Seleção de amostras baseada em *transfer learning* e autoaprendizagem para tarefas *zero-shot* *

Matheus V. Todescato[1], **Joel L. Carbonera**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{mvtodescato,jlcarbonera}@inf.ufrgs.br

Abstract. *In environments with a scarcity of labeled visual data, such as in medical or geological contexts, deep learning models suffer from performance issues. In this context, we developed a pipeline using transfer learning and self-learning to enable the training of a classifier that can be used for various zero-shot tasks such as image classification or out-of-distribution (OOD) detection. Experimental evaluations demonstrate that our approach outperforms the state-of-the-art for these tasks.*

Resumo. *Em ambientes com escassez de dados visuais rotulados, como no contexto médico ou geológico, modelos de aprendizado profundo sofrem com problemas de desempenho. Nesse contexto, desenvolvemos um pipeline utilizando transfer learning e auto-aprendizagem para viabilizar o treinamento de um classificador, podendo ser utilizado para diferentes tarefas zero-shot como classificação de imagens ou detecção de amostras de fora da distribuição (OOD). Avaliações experimentais demonstram que nossa abordagem supera o estado-da-arte dessas tarefas.*

1. Introdução

Os avanços recentes em aprendizado profundo melhoraram significativamente o desempenho da classificação de imagens, tornando essas técnicas dominantes na pesquisa atual. Métodos de classificação baseados em aprendizagem profunda agora são empregados em diversas áreas. Em geral, modelos complexos de aprendizagem profundos requerem conjuntos de dados anotados para um desempenho eficaz. No entanto, dados visuais (por exemplo, imagens) anotados são escassos em muitas aplicações práticas, criando um obstáculo significativo para a implantação bem sucedida desses modelos. Nesse contexto, geralmente chamados de *zero-shot* (sem dados rotulados), modelos de visão e linguagem (VLMs) demonstraram um grande potencial de desempenho. Esses modelos são treinados em grandes conjuntos de dados de pares de imagem e texto da internet, realizando um processo de aprendizado contrastivo, aprendendo a comparar imagens com suas legendas corretas. O modelo CLIP [Radford et al. 2021] se destaca nesse contexto.

Além desse tipo de modelo, destaca-se a utilização de *transfer learning*, aplicando modelos visuais pré-treinados em grandes conjuntos de dados de imagens

*Este trabalho foi apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de financiamento 001; e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

[Torrey and Shavlik 2010] para extrair características informativas dos dados sem a necessidade de treinamento específico [Zhuang et al. 2020]. Com esses características é possível realizar o treinamento em um classificador de imagens simples, obtendo alto desempenho na tarefa de classificação [Todescato and Carbonera 2024].

Nesse contexto, propomos a utilização de *transfer learning*, aplicando a CLIP e um modelo visual pré-treinado em um ciclo de treinamento autossupervisionado para seleção de amostras para treinamento de um classificador simples, podendo aplicar o mesmo a diferentes tarefas como classificação de imagens e detecção de amostras de fora da distribuição (*OOD Detection*). Experimentos demonstram que nossa abordagem supera o método *zero-shot* de base em classificação de imagens e detecção de OOD por uma margem considerável.

2. Trabalhos relacionados

Para tarefas *zero-shot* relacionadas a imagens, a utilização da CLIP é destaque. Sua capacidade de representar imagens e textos em um espaço latente compartilhado permite a comparação de um conjunto de nomes de classes e imagens, sendo a base da tarefa. Em classificação de imagens, diversos trabalhos buscam melhorar o desempenho do CLIP, visando principalmente dois nichos: gerar melhores *prompts* usando LLMs[Saha et al. 2024] ou ajustar o modelo por meio de novo treinamento [Mirza et al. 2023]. Já para a tarefa de detecção de OOD, alguns trabalhos buscam outros caminhos: treinamento adicional de modelos auxiliares [Esmaeilpour et al. 2022] e aplicação de melhorias a saída do modelo [Ming et al. 2022, Miyai et al. 2025]. Porém, todas essas abordagens apresentam uma mesma linha de limitações: alta dependência da semântica das classes. Além disso, alguns ainda exigem treinamento adicional em conjuntos de dados genéricos. Essas características fazem com que conjuntos de dados de domínios específicos tenham perda de desempenho. Com a utilização do modelo visual como fonte de informação para o classificador, a nossa abordagem busca melhorar esse aspecto, criando um ciclo colaborativo que mitiga esse tipo de problema.

3. Abordagem proposta

Nosso pipeline integra: (1) **Codificador de Imagem/Texto** (CLIP) para seleção de sementes baseada em similaridade; (2) **Extrator de Características** (neste caso, ViT-G-32[1]) para características visuais de alta qualidade; e (3) **Classificador Leve** treinado iterativamente em pseudorrótulos. A Figura 1 apresenta todo o *pipeline*, que é dividido em três etapas, como apresentado na sequência.

Etapa A: Seleção de Sementes — O CLIP codifica rótulos e imagens em um espaço compartilhado, sendo usado para realizar uma avaliação de similaridade entre o nome das classes e as imagens do *dataset*. Os melhores candidatos para cada classe são selecionados, produzindo um *ranking*. Os k melhores formam o conjunto de treinamento inicial ($SEED$).

Etapa B: Treinamento do Classificador — Um classificador linear é treinado em $SEED$ e depois refinado em ciclos: cada ciclo adiciona amostras previstas com segurança das classificações e ajusta o classificador até que os limites de perda ou ciclo sejam atingidos.

[1] `https://github.com/huggingface/pytorch-image-models`

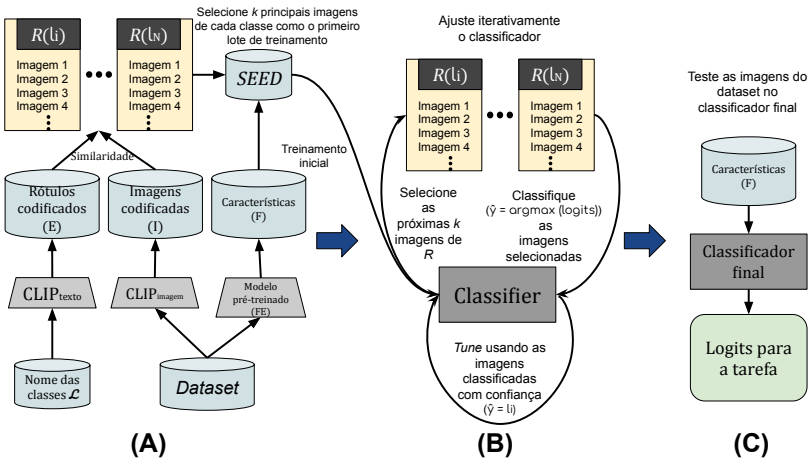

Figura 1. *Pipeline* da abordagem

Tabela 1. Performances das abordagens avaliadas na tarefa de classificação de imagens

	IMAGENET	CARS	CIFAR10	CIFAR100	CALTECH101	CALTECH256	Média
CLIP	61,15	57,62	87,65	59,54	81,45	84,04	71,91
AdaptCLIPZS (A)	63,11	57,73	87,40	61,52	85,68	85,50	73,49
Nossa abordagem	**68,29**	**79,63**	**95,07**	**80,14**	**87,12**	**89,75**	**83,33**

Etapa C: Classificador final — O classificador final é então utilizado para realizar a tarefa com base em sua saída. Para a tarefa de classificação de imagens, seleciona-se a classe com maior probabilidade. Para detecção de OOD, seus *logits* de saída são usados para calcular três critérios diferentes: o valor *logit* mais alto, a diferença entre os dois *logits* superiores e o desvio padrão de todos os *logits*. Com esses critérios então é possível calcular a possibilidade (pontuação) da imagem ser de fora da distribuição.

4. Metodologia e resultados

Avaliamos nossa abordagem em duas tarefas distintas: classificação de imagens e detecção de OOD. Para uma avaliação robusta utilizamos conjuntos de dados altamente utilizados na literatura (como em [Todescato and Carbonera 2024]), desde domínios genéricos até domínios refinados, usando apenas nomes de classes como supervisão. As abordagens de base incluem *zero-shot* CLIP e abordagens do estado-da-arte. As medições utilizadas foram a precisão top-1 para classificação e AUROC para detecção de OOD e estão reportadas nas Tabelas 1 e 2, respectivamente.

Nossa abordagem supera consistentemente a CLIP e as abordagens do estado-da-arte, tanto na classificação quanto na detecção de amostras OOD. Destacam-se os maio-

Tabela 2. Performances das abordagens avaliadas na tarefa de detecção de amostras OOD em 5 *splits*

	CARS	GEO	CIFAR10	CIFAR100	CALTECH101	CALTECH256	Média
MCM	63,93±4,5	47,10±7,0	89,26±4,0	78,10±1,7	89,96±8,7	91,15±6,0	76,58
GL-MCM	63,16±4,3	45,36±6,7	88,34±4,9	78,94±1,7	91,53±6,5	91,24±5,9	76,43
ZOC	72,50±6,8	46,41±7,2	**93,00±1,7**	82,10±2,1	90,06±6,5	91,50±4,3	79,26
Nossa abordagem	**73,19±6,6**	**58,17±6,3**	89,26±3,2	**85,60±1,2**	**92,80±5,0**	**93,86±2,7**	**82,15**

res ganhos em conjuntos de dados refinados onde as características visuais são cruciais, destacando os pontos fortes que alavancam a autoaprendizagem e o *transfer learning*, combinando o potencial da CLIP com um extrator de características visuais poderoso.

5. Conclusão

O método é simples de implementar, livre de rótulos e adaptável a novos domínios. Seu design modular permite a integração futura de codificadores, extratores ou classificadores aprimorados sem alterar o ciclo de autoaprendizagem. O design para ambientes com escassez de dados ajuda a superar um desafio frequente em aplicações do mundo real e uma grande barreira para a adoção mais ampla da IA. Apesar de seus pontos fortes, nossa abordagem ainda depende do desempenho da CLIP e da condição semântica do conjunto de dados (sejam palavras distintas ou comuns). Trabalhos futuros se concentrarão em refinar esses aspectos para melhorar a robustez e a adaptabilidade.

Referências

Esmaeilpour, S., Liu, B., Robertson, E., and Shu, L. (2022). Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6568–6576.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. (2022). Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102.

Mirza, M. J., Karlinsky, L., Lin, W., Possegger, H., Kozinski, M., Feris, R., and Bischof, H. (2023). Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections. *Advances in Neural Information Processing Systems*, 36:5765–5777.

Miyai, A., Yu, Q., Irie, G., and Aizawa, K. (2025). Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, pages 1–11.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Saha, O., Van Horn, G., and Maji, S. (2024). Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17542–17552.

Todescato, M. V. and Carbonera, J. L. (2024). Investigating performance patterns of pre-trained models for feature extraction in image classification. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1024–1031. IEEE.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Zhuang, F. et al. (2020). A comprehensive survey on transfer learning. In *Proceedings of the IEEE 109*, pages 43–76, 1.