Guia Para Criação de Dataset de Imagens Para modelos de Aprendizagem Profunda

Editado por: Dr. Arnaldo de Carvalho Junior em Agosto 09, 2024.

Uma das etapas mais importantes para se criar um modelo de aprendizagem profunda (*deep learning* - DL) é criar um *dataset* (conjunto de dados) de imagens em grande escala para treinar o modelo.

Como coletar, rotular e armazenar imagens de maneira eficiente e eficaz?

Neste guia abrangente, será abordado o passo a passo do processo de criação do *dataset* de imagem, incluindo desde o planejamento e preparação até os detalhes de anotação e organização.

1. Compreendendo As Necessidades do *Dataset* de Imagem

Quando se trata de construir um d*ataset* de imagem, o primeiro passo é entender o propósito que o *dataset* servirá. Se o propósito é para pesquisa, um modelo de inteligência artificial (IA) ou DL, ou diversão, saber o propósito ajudará a determinar que tipo de dados são necessários e quanto deles.

Por exemplo, se o objetivo do *dataset* de imagem for um projeto de pesquisa, os dados devem ser de alta qualidade e ter um nível específico de detalhe que auxiliará no estudo.

Por outro lado, se o propósito é apenas para diversão, então o *dataset* pode ser mais relaxado, variado e aberto. Se um modelo de IA ou DL usar o *dataset*, ele deverá ser ***estruturado*** para facilitar o aprendizado e o ***reconhecimento de padrões*** nas imagens pelo sistema. Portanto, a clareza sobre o propósito das imagens é crucial na construção de um *dataset* bem-sucedido e eficaz.

Ao construir um *dataset* de imagem, uma das primeiras coisas a considerar é o tamanho e a complexidade dos dados. Como tal, é importante determinar quantas imagens são necessárias e quais tipos de imagens atendem às necessidades.

É importante lembrar que o tipo de imagens no *dataset* impactará significativamente o quão ***preciso*** e ***eficaz*** será qualquer modelo de aprendizado de máquina (*machine learning* – ML) que se treinar nele.

Portanto, gastar tempo pesquisando e planejando o tamanho e a complexidade ideais do *dataset* de imagens é crucial para obter resultados confiáveis.

Em seguida, ao fazer um *dataset* de imagem, é preciso considerar os requisitos de qualidade que as imagens precisam atender para fornecer resultados precisos.

A resolução das imagens é uma grande parte se descobrir o quão bom é o *dataset*. Podem ser necessárias imagens de alta ou baixa resolução dependendo do foco da pesquisa ou projeto. Além disso, incluir recursos específicos, como cor, orientação e outras características visuais, é fundamental.

Esses recursos ajudam a criar um *dataset* mais abrangente e detalhado que pode ser utilizado para diversos fins. O desenvolvimento de um *dataset* de imagem requer uma consideração cuidadosa de seus requisitos e recursos de qualidade, o que garante que ele atenda ao propósito pretendido.

2. Preparando Recursos Para a Escalada

Como a importância do *dataset* de imagens continua a crescer em diferentes áreas de pesquisa, é essencial preparar recursos para dimensionar esses conjuntos de dados. Isso envolve entender as várias ferramentas e técnicas disponíveis para gerenciar e armazenar grandes quantidades de dados.

a) Gerenciamento: Os conjuntos de dados de imagens requerem gerenciamento adequado para garantir seu uso prático, seja para visão computacional, ML ou outras aplicações.

b) Preparação: Gerenciar e preparar dados de imagem requer uma consideração cuidadosa de fatores como ***anotação***, ***rotulagem*** e ***limpeza de dados***.

c) Armazenamento: Um aspecto crucial do processo de gerenciamento é selecionar os métodos de armazenamento mais adequados, como sistemas de arquivos distribuídos ou armazenamento em nuvem, que possam lidar com eficiência com o volume e a complexidade dos dados. O armazenamento em nuvem não apenas torna os dados mais fáceis de encontrar e acessar, mas também torna os dados mais seguros e confiáveis.

d) Segurança: Com o armazenamento em nuvem, os dados são criptografados e protegidos contra acesso não autorizado. Isso reduz o risco de violações de dados.

e) Ferramentas de Otimização: A escala de *dataset* de imagens requer uma compreensão abrangente das ferramentas e técnicas necessárias para otimizar o armazenamento, processamento e preparação de dados para análise.

As plataformas de armazenamento em nuvem oferecem soluções que podem ser ampliadas à medida que o armazenamento de dados cresce. Esforços colaborativos entre pesquisadores e partes interessadas também são possíveis, pois o armazenamento em nuvem permite fácil compartilhamento e colaboração em um ambiente seguro. À medida que o volume de conjuntos de dados de imagem aumenta, aproveitar o armazenamento em nuvem continua sendo crucial para gerenciar e armazenar arquivos grandes, mantendo a segurança e a acessibilidade dos dados.

3. Estratégias de Aquisição de Imagens

A aquisição de imagens de alta qualidade é um passo essencial na construção de um *dataset* de imagens. Requer consideração cuidadosa de vários fatores, como fonte, formato e tamanho das imagens.

a) Fonte: A origem das imagens deve ser confiável para garantir autenticidade e credibilidade. Além disso, o formato deve ser compatível com o *dataset* e as ferramentas que serão utilizadas para análise.

b) Tamanho das Imagens: é outro fator crítico que deve ser considerado. Imagens grandes podem consumir recursos significativos de memória, enquanto imagens pequenas podem precisar fornecer mais detalhes para análise. Portanto, é necessário equilibrar a qualidade da imagem e o tamanho do arquivo. A seleção de imagens deve ser abrangente e representativa para garantir que o *dataset* represente com precisão a população ou os fenômenos pretendidos.

c) Aquisição de imagens: é um processo crítico que requer deliberação cuidadosa e atenção aos detalhes para garantir que o *dataset* seja preciso e confiável. O processo de aquisição deve incluir de forma abrangente as variações de ângulos, condições de iluminação e objetos.

Por fim, criar um *dataset* de imagens é crucial em muitos campos, como visão computacional e ML. Isso pode ser conseguido através da aquisição manual ou automatizada de imagens de diversas fontes, como bancos de dados *on-line*, repositórios, pesquisas estruturadas na web ou configurações personalizadas de câmeras.

É importante notar que a criação de um *dataset* de imagem não consiste apenas em coletar, mas também em garantir que eles sejam de alta qualidade, diversificados e relevantes para a tarefa. Depois que o *dataset* de imagem é criado, ele pode ser usado para diversas aplicações, como reconhecimento de objetos, classificação de imagens e análise de cena. Concluindo, a criação de um *dataset* é fundamental para o avanço do campo da visão computacional e do ML.

4. Pré-processamento e Preparação de Imagens

Ao se trabalhar em um projeto com um conjunto de dados de imagens, é essencial configurar uma maneira confiável de lidar com os dados.

O pré-processamento e a preparação são etapas integrais para atingir esse objetivo. Essas etapas envolvem a ***manipulação*** e ***otimização*** das imagens para garantir que eles estejam em um formato que seja propício aos requisitos do projeto.

Durante a fase de pré-processamento, tarefas como **filtragem**, **redimensionamento** e **normalização** geralmente são feitas para garantir que todas tenham a mesma qualidade e formato. Esta etapa garante que os dados estejam prontos para análise, permitindo que os pesquisadores extraiam ***insights*** dos dados de forma mais eficaz.

O pré-processamento faz também com que os ***modelos de ML*** mais precisos, o que os torna melhores em tarefas como reconhecimento. Um bom pré-processamento é uma das etapas mais críticas para garantir que um projeto de conjunto de dados atenda aos seus objetivos e atenda a altos padrões de qualidade e precisão.

Além disso, é imperativo observar que o manuseio adequado de um *dataset* é crucial para seu uso bem-sucedido em aplicativos de ML e visão computacional. O ***redimensionamento*** e ***recorte*** durante a fase de preparação devem ser realizados com muito cuidado para garantir que os dados representem com precisão os objetos e cenas dos quais foram inicialmente retirados.

É importante ao usar o *dataset* para tarefas como detecção ou classificação de objetos, observar que a integridade do sistema afeta diretamente a precisão dos algoritmos. Para fazer um *dataset* confiável e útil, deve-se prestar muita atenção a cada detalhe durante a preparação.

5. Etapas de Validação e Garantia de Qualidade

O *dataset* é crucial para muitos setores e campos de pesquisa, incluindo ML, visão computacional e IA.

a) Verificação: Para garantir dados da mais alta qualidade para processos de validação e garantia de qualidade, é crucial examinar minuciosamente e verificar a precisão de cada *dataset* antes de usar. Nesse processo, cada um é analisado, seus metadados são verificados e garante que atenda a padrões específicos. O *dataset* deve ser preciso e consistente para que os modelos de ML sejam confiáveis e robustos.

b) Testes de Qualidade e Validação: Dados de má qualidade podem levar a previsões imprecisas, o que pode ser prejudicial ao desempenho geral do sistema. Portanto, é essencial priorizar a qualidade dos dados ao trabalhar com conjuntos de dados para alcançar resultados precisos e confiáveis. Ao lidar com *dataset*, é crucial garantir que os dados dentro deles atendam aos padrões de qualidade esperados. Uma maneira de garantir isso é executando testes automatizados no conjunto de dados.

c) Identificação de Anomalias e Inconsistências: Isso ajuda a identificar quaisquer anomalias ou inconsistências nos dados, que podem então ser abordadas antes que se tornem problemas significativos. Recomenda-se também a realização de testes manuais para verificar a precisão de quaisquer detalhes adicionais incluídos no conjunto de dados, como rótulos ou anotações. Isso é especialmente importante para aplicações sensíveis ou essenciais, como imagens médicas ou carros autônomos. Aproveitar o tempo para testar e confirmar a precisão de um *dataset* pode economizar tempo e dinheiro no longo prazo e torná-lo mais confiável e confiável em geral.

d) Manutenção Periódica: É essencial verificar o conjunto de dados com frequência para garantir que sua precisão permaneça a mesma ao longo do tempo. Mudanças podem acontecer nos conjuntos de dados por vários motivos, como quando os dados são corrompidos ou surgem novas tecnologias. Testes de validação e garantia de qualidade devem ser feitos regularmente e quaisquer alterações necessárias devem ser feitas para manter resultados precisos. Além disso, recomenda-se estabelecer um protocolo para manutenção e atualização regular de *dataset.* Essa prática garante que os usuários do *dataset* tenham acesso a dados confiáveis e atualizados.

Seguindo um processo de manutenção bem planejado, as organizações podem usar seus *datasets* como uma ferramenta confiável para obter *insights* e tomar decisões inteligentes.

Em última análise, o sucesso de qualquer *dataset* depende do compromisso com a sua melhoria e manutenção contínuas.

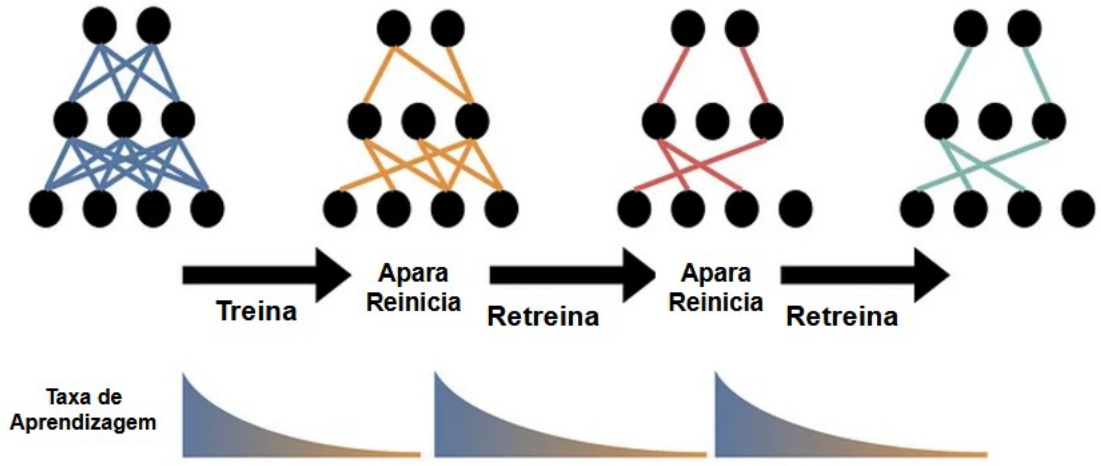
6. Pós-processamento e Finalização do *Dataset*

Um *dataset* é uma coleção usada para diversos fins, incluindo treinamento de algoritmos de ML ou realização de pesquisas.

O pós-processamento do *dataset* é a etapa final e crucial na preparação do conjunto de dados para uso. Envolve um conjunto de etapas de garantia de qualidade que garantem a ***correção***, ***precisão*** e ***confiabilidade*** dos dados.

O pós-processamento inclui várias técnicas, como ***correção***, ***normalização de dados***, ***filtragem*** e ***segmentação***. Esses métodos garantem que sejam de alta qualidade, livres de ruído e mantenham o nível de consistência desejado. Por causa disso, o pós-processamento desempenha um papel significativo para garantir que o *dataset* esteja pronto para ser usado, confiável e útil no trabalho em direção à meta estabelecida para ele.

As tarefas de pós-processamento, como ***rotulagem***, ***categorização*** e ***criação de metadados*** podem ser tediosas e demoradas, mas são fundamentais para tornar o *dataset* valioso para aplicativos de ML e visão computacional.



O *dataset* é melhorado tanto em termos da sua qualidade como da sua fiabilidade após a eliminação do duplicado e a execução de verificações de qualidade. Os metadados fornecem informações essenciais sobre e contribuem para resultados de pesquisa mais precisos e úteis. O pós-processamento é uma etapa essencial na criação de um conjunto de dados preciso e valioso que pode ser usado em vários contextos e aplicações.

7. Conclusão

Criar um *dataset* em grande escala para os modelos de DL pode inicialmente parecer assustador. Ainda assim, se for empregada a estratégia apropriada, pode-se concluir a tarefa de forma rápida e eficiente. Se este guia for seguido, o resultado será um *dataset* que contribuirá para o desenvolvimento de modelos de ML confiáveis com mais frequência. Além disso, a ordem e a limpeza em relação a tudo e qualquer coisa associada à anotação e armazenamento estarão mantidos. Seguindo as diretrizes descritas neste guia pode-se construir um *dataset* que impulsionará significativamente os esforços de DL.

REFERÊNCIAS

Xoffshoringt, A Comprehensive Guide to Creating a Large-Scale Image Dataset for Deep Learning Models, Medium, 2024. Disponível em: <https://medium.com/@24x7offshoringt/a-comprehensive-guide-to-creating-a-large-scale-image-dataset-for-deep-learning-models-e2922a9f36b1>. Acessado em Ago 09, 2024.

24x7Ofshoring. A Comprehensive Guide to Creating a Large-Scale Image Dataset for Deep Learning Models? 2021. Disponível em: <https://24x7offshoring.com/a-comprehensive-guide-to-creating-a-large-image/?feed_id=32040&_unique_id=65dff239aa281>. Acessado em Ago 09, 2024.