ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Projeto 2

Ambiente Virtuais Interativos e Inteligentes

Trabalho realizador por:

Duarte Valente | A47657

João Valido | A51090

Docente:

Eng. Arnaldo Abrantes

Curso: MEIM

2023

Índice

1. Introdução 4

a) Descrição 4

b) Âmbito 4

c) Documentação 4

A. Implementação 5

1. Implementação Geral 5

2. Implementação Detalhada 6

i. Modelo BasicUNet 6

ii. Modelo UNet2DModel 7

iii. Modelo UNet sem skip connections 8

iv. Resultados e Comparações 8

v. Previsões dos modelos 9

vi. Amostragem 10

3. Conclusão de comparações 11

B. Exploração de ferramentas Stable Diffusion 12

1. Scribble Diffusion 12

2. DreamStudio 13

3. Stable Diffusion 15

4. Playground AI 16

5. Comentário 17

Índice de ilustrações

Figura 1 - Arquitetura do modelo 6
Figura 2 - Resultados Obtidos 8
Figura 3 - Previsão Unet 9
Figura 4 - Previsão Unet sem skip connection 9
Figura 5 - Previsão Unet2DModel 9
Figura 6 - Amostragem Unet 10
Figura 7- Amostragem Unet2DModel 10
Figura 8- Amostragem Unet sem skip connections 11
Figura 9 - Scribble Imagem Gerada 1 12
Figura 10 - Scribble Imagem Gerada 2 12
Figura 11 - DreamStudio Imagens Gerada 13
Figura 12 - DreamStudio Opções de prompt 14
Figura 13 - Stable Diffusion Imagens Geradas 15
Figura 14 - Playgound AI Imagens Geradas 16

1.Introdução

a) Descrição

Este trabalho tinha como objetivo, tendo como base o *notebook* Diffusion Models from Scratch, apresentado em aula, treinar uma rede neural para remover ruído de imagens. Isso será feito através da definição do conjunto de imagens usadas para treino e adição de ruído a essas imagens. Em seguida, será treinada uma rede neural capaz de remover o ruído adicionado, e o modelo treinado será usado para produzir novas imagens

b) Âmbito

Este relatório insere-se no âmbito da realização da ficha laboratorial unidade curricular de Ambientes Virtuais Interativos e Inteligentes, do Mestrado em Engenharia Informática e Multimédia do DEETC do ISEL.

c) Documentação

Documentação de apoio à unidade curricular de Ambientes Virtuais Interativos e Inteligentes.

Notebook Diffusion Models from Scratch.

A. Implementação

1. Implementação Geral

O notebook "Diffusion Models from Scratch" contém uma implementação em Python de um modelo de difusão de imagem, que é uma técnica usada para remover ruído de imagens. A implementação é baseada em uma rede neural que é treinada para modelar o processo de difusão de uma imagem.

O primeiro passo é definir o conjunto de imagens que serão usadas para o treino da rede neural. Foi então escolhido o conjunto de dados de imagens de carros, retirado do dataset ‘cifar10’. Esse conjunto de dados contém cerca de 5000 imagens de 32x32 a cores de carros, que podem ser usadas para treinar a rede neural.

O próximo passo é adicionar ruído às imagens. Isso pode ser feito de várias maneiras, mas seguindo o notebook, o ruído será adicionado como uma perturbação gaussiana aleatória. O grau de ruído adicionado pode ser ajustado para controlar a quantidade de ruído presente nas imagens.

Depois que o ruído é adicionado às imagens, a rede neural pode ser treinada para remover o ruído. Isso é feito usando o modelo de difusão de imagem implementado no notebook. A rede neural é treinada para modelar o processo de difusão de uma imagem com ruído, a fim de remover o ruído e restaurar a imagem original.

Uma vez que o modelo é treinado, ele pode ser usado para produzir novas imagens. Isso é feito através da amostragem de ruído aleatório e aplicação do modelo de difusão de imagem treinado. O resultado é uma imagem gerada que é uma versão restaurada da imagem original com ruído adicionado.

2. Implementação Detalhada

i. Modelo BasicUNet

A arquitetura UNet é composta por um caminho de "compressão", no qual os dados são reduzidos, e um caminho de "expansão", no qual eles são expandidos de volta à dimensão original. No entanto, o UNet também possui conexões (skip connections) entre as camadas não adjacentes que permitem a troca de informações e gradientes entre elas.

Um modelo UNet, pode ser bastante complexo e elaborado. No entanto, para este trabalho, foi utilizado o modelo sugerido e estudado em aula. Modelo esse, que receberá uma imagem com três canais e os processará através de três camadas convulsionais no caminho descendente e outras três no caminho ascendente. As conexões entre camadas não adjacentes mencionadas anteriormente conectam assim as camadas descendentes e ascendentes.

Diferentemente de UNets mais complexas, é utilizado o *max pooling* para "compressão" dos dados (downsampling) e *nn.Upsample* para a "expansão" dos dados (upsampling). Essa abordagem permite assim preservar as informações essenciais durante o processo.

Assim, a figura que se segue representa um exemplo da arquitetura do modelo que conta com três canais de entrada e três canais de saída, um para cada componente RGB da imagem:

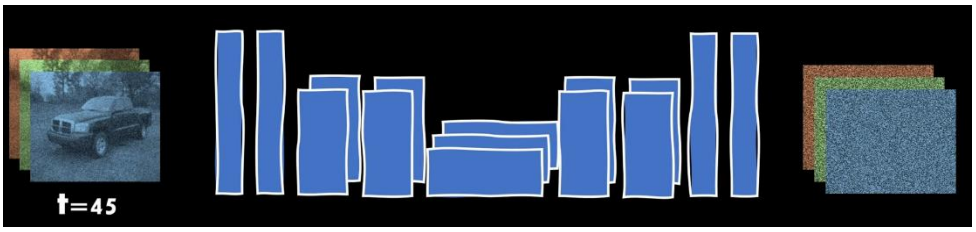


Figura 1 - Arquitetura do modelo

Para treinarmos o modelo, fornecemos como entrada a imagem corrompida e o modelo deve fornecer a sua melhor estimativa de como seria a sum aparência original. Em seguida, comparamos essa estimativa com a imagem original antes de ser corrompida utilizando o erro médio quadrático e ajustamos os parâmetros do modelo de forma a o melhorar.

Ordem do processo de treino:

1) Obtenha os dados para treino, no nosso caso um batch de 128 imagens.
2) É adicionado ruido aleatório a essas mesmas imagens.
3) São fornecidas as imagens com ruido ao modelo.
4) São comparadas as previsões retornadas pelo modelo com as imagens originais antes de ser adicionado o ruido.
5) Por fim, são ajustados os parâmetros do modelo.

ii. Modelo UNet2DModel

Este modelo é mais avançado em comparação com o BasicUNet, pois o processo de corrupção das imagens é feito de forma diferente. O objetivo do treino é diferente, envolvendo a previsão do ruído em vez da imagem sem ele. E é tambem condicionado com base na quantidade de ruído presente através do condicionamento do *timestep*.

Assim, podemos afirmar que o UNet2D é uma versão melhorada do modelo UNet pois conta com melhorias como:

- São feitas normalização em grupo às entradas de cada bloco.
- São utilizadas camadas de *Dropout (*define aleatoriamente as unidades de entrada como 0 com uma certa frequência em cada etapa durante o tempo de treino, o que ajuda a evitar o overfitting*)* para um treino mais suave.
- São incorporadas múltiplas camadas *Resnet (*arquitetura (CNN) projetada para suportar centenas ou milhares de camadas convulsionais*)* por bloco.
- A atenção é aplicada (geralmente apenas em blocos de resolução inferior).
- O modelo é condicionado ao timestep.
- São utilizados blocos de downsampling e upsampling com parâmetros possíveis de serem aprendidos pelo modelo tambem.

iii. Modelo UNet sem skip connections

As conexões (skip connections) desempenham um papel crucial na preservação e propagação de informações em diferentes níveis ou camadas da rede como a nível de:

Fluxo de informação e localização: Estas conexões permitem o fluxo de informações entre os caminhos de "compressão" e “expansão”. Ao fazê-lo, as conexões possibilitam que a rede localize recursos de forma eficaz.

Solução para problemas de estrangulamento: O caminho de "compressão" da arquitetura UNet utiliza camadas convolucionais e de pooling para reduzir as dimensões espaciais da entrada. Embora esse processo de redução ajude a capturar informações semânticas de alto nível, também leva à perda de detalhes espaciais. As conexões de salto ajudam a resolver esse problema de estrangulamento, fornecendo atalhos para o fluxo dos gradientes na direção oposta. Dessa forma, a rede pode recuperar detalhes espaciais e manter uma melhor compreensão das estruturas locais durante o caminho de expansão.

Fluxo de gradientes e estabilidade de treino: Estas conexões ajudam tambem a aliviar problemas com os gradientes durante o treino. Ao fornecer conexões diretas entre as camadas de "compressão" e “expansão”, os gradientes podem fluir facilmente entre diferentes níveis da rede. Isso possibilita um treino estável e eficiente, permitindo que o modelo aprenda e otimize os parâmetros de forma eficaz.

Resumindo, as conexões (skip connections) no modelo UNet facilitam o fluxo de informações, auxiliam na localização de recursos, resolvem problemas de estrangulamento, promovem a reutilização de recursos e melhoram o fluxo de gradientes e a estabilidade do treino. Estas conexões permitem que o UNet obtenha resultados de segmentação precisos e detalhados.

iv. Resultados e Comparações

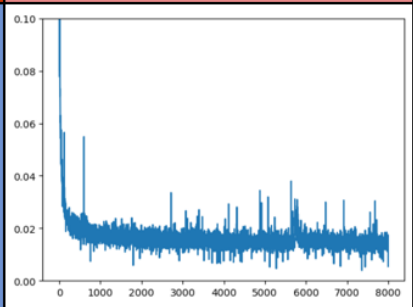
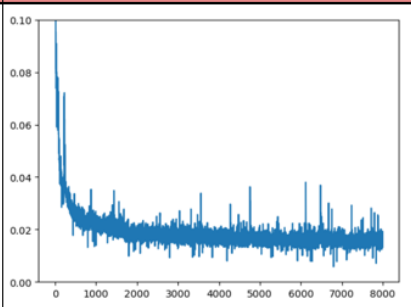
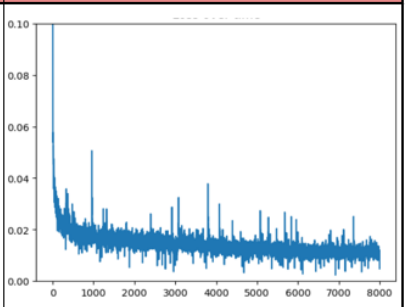
Modelo	U-Net	U-Net sem skip conections	UNet2DModel
Gráfico de evolução			
Loss / 200 epocas	0.014049	0.015549	0.010834

Figura 2 - Resultados Obtidos

v. Previsões dos modelos:

Unet

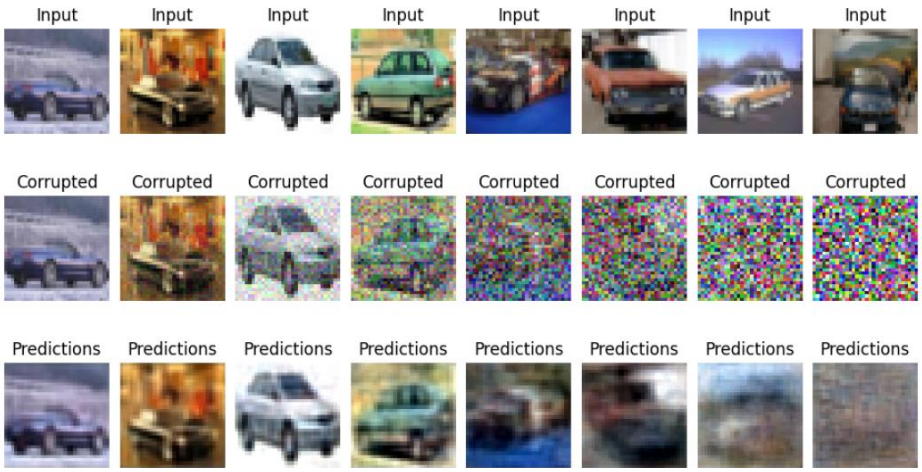


Figura 3 - Previsão Unet

Unet2DModel

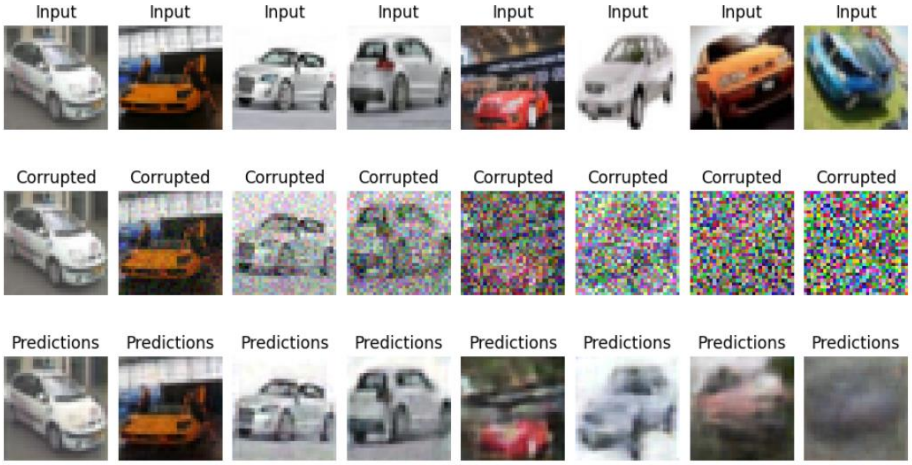


Figura 5 - Previsão Unet2DModel

Unet sem skip connection

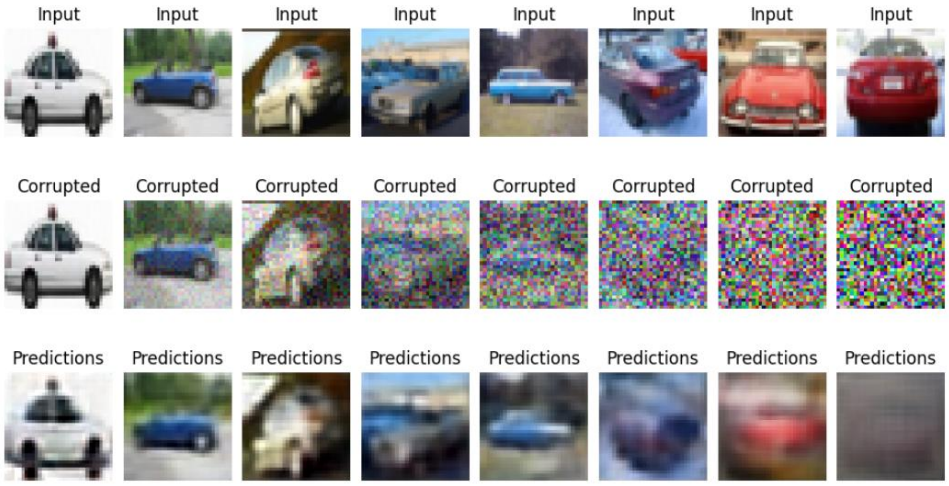


Figura 4 - Previsão Unet sem skip connection

vi. Amostragem

Para gerar novas imagens, começamos por gerar um ruído aleatório, e analisamos as previsões do modelo, mas só avançando uma pequena distância em direção a essa previsão. Assim, teremos uma imagem com muito ruido, onde talvez se consiga encontrar alguma estrutura. Estrutura essa que podemos utilizar no modelo para obter uma nova previsão. A esperança é que essa nova previsão seja ligeiramente melhor do que a primeira, permitindo-nos continuar o processo com essa nova imagem melhorada.

Ao repetir este processo algumas vezes e conseguimos obter assim obter uma amostra de imagem.

Algumas amostras dos modelos:

Unet

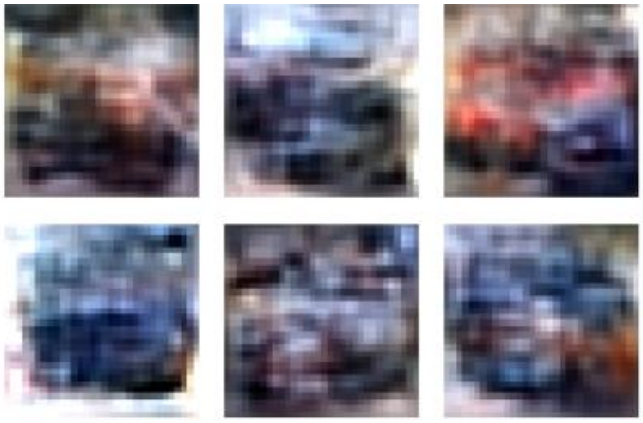


Figura 6 - Amostragem Unet

Unet2DModel

Figura 7- Amostragem Unet2DModel

Unet sem skip connections

Figura 8- Amostragem Unet sem skip connections

3. Conclusão de comparações

Claramente que o modelo que mais se destacou foi o UNet2D como previsto, pois o loss calculado do modelo é menor que do modelo BasicUNet, as imagens resultantes da remoção do ruido pelo modelo tambem são mais semelhantes às do modelo BasicUNet e as imagens geradas aleatoriamente tambem se parecem mais semelhantes com automóveis.

Em relação ao skip connections, reparamos que a diferença existiu, no entanto não foi tão notória como entre os modelos BasicUNet e UNet2D.

No geral, apesar dos resultados obtidos não serem perfeitos, ficamos a perceber o funcionamento destes modelos e como as novas imagens estavam a ser geradas e tinham parecenças com automóveis. O que prova que com um maior poder de processamento e talvez umas melhorias nos modelos poderíamos perfeitamente gerar imagens únicas de automóveis.

B. Exploração de ferramentas Stable Diffusion

1. Scribble Diffusion

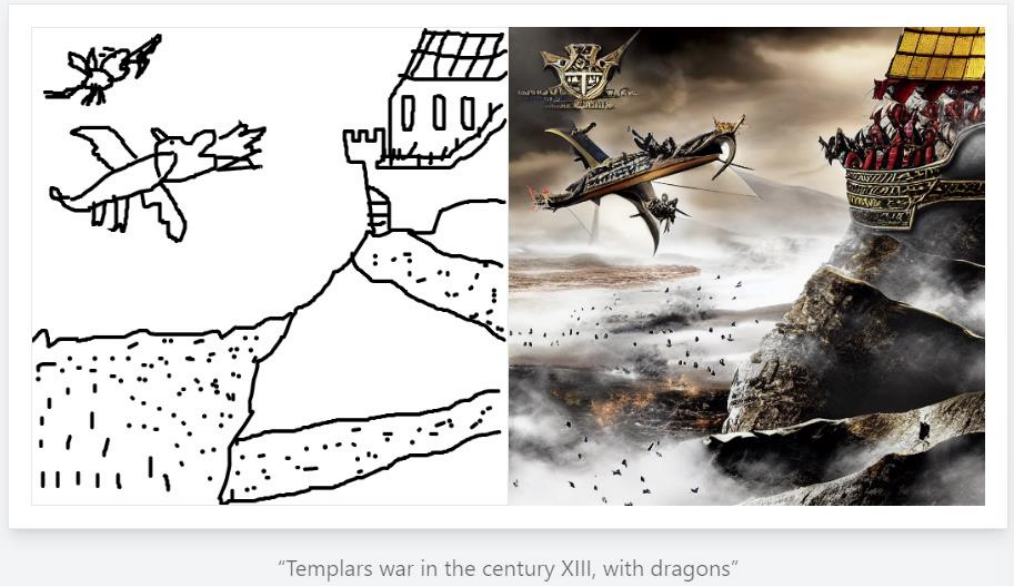


Figura 9 - Scribble Imagem Gerada 1

Figura 10 - Scribble Imagem Gerada 2

2. DreamStudio

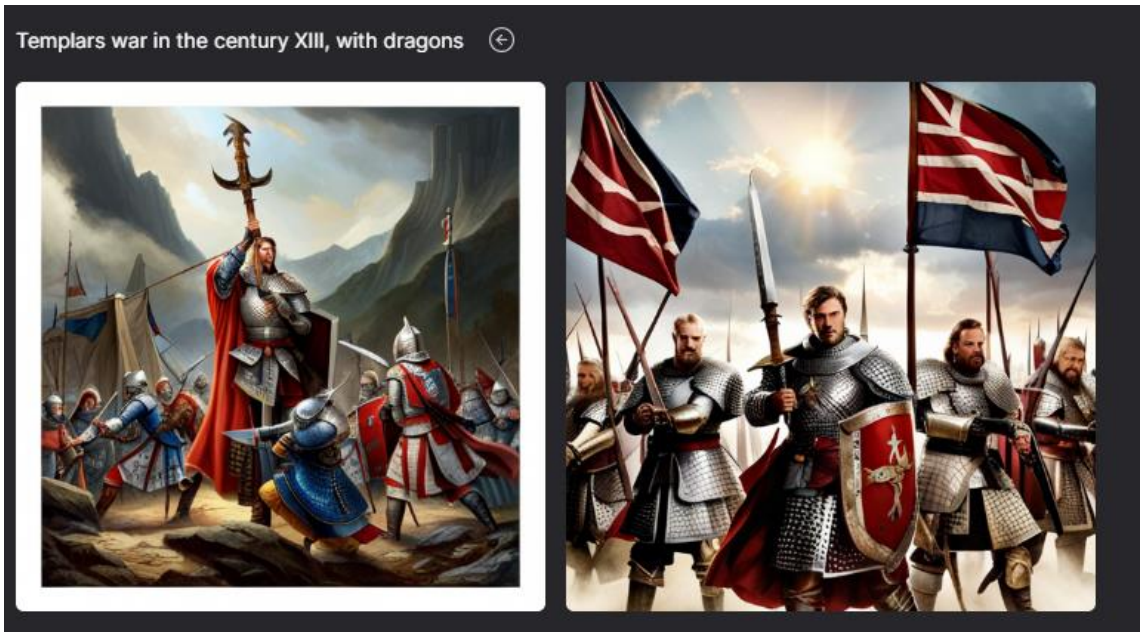


Figura 11 - DreamStudio Imagens Gerada

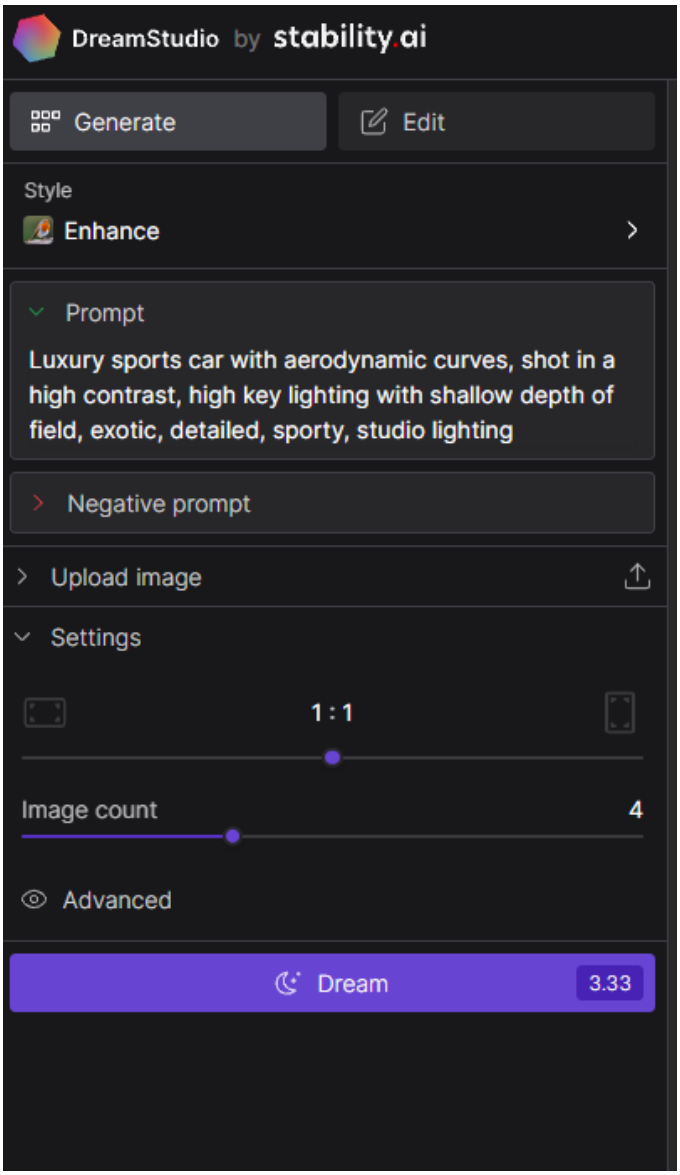


Figura 12 - DreamStudio Opções de prompt

3. Stable Diffusion

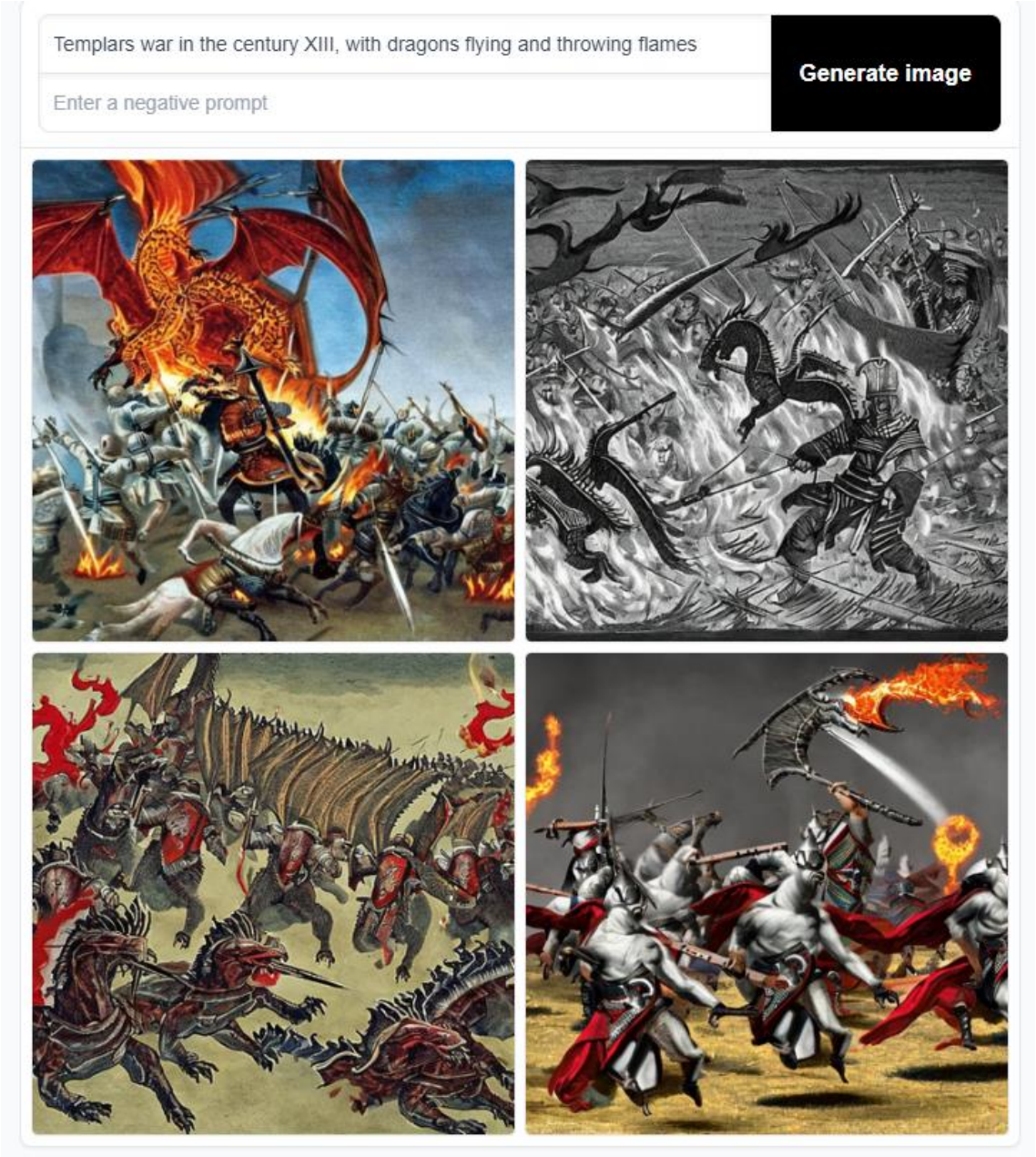


Figura 13 - Stable Diffusion Imagens Geradas

4. Playground AI

Figura 14 - Playgound AI Imagens Geradas

5. Comentário

Os sites Stable Diffusion Online, DreamStudio, Playground e Scribble Diffusion podem ter objetivos e funcionalidades diferentes, mas todos compartilham o objetivo geral de fornecer uma experiência de usuário agradável e funcional. Stable Diffusion Online pode ser útil para designers e desenvolvedores de sites que desejam criar e testar designs e layouts, enquanto DreamStudio pode ser uma plataforma mais ampla para pessoas criativas que desejam explorar ferramentas de design e criar projetos multimedia. Playground, já é mais uma biblioteca de imagens geradas através de inteligência artificial que através da barra de pesquisa permite encontrar uma imagem o mais próxima possível da descrição introduzida. Enquanto Scribble Diffusion pode ser útil para artistas digitais que desejam criar e compartilhar desenhos e ilustrações. Em termos de capacidades e limitações, cada um desses sites pode ter pontos fortes e fracos, dependendo das necessidades do usuário. Por exemplo, Stable Diffusion Online pode ter recursos avançados de design, mas pode não ser adequado para usuários iniciantes, como por exemplo a introdução da prompt negativa. DreamStudio pode ter uma ampla gama de ferramentas, mas pode ser mais complexo e exigir mais conhecimento técnico, tem ainda um limite de imagens possíveis de gerar de forma gratuita. Scribble Diffusion pode ter recursos avançados de desenho, mas pode ser menos adequado para outras formas de criação, tendo em conta que a imagem é gerada tendo como entrada obrigatória um rascunho elaborado pelo utilizador. Em resumo, a escolha do site depende das necessidades e habilidades do utilizador. É importante considerar as capacidades e limitações de cada plataforma para escolher a melhor opção para o seu projeto ou objetivo específico. Tendo em conta a nossa experiência o Scribble Diffusion é o que melhor deixa expressar a criatividade do utilizador, uma vez que a maior parte da imagem é gerada tendo em conta o rascunho efetuado, o que por outro lado pode não ser o mais prático, pois depende das habilidades de desenho do utilizador.