimensional latent space. A plot of the samples themselves in latent space can be found in Figure 6

also sorts the images after brightness (brighter samples are in the center whereas darker samples are located farther outside).

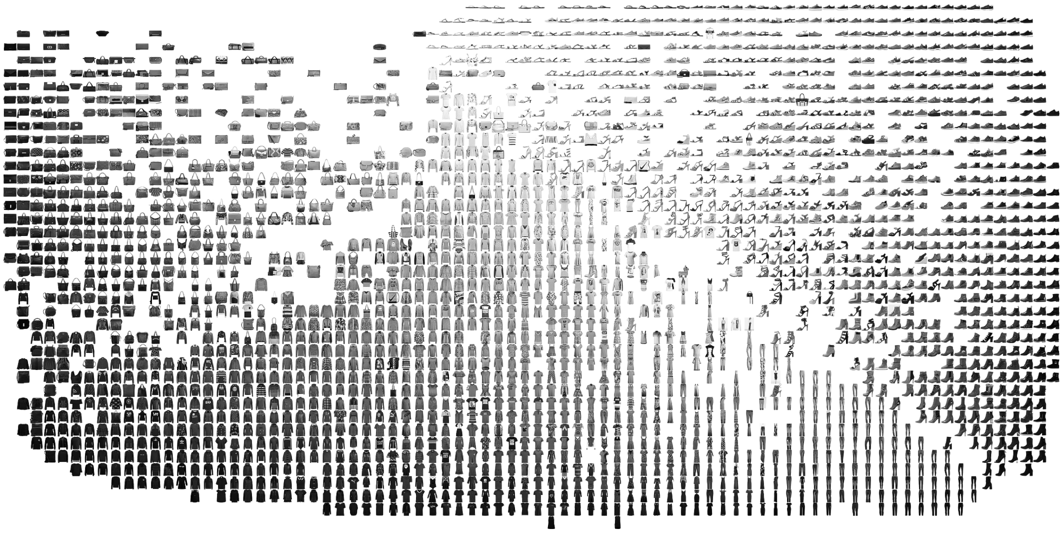

Figure 6: Test samples of the Fashion MNIST set plotted in the two dimensional latent space. Best viewed in electronic form

4 Conclusion

In this work, we motivated a new encoder discriminator architecture by taking inspiration from the game ”Taboo” and using the concept of information entropy. First promising experiments on the MNIST and Fashion MNIST dataset show that this approach is capable of clustering samples of the same class. We thus conclude that this encoder discriminator networks can learn unsupervised meaningful abstractions that might be helpful in tasks like image classification or scene understanding.

In the future this architecture could be applied to larger datasets such as ImageNet and a higher dimensional latent space in order to perform for example unsupervised transfer learning (Pan et al. (2010)) for classification tasks. Another application could be to use the output of the encoder network as the source of latent noise for generator networks in GANs. By passing also this latent description to the discriminator, the problem of mode collapse could be addressed.

REFERENCES

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103. ACM, 2008.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.