Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Kaio Freitas Da Dalt

Comparação de modelos de IA para classificação de alimentos

Curitiba
2025

Kaio Freitas Da Dalt

Comparação de modelos de IA para classificação de alimentos

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Paulo Ricardo Lisboa de Almeida

Curitiba
2025

Comparação de modelos de IA para classificação de alimentos

Kaio Freitas Da Dalt[1]

[1]Aluno do programa de Especialização em Data Science Big Data*

Este trabalho compara o desempenho de três modelos de inteligência artificial (ViT-Base, Swin-Base e SigLIP2-Base) na tarefa de classificação de alimentos, utilizando o conjunto de dados Food-101. Foram avaliadas a acurácia, o custo computacional (parâmetros, FLOPs e latência de inferência) e as diferenças arquiteturais, com foco nos mecanismos de atenção e no tipo de pré-treinamento. Os resultados mostraram que todos os modelos atingiram alta acurácia (acima de 89%), com o Swin-Base apresentando o melhor equilíbrio entre desempenho e eficiência. O SigLIP2-Base obteve a maior precisão Top-1 e F1-macro, enquanto o ViT-Base destacou-se no Top-5. A análise fornece subsídios para a escolha de modelos eficientes em aplicações de visão computacional para reconhecimento de alimentos.
Palavras-chave: visão computacional, classificação de imagens, Food-101, ViT-Base, Swin-Base, SigLIP2

This study compares the performance of three artificial intelligence models (ViT-Base, Swin-Base, and SigLIP2-Base) for food classification using the Food-101 dataset. The evaluation considered accuracy, computational cost (parameters, FLOPs and inference latency), and architectural differences, focusing on attention mechanisms and pre-training strategies. Results showed that all models achieved high accuracy (above 89%), with Swin-Base providing the best balance between performance and efficiency. SigLIP2-Base achieved the highest Top-1 and macro F1 scores, while ViT-Base stood out in Top-5 accuracy. This analysis offers insights for selecting efficient models for food recognition tasks in computer vision.
Keywords: computer vision, image classification, Food-101, ViT-Base, Swin-Base, SigLIP2

1. Introdução

A classificação automática de imagens de alimentos é uma tarefa desafiadora de visão computacional, com aplicações que vão desde recomendações nutricionais até robótica de cozinha. O conjunto Food-101 é amplamente utilizado como *benchmark* nesse domínio: ele contém 101 categorias diferentes de pratos e 101.000 imagens (cada classe tem 750 imagens de treino e 250 de teste) [1]. A variedade visual e o ruído (imagens não limpas) tornam a tarefa desafiadora. Nos últimos anos, arquiteturas baseadas em *transformers* demonstraram excelente desempenho em visão computacional. Por exemplo, o Vision Transformer (ViT) provou que é possível aplicar diretamente um encoder transformer em patches de imagem [2]. Outros modelos recentes, como o Swin Transformer, introduzem mecanismos hierárquicos com atenção restrita por janelas deslizantes [4].

Neste trabalho, comparamos três modelos pré-treinados específicos para Food-101 obtidos no Hugging Face: um ViT-Base, um Swin-Base e um SigLIP2-Base. O objetivo é avaliar quantitativamente, a precisão e o custo computacional de cada arquitetura no conjunto Food-101, sem realizar re-treinamento adicional. Relatamos as métricas de acurácia (Top-1 e Top-5) e F1, bem como métricas como número de parâmetros, FLOPs e latência de inferência. Essas informações são cruciais para decidir qual modelo usar em diferentes cenários, considerando requisitos de desempenho e recursos disponíveis.

2. Revisão da Literatura

Vision Transformer (ViT) — O ViT é um modelo pioneiro que adapta o mecanismo de atenção originalmente desenvolvido para Processamento de Linguagem Natural (PLN) às tarefas de visão computacional [2]. Nesse modelo, cada imagem é dividida em blocos fixos (*patches*) de 16×16 pixels, que são linearmente projetados em vetores chamados de *embeddings*. Um token especial, chamado `[CLS]`, é adicionado à sequência de entrada para representar a infor-

*kaiodadalt@gmail.com

mação global da imagem. Posicionamentos absolutos também são incorporados para indicar a ordem dos *patches*, e toda a sequência é processada por camadas de autoatenção do Transformer. A saída do token `[CLS]` é então passada para uma camada final do tipo MLP (perceptron multicamada) para prever a classe da imagem. O ViT elimina a necessidade de convoluções tradicionais e permite capturar dependências de longo alcance desde as primeiras camadas. No entanto, o modelo exige grandes volumes de dados para treinamento eficaz. Estudos demonstram que, quando pré-treinado em conjuntos massivos (como ImageNet-21k) e ajustado posteriormente (*fine-tuning*), o ViT pode superar arquiteturas convolucionais em tarefas de classificação [3].

SigLIP2-Base — O SigLIP2 é um modelo multimodal desenvolvido pelo Google DeepMind, que combina representações visuais e textuais. Ele é uma evolução dos modelos CLIP, e introduz novos objetivos de treinamento que enriquecem a representação semântica das imagens [5]. O modelo *siglip2-base-patch16-224* utiliza uma arquitetura baseada no ViT-Base para o codificador de imagens, com *patches* de 16×16, e um codificador textual paralelo. Ambos são treinados conjuntamente usando funções de perda mais complexas, como *sigmoid loss* e auto-destilação (*self-distillation*). Para tarefas de classificação, o SigLIP2 utiliza apenas o codificador visual e uma camada linear de saída adaptada para o número de classes. Seu diferencial está na qualidade das representações visuais aprendidas por meio do treinamento multimodal, o que pode trazer benefícios mesmo quando o texto não é utilizado diretamente.

Swin Transformer (Base) — O Swin Transformer é uma variante hierárquica dos Transformers para visão computacional [4]. Ao contrário do ViT, que aplica atenção global diretamente a toda a imagem, o Swin organiza a imagem em uma estrutura piramidal semelhante às redes convolucionais, aplicando atenção local restrita a janelas deslizantes (*shifted windows*). Essas janelas se sobrepõem parcialmente entre camadas, permitindo a propagação de informações globais ao longo da rede. Essa arquitetura reduz a complexidade computacional de $O(n^2)$ para $O(n)$, onde n é o número de pixels da imagem, o que a torna mais eficiente. Além disso, o Swin atinge resultados competitivos em tarefas de classificação, detecção e segmentação, combinando desempenho robusto com menor custo computacional.

3. Protocolo Experimental

3.1. Dataset utilizado

Nesta pesquisa, vamos utilizar o conjunto de dados Food-101, proposto por Bossard et al. [1], que é amplamente utilizado como benchmark em tarefas de classificação de alimentos. Ele contém imagens coloridas (RGB) de pratos divididos em 101 classes, sendo cada classe correspondente a um tipo distinto de alimento, como *pizza*, *spaghetti bolognese*, *cheesecake*, entre outros.

Cada classe possui exatamente 1.000 imagens, das quais 750 são destinadas ao treinamento e 250 ao teste, totalizando:

- 75.750 imagens de treino
- 25.250 imagens de teste
- 101.000 imagens no total

As imagens foram coletadas do site `FoodSpotting`, uma antiga rede social voltada ao compartilhamento de fotos de pratos, o que confere ao dataset alta variedade de estilos visuais, resoluções e condições de iluminação. Embora o serviço tenha sido descontinuado, ele forneceu imagens realistas, com ruído típico de ambientes não controlados, o que torna a base representativa para aplicações do mundo real.

A Tabela 1 resume as principais características do dataset.

3.2. Modelos Avaliados

Foram avaliados três modelos de classificação de imagens disponíveis no repositório Hugging Face, todos previamente ajustados no dataset Food-101 pelos próprios autores, sem nenhum re-treinamento adicional:

- ViT-Base (`nateraw/food`): Vision Transformer com *patch size* 16 [2].
- Swin-Base (`skylord/swin-finetuned-food101`): versão base do Swin Transformer com atenção por janelas móveis [4].
- SigLIP2-Base (`prithivMLmods/Food-101-93M`): modelo baseado em ViT-Base, treinado com técnicas de aprendizado multimodal [5].

3.3. Métricas Avaliadas

Para uma avaliação completa dos modelos, definimos métricas de precisão e de custo computacional:

Característica	Descrição
Número de classes (categorias)	101 tipos de alimentos
Imagens por classe	1.000 (750 treino / 250 teste)
Total de imagens de treino	75.750
Total de imagens de teste	25.250
Número total de imagens	101.000
Tipo de imagem	Coloridas (RGB)
Origem das imagens	Site `foodspotting.com` (rede social de fotos de pratos)
Dimensão média	Aproximadamente 512×512 px
Formato original	JPG, com variação de resolução e qualidade

Tabela 1: Resumo das características do dataset Food-101.

- Top-1 Accuracy: Fração de predições em que a classe com a maior probabilidade é a classe correta.
- Top-5 Accuracy: Fração de predições em que a classe correta está entre as cinco com maior probabilidade.
- Macro F1-Score: Média aritmética das pontuações F1 de cada classe, garantindo que todas as classes tenham o mesmo peso na avaliação final.
- Parâmetros (Milhões): Número total de pesos treináveis do modelo, indicando seu tamanho em memória.
- GFLOPs: Giga Floating Point Operations per Second. Medida teórica da complexidade computacional do modelo para processar uma única imagem.
- Latência de Inferência (ms): Tempo de processamento para uma única imagem (batch size = 1), medido em milissegundos. Para garantir a robustez da medição, foram realizadas 5 execuções para cada modelo. Antes de cada série de medições, foi conduzido um aquecimento (warm-up) de 10 inferências para estabilizar os clocks da GPU. Os resultados são apresentados como média ± desvio padrão.

3.4. Procedimento de Avaliação

A avaliação dos modelos foi conduzida utilizando o conjunto de teste oficial do Food-101. O ambiente computacional consistiu em uma máquina virtual do Google Colab equipada com uma GPU NVIDIA T4.

Para garantir que cada modelo recebesse os dados no formato exato para o qual foi treinado — um fator crítico para a reprodutibilidade dos resultados —, o pré-processamento foi realizado utilizando o `AutoImageProcessor` específico de cada modelo, disponibilizado pela biblioteca `transformers`. Este procedimento aplica automaticamente as transformações corretas, como redimensionamento e normalização, de acordo com a configuração original do modelo.

O processo de inferência para o cálculo das métricas de precisão foi realizado em lotes (*batches*) de 128 imagens para otimizar a vazão (throughput) na GPU. Todos os modelos foram executados em modo de avaliação (`eval()`), sem qualquer treinamento ou ajuste fino adicional, assegurando condições de teste idênticas para todas as arquiteturas.

4. Resultados e Discussão

Esta seção apresenta os resultados quantitativos da avaliação e uma discussão sobre as observações, focando tanto na precisão preditiva quanto na eficiência computacional dos modelos. A Tabela 2 consolida todas as métricas aferidas.

4.1. Análise do Desempenho Preditivo

O modelo Swin-Base alcançou o melhor desempenho nas métricas de Top-1 Accuracy e Macro F1-Score, indicando ser o classificador mais preciso entre os avaliados. O ViT-Base, por sua vez, obteve a maior Top-5 Accuracy, sugerindo que, mesmo quando não acerta a primeira predição, a resposta correta frequentemente se encontra entre suas principais hipóteses.

É relevante notar que a acurácia obtida para o Swin-Base (90.34%) foi inferior à reportada pelo autor em seu repositório (92.14%). Essa discrepância é esperada e pode ser atribuída ao fato de que os autores comumente reportam a acurácia máxima atingida no conjunto de validação, enquanto este trabalho avaliou o desempenho no conjunto de teste final, que representa um desafio de generalização maior e mais realista.

	Métricas de Precisão			Métricas de Custo Computacional		
Modelo	**Top-1 Acc. (%)**	**Top-5 Acc. (%)**	**Macro F1**	**Params (M)**	**GFLOPs**	**Latência (ms)**
ViT-Base	89.14	**97.70**	0.8915	**85.88**	16.86	21.81 ± 5.79
Swin-Base	**90.34**	97.57	**0.9034**	86.85	**15.15**	43.73 ± 5.95
SigLIP2-Base	89.68	96.73	0.8966	92.96	16.78	**21.70 ± 5.29**

Tabela 2: Comparativo de precisão e custo computacional dos modelos no conjunto de teste do Food-101. A latência é reportada como média ± desvio padrão de 5 execuções. O melhor resultado em cada métrica está em negrito.

4.2. Análise do Custo e Eficiência Computacional

A análise das métricas de custo revela um panorama mais complexo. Uma observação notável é a disparidade significativa na latência de inferência, apesar da relativa semelhança em parâmetros e FLOPs. O modelo Swin-Base, embora seja o mais eficiente em termos de operações teóricas (menor GFLOPs), apresentou uma latência (43.73 ms) superior ao dobro dos modelos baseados em ViT (aprox. 21.7 ms).

Esta diferença pode ser atribuída diretamente à arquitetura de cada modelo e sua adequação ao hardware da GPU. Modelos ViT, como o `nateraw/food` e o `prithivMLmods/Food-101-93M`, baseiam-se em grandes e densas multiplicações de matrizes no seu mecanismo de atenção. Essas operações são massivamente paralelizadas e altamente otimizadas nas bibliotecas de software (cuDNN) para GPUs, resultando em um tempo de execução prático muito baixo.

Em contrapartida, a arquitetura do Swin Transformer fragmenta o cálculo da atenção em janelas locais menores e realiza operações de deslocamento para agregar informação. Embora isso reduza o número teórico de FLOPs, essa sequência de operações menores e mais complexas gera um maior overhead de acesso à memória e não satura os núcleos da GPU com a mesma eficiência. Portanto, conclui-se que a maior latência do Swin-Base não se deve a uma maior complexidade de cálculo, mas a uma menor otimização de sua arquitetura para a execução prática no hardware avaliado, caracterizando um *trade-off* entre acurácia e velocidade de inferência.

5. Conclusão

Este trabalho conduziu uma análise comparativa de três modelos Transformer pré-treinados — ViT-Base, Swin-Base e ViT-Base-SigLIP — na tarefa de classificação de alimentos. A avaliação revelou que, embora todos os modelos demonstrem alta capacidade preditiva, não existe uma única solução superior. O modelo Swin-Base consolidou-se como a arquitetura mais precisa, alcançando a maior acurácia Top-1 (90.34%) e Macro F1-Score (0.9034). Contudo, essa precisão superior vem ao custo de uma eficiência prática reduzida, manifestada pela maior latência de inferência (43.73 ms) — mais que o dobro dos seus concorrentes. Conforme discutido, tal comportamento é um reflexo direto de sua arquitetura de atenção em janelas, que, apesar de teoricamente eficiente em FLOPs, é menos otimizada para o paralelismo de hardware de GPUs modernas.

Em contrapartida, os modelos baseados na arquitetura ViT demonstraram notável velocidade. O ViT-Base-SigLIP emergiu como o modelo mais rápido em inferência (21.70 ms), tornando-o ideal para aplicações sensíveis à latência. O ViT-Base, por sua vez, além de apresentar desempenho similar, destacou-se por ser o modelo mais enxuto em número de parâmetros (85.88 M) e obter a melhor acurácia Top-5 (97.70%), caracterizando-se como uma solução de grande equilíbrio.

Portanto, a escolha do modelo ideal é estritamente dependente dos requisitos da aplicação. Para sistemas onde a máxima precisão é o critério decisivo, o Swin-Base é a escolha indicada. Para cenários que demandam respostas em tempo real, como aplicações interativas, o ViT-Base-SigLIP representa a opção mais adequada. Já o ViT-Base se posiciona como uma solução versátil e de excelente custo-benefício geral.

Em suma, este estudo reforça que a análise de modelos de inteligência artificial deve transcender métricas isoladas. A avaliação conjunta de precisão e eficiência prática é fundamental, e os resultados demonstram que indicadores teóricos como GFLOPs nem sempre se traduzem em menor latência.

Limitações e Trabalhos Futuros

Embora este estudo tenha fornecido uma análise sobre o desempenho de modelos Transformers no Food-101, algumas limitações devem ser consideradas.

Primeiramente, todos os modelos foram avaliados apenas no conjunto de teste do próprio Food-101, sem

validação cruzada ou generalização para outros datasets de classificação de alimentos. Isso limita a capacidade de extrapolação dos resultados para cenários de aplicação no mundo real.

Além disso, o estudo se concentrou exclusivamente em inferência, utilizando modelos previamente treinados. Não foram realizados experimentos de treinamento supervisionado ou ajustes adicionais, o que pode impactar significativamente o desempenho, especialmente em contextos com dados específicos ou distribuições distintas.

Outro ponto relevante é que as métricas de custo computacional foram obtidas considerando apenas o ambiente de execução utilizado (GPU com CUDA), o que pode variar significativamente dependendo da infraestrutura de hardware.

Como trabalhos futuros, recomenda-se:

- Avaliar os modelos em outras base de dados de classificação de alimentos, explorando sua capacidade de generalização.
- Realizar o ajuste fino com diferentes tamanhos de conjuntos de dados para avaliar a sensibilidade dos modelos ao tamanho do conjunto de treino.
- Investigar a aplicação de técnicas de compressão e quantização de modelos para reduzir latência e tamanho de arquivo, especialmente para cenários embarcados.
- Incluir métricas adicionais, como consumo de memória VRAM durante a inferência e tempo de carregamento do modelo.

Essas futuras abordagens podem ampliar a compreensão dos trade-offs envolvidos e orientar escolhas mais assertivas em projetos de visão computacional aplicada à classificação de alimentos.

Agradecimentos

Agradeço ao Professor Paulo Ricardo Lisboa de Almeida pela orientação e apoio técnico durante o desenvolvimento deste trabalho. Estendo meus agradecimentos à Universidade Federal do Paraná e aos professores da Especialização em Ciência de Dados e Big Data pela qualidade do ensino. Agradeço também à minha tia Nenzinha, que me acolheu em sua casa nos primeiros meses do curso. Sua generosidade e apoio nesse momento de transição foram essenciais para que eu pudesse iniciar essa nova etapa da minha vida acadêmica.

De forma especial, agradeço à minha mãe Silvana, ao meu pai Gilson e ao meu irmão Kauã, que me apoiaram nessa decisão e me incentivaram desde o início. Sou profundamente grato por todo amor, apoio e confiança que sempre depositaram em mim. Agradeço ainda à minha noiva Bruna, que além de me apoiar durante todo o percurso, aceitou embarcar comigo nessa mudança de cidade. Sua parceria e compreensão tornaram possível a realização deste projeto.

Referências

[1] L. Bossard, M. Guillaumin, and L. Van Gool, *Food-101 – Mining Discriminative Components in Food Images*, European Conference on Computer Vision (ECCV), 2014.

[2] A. Dosovitskiy *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, International Conference on Learning Representations (ICLR), 2021.

[3] A. Steiner *et al.*, *How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers*, Advances in Neural Information Processing Systems (NeurIPS), 2021.

[4] Z. Liu *et al.*, *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[5] C. Chen *et al.*, *SigLIP: Scaling and Improving Vision-Language Pretraining*, International Conference on Learning Representations (ICLR), 2024.

[6] A. Radford *et al.*, *Learning Transferable Visual Models From Natural Language Supervision*, Proceedings of the International Conference on Machine Learning (ICML), 2021.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, Advances in Neural Information Processing Systems (NeurIPS), 2012.