Relatório — Variational Autoencoders

Leonardo Loureiro Costa
leonardo.costa@unifesp.br

***Resumo*—O objetivo desse relatório foi explorar o efeito do tamanho do espaço latente em autoencoders variacionais no dataset MNIST. Cinco redes foram treinadas com espaços latentes de 2, 4, 8, 16 e 32 dimensões, respectivamente.**

Os resultados mostraram que quanto maior o espaço latente, dada uma mesma topologia e mesmas condições de treino, melhor a qualidade da reconstrução das imagens.

Além disso, foi observado agrupamento de dígitos da mesma classe no espaço latente e regularidade nas distribuições, sugerindo que o autoencoder aprendeu a codificar eficientemente as características distintivas de cada classe e que imagens morfologicamente próximas têm coordenadas próximas no espaço latente.

I. Introdução

A. Autoencoders e VAE

Autoencoders são um tipo de rede neural que visa aprender uma representação compacta de um conjunto de dados. Eles são compostos por duas partes: um Encoder que codifica os dados de entrada em uma representação compacta, e um Decoder que tenta reconstruir os dados de entrada a partir desta representação.

Os autoencoders variacionais são uma variante desta técnica que adiciona uma camada adicional de incerteza na representação compacta dos dados, através da utilização de distribuições probabilísticas em vez de valores determinísticos. Isso permite que eles capturem mais facilmente padrões complexos nos dados e tenham uma maior capacidade de generalização.

Este relatório estudará o uso de Autoencoders variacionais no dataset MNIST, analisando o que acontece quando o espaço latente muda de dimensão.

B. Base de dados MNIST

O conjunto MNIST — Modified National Institute of Standards and Technology — é uma grande base de dados de dígitos manuscritos comumente usada para treinar diversos sistemas de processamento de imagens.

A base de dados também é amplamente utilizada para treinamento e teste no campo da aprendizagem de máquina; possui 70 mil exemplares de imagens 28 por 28 píxeis, de fundo preto com algarismos de 0 a 9 em branco.

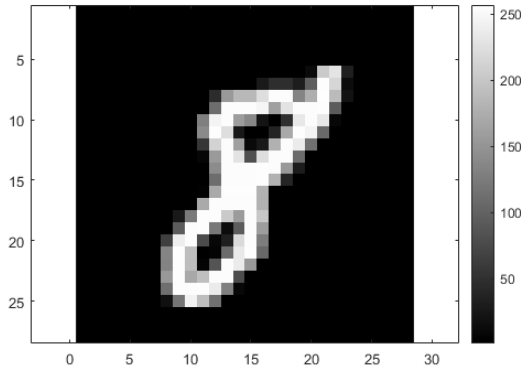

Figura 1: Digito 8, obtido a partir da base de dados MNIST

II. Metodologia

A. Formatação dos dados

Primeiro o conjunto de dados será formatado e seccionado em treino, e teste. As imagens de 28x28, elementos do conjuntos MNIST, serão transformadas em um vetor único, de dimensão (1, 784) e, após, serão normalizadas, cada elemento terá seu valor dividido por 255.

Os vetores normalizados são divididos em 2 partes, $\frac{6}{7}$ para treino e $\frac{1}{7}$ para teste; 60 mil e 10 mil, respectivamente.

B. Modelo de rede neural

Os vetores normalizados alimentam a rede neural descrita na figura 2. O vetor é representado por X,

a rede neural almeja reduzir a dimensionalidade de X em uma representação mínima e, a partir disso, reconstruir $\hat{X}$ de forma que $X \approx \hat{X}$.

O Encoder e o Decoder são constituídos apenas de 1 camada densa cada, totalmente conectada, de 512 neurônios de ativação "ReLu".

As camadas μ_x e σ_x representam a média e o logaritmo do desvio padrão, respectivamente; são conectadas à camada anterior, de 512 neurônios, e é a partir delas que o algoritmo de "sample"gera o vetor de espaço latente.

Esse algoritmo calcula uma amostra da distribuição probabilística da representação latente. Isso é feito multiplicando um tensor gerado aleatoriamente — de média 0, desvio padrão 0.1 — por ${\sigma_x}^2$ e somando o resultado ao vetor μ_x.

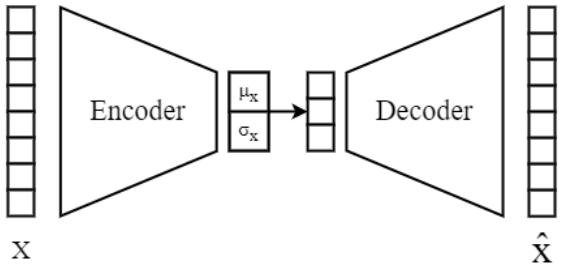

Figura 2

A função de custo utilizada é composta por duas partes: a perda de reconstrução e a perda de divergência de Kullback-Leibler (KL).

A perda de reconstrução mede o quão bem o Decoder consegue reconstruir os dados de entrada a partir da representação compacta gerada pelo Encoder.

A perda de KL mede a distância entre a distribuição probabilística da representação latente gerada pelo Encoder e uma distribuição normal padrão.

Ela é usada para garantir que a representação latente seja suficientemente variável. A função de custo final é a média da soma dessas duas perdas.

C. Treinamento

Cinco Redes serão treinadas, todas com a mesma topologia, porém com tamanhos diferentes de espaço latente — 2, 4, 8, 16, e 32, respectivamente.

O treinamento consiste em 10 épocas no conjunto de treino com "batch size"32. O otimizador usado em todos os treinos é o Adam.

III. Objetivos

- Explorar o efeito de diferentes espaços latentes na reconstrução de novas imagens.
- Visualizar os diversos espaços latentes em apenas 2 dimensões utilizando T-SNE

IV. Resultados

A figura 3 exibe 6 imagens, da esquerda pra direita temos: a imagem original, e as 5 imagens reconstruídas a partir das diferentes redes treinadas, com espaços latentes de tamanhos 2, 4, 8, 16 e 32, respectivamente.

Figura 3: Algarismo dois reconstruído por Autoencoders de espaços latentes de diferenes dimensões.

Observa-se que quanto maior o espaço latente, mais próximo é X de $\hat{X}$, isto é: mais a imagem reconstruída se aproxima da imagem original.

A figura 4 Ilustra em um gráfico 2D pontos referentes as coordenadas de índice 0 e 1 do espaço latente da primeira rede treinada.

Observa-se que há certa clusterização de dígitos da mesma classe.

Nota-se também que há regularidade nas distribuições; há continuidade, pois pontos próximos no plano cartesiano representam imagens morfologicamente próximas em suas reconstruções. Pontos distantes, como o cluster 0 é do cluster 1 e do cluster 7 indicam, analogamente, pouca proximidade morfológica entre figuras reconstruídas.

Essa regularidade evidencida-se na figura 5, onde os pontos (-2, -2) e (2, 2), do espaço latente do primeiro modelo foram interpolados de forma a criar uma grade 9 por 9 de imagens reconstruídas a partir de um espaço latente não presente nos conjuntos de treino e validação, formando um "gradiente"entre números diferentes.

Observa-se a presença de todos os algarismos, alguns mais que outros, nota-se também como um mescla-se com outro.

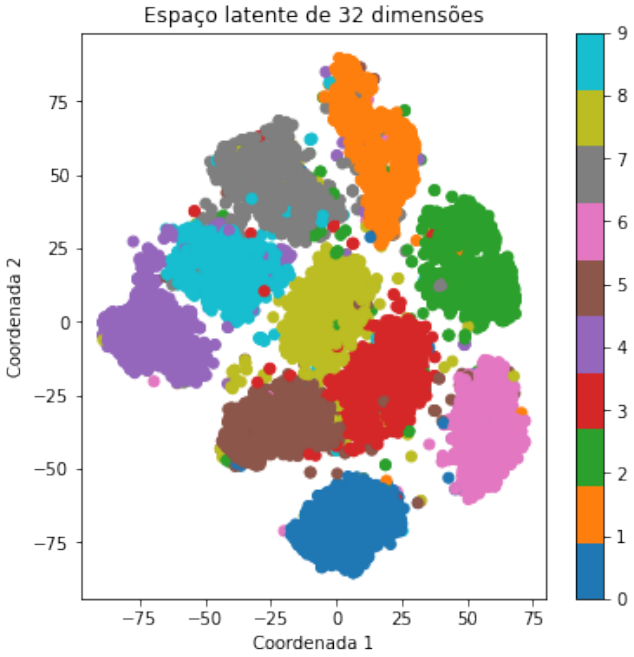

Figura 4

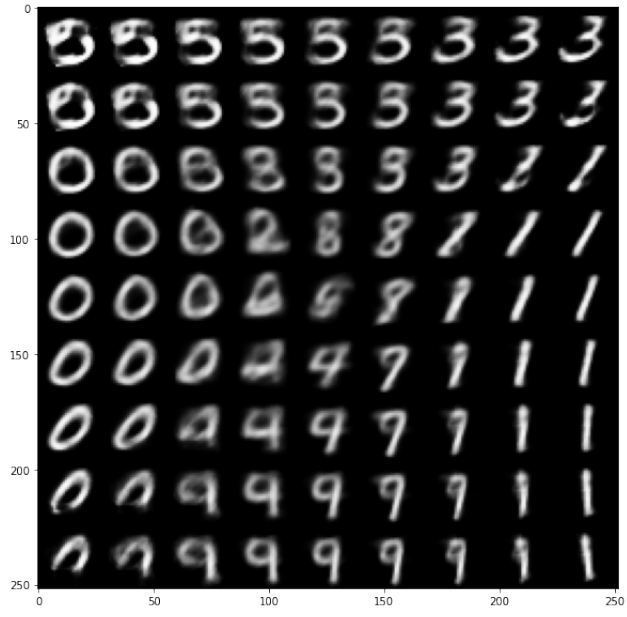

Figura 5

V. Conclusão

Com base nos resultados apresentados, conclui-se que o tamanho do espaço latente é um fator determinante na precisão da reconstrução de imagens pelo autoencoder.

Quanto maior o espaço latente, melhor a qualidade da reconstrução, pois há uma maior capacidade de armazenamento de informação.

Além disso, observou-se que há uma certo agrupamento de dígitos da mesma classe no espaço latente, o que sugere que o autoencoder aprendeu a codificar eficientemente as características distintivas de cada classe.

Também foi observada regularidade nas distribuições dos pontos no espaço latente, o que indica que imagens morfologicamente próximas têm coordenadas próximas no espaço latente.

Referências

[1] Digitos mnist, 2022.

[2] Francois Chollet. Building autoencoders in keras. https://blog.keras.io/building-autoencoders-in-keras.html, 2015. Accessed: 2022-12-30.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[4] Carl Doersch. Understanding variational auto-encoders (vaes). https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73, 2016. Accessed: 2022-12-30.

[5] Simon S. Haykin. *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition, 2009.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.