O impacto de transformações de imagens no contexto de abordagens de self-supervised learning utilizando contrastive learning

Misael S. de Rezende, Jesimon Barreto, William R. Schwartz
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte - MG, Brasil
Email: misaelsr1@gmail.com, jesimonbarreto@dcc.ufmg.br, william@dcc.ufmg.br

Esta pesquisa investiga o impacto das transformações de imagens no contexto da aprendizagem auto-supervisionada, especialmente quando combinadas com técnicas de aprendizado contrastivo. Nosso objetivo é avaliar como várias transformações de imagens influenciam a qualidade das representações aprendidas e, consequentemente, o desempenho geral do modelo. Ao focar nas limitações de métodos existentes, incluindo o modelo LEWEL, nosso estudo busca aprofundar a compreensão dos efeitos das transformações de imagens na aprendizagem auto-supervisionada. Através de experimentos no conjunto de dados ImageNet-100, exploramos as implicações das transformações nas representações e sua transferibilidade para classificação linear.

***Abstract*—This research investigates the impact of image transformations in the context of learning self-supervised, especially when combined with contrastive learning techniques. Our objective is evaluate how various image transformations influence the quality of the learned representations and, consequently, the overall performance of the model. By focusing on the limitations of existing methods, including the LEWEL model, our study seeks to deepen the understanding of the effects of transformations of images in self-supervised learning. Across experiments on the ImageNet-100 dataset, we explored the implications of transformations in representations and their transferability to linear classification.**

I. Introdução

Nos últimos anos, uma abordagem que tem ganhado mais atenção da literatura é a aprendizagem auto-supervisionada, também conhecida como *self-supervised learning* (SSL). Ela se insere na categoria da aprendizagem de representação não supervisionada e possibilita uma alternativa promissora para o desafio da representação de imagens. O que torna o SSL promissor é a não necessidade de rotulagem prévia dos dados, ou seja, o treinamento de um modelo seguindo essa abordagem utiliza os próprios dados para guiar o processo de aprendizagem [1]. Ainda mais, a sua capacidade de aproveitar a grande quantidade de dados disponível atualmente de forma eficiente também contribui para atração da atenção da literatura para essa abordagem. A eficácia dessa abordagem tem alcançado resultados impressionantes, chegando a rivalizar com métodos que dependem de dados rotulados.

Recentemente, alguns trabalhos como [2]–[4] tem resultados muito próximos aos melhores modelos de abordagens supervisionadas. O trabalho de [4] é um exemplo, onde os autores propõem uma abordagem que usa parte dos conceitos propostos em [2]. Os autores de [4] propõem uma abordagem usando aprendizagem auto-supervisionada, que é adaptativa para agregar informação espacial de características da projeção global de uma rede neural, resultando em uma contribuição significativa para a literatura da área.

No entanto, mesmo com esses avanços, permanece uma questão importante a ser explorada: o impacto das transformações de imagens no contexto do *self-supervised learning*, especialmente quando combinadas com técnicas de contrastive learning [1]. Nossa pesquisa se propõe a investigar essa questão, avaliando como diferentes transformações de imagens afetam não apenas a qualidade das representações aprendidas, mas também a transferência dessas representações para downstream tasks (aplicações dos modelos pré-treinados usando a abordagem SSL).

II. Trabalhos Relacionados

Dentro do vasto campo de SSL, a abordagem contrastiva se destaca como uma estratégia proeminente [1], [3], [5], [6]. E tem sido amplamente adotada para aprender representações robustas, especialmente no contexto de imagens. A premissa central da aprendizagem contrastiva é, durante o treinamento e usando alguma função de similaridade, aproximar pares de amostras positivas (que compartilham características semelhantes) e distanciar pares de amostras negativas (que são dissimilares). Entretanto, a aprendizagem contrastiva precisa ser combinada junto a outras técnicas para auxiliar na aprendizagem por representação. Um método comum para realizar isso em imagens é aplicar transformações diversas, gerando pares positivos e negativos a partir de uma imagem de referência [7].

O papel das transformações na aprendizagem auto-supervisionada transcende a geração convencional de pares positivos e negativos. Estas transformações desempenham um papel crucial ao introduzir variações significativas nos dados (Figura 1, [5]), desafiando o modelo a aprender características robustas e invariantes diante de alterações na apresentação visual. Contudo, isso não significa que um número alto de transformações implique em melhores resultados. Alguns

trabalhos focam em usar regras pré-definidas ou encontrar transformações ideais para as tarefas de aplicação. Outros trabalhos [4] defendem que deve-se evitar adicionar informações específicas à arquitetura de treino, mas sim alterar a arquitetura para que o processo de treinamento já seja invariante a downstream task, ou seja, tornar a rede independente da tarefa de aplicação.

O trabalho de [4], apresentou a *LEWEL*, um framework que adapta a aprendizagem contrastiva para alinhar representações em baixo nível, aprimorando o processo de aprendizagem. Apesar de seus méritos, a influência específica de cada transformação nas imagens no desempenho da LEWEL permanece inexplorada. Esta lacuna motiva diretamente nossa pesquisa, que busca responder essa pergunta, examinando como diferentes transformações afetam as representações geradas e, consequentemente, a performance do modelo.

Assim, a compreensão das interações entre transformações de imagens e a aprendizagem auto-supervisionada torna-se essencial. Nosso trabalho se alinha a esses esforços, concentrando-se na investigação desses tópicos. Ao abordar essa questão central, esperamos trazer uma contribuição significativa para o campo em constante evolução de aprendizagem por representação e aprendizagem contrastiva.

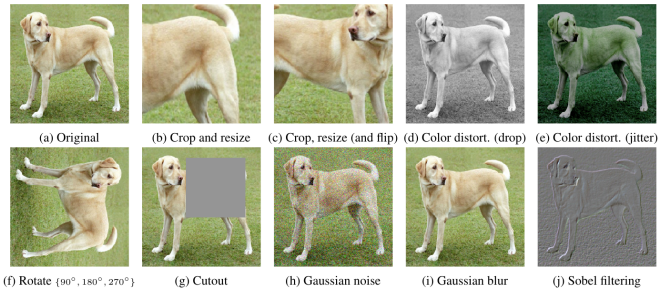

Fig. 1. Exemplo de algumas transformações possíveis aplicadas em uma imagem. Reprodução de [5].

III. Metodologia

A metodologia seguida nesse trabalho centra-se na realização de experimentos que exploram diversas transformações aplicadas a imagens, o que não foi devidamente abordado tanto no artigo base, como em alguns outros artigos da literatura. As transformações consideradas estão definidas em III-B. O treinamento da rede é conduzido separadamente para cada transformação, permitindo uma avaliação tanto quantitativa quanto qualitativa do impacto dessas alterações nas representações geradas pela rede neural. Para simplificação, apenas a aplicação (do inglês *downstream task*) de classificação linear foi avaliada neste trabalho.

As avaliações quantitativas abrangeram análises da função de perda e acurácia no conjunto de validação após o treinamento, enquanto as avaliações qualitativas envolveram a análise de mapas de calor gerados pelo Grad-CAM [8] para identificar características relevantes nas transformações que afetam o aprendizado da rede.

TABELA I
Pré-treino com acurácia 1 média

Transformação	Self-supervised
Todas transformações	**58.9** [49.4, 68.3]
Resized Crop	32.3 [28.5, 36.0]
Color Jitter	11.5 [9.3, 13.7]
Gaussian Blur	9.7 [9.5, 10.0]
Solarize	9.3 [9.1, 9.5]
Grayscale	7.5 [7.3, 7.7]
Horizontal Flip	7.5 [5.0, 10.0]

O dataset escolhido é o dataset ImageNet-100 (IN-100), seguindo [4], [9]. Esse dataset foi escolhido por ser um subconjunto menor (10% do dataset completo) que o dataset ImageNet-1K [10], permitindo um tempo de treinamento menor e resultados mais rápidos. O treinamento segue os hiperparâmetros propostos para LEWEL nos estudos de ablação para a variante LEWEL$_B$, com ajustes específicos, como 100 épocas na etapa de self-supervised learning e tamanho do batch de 128, sendo que esse último, tanto no treinamento SSL quanto na classificação linear.

A. Implementação

A função de perda escolhida foi a variante LEWEL$_B$ para fazer os experimentos. A etapa de validação dos treinamentos, tanto na etapa da aprendizagem auto-supervisionada quanto na etapa de classificação linear segue o trabalho dos autores. Dessa maneira, a etapa de pré-treino foi feita com treinamento no conjunto de treino do dataset, e validação no conjunto de validação, usando Knn como avaliação, com parâmetros iguais ao artigo base. Foi escolhido a priori, uma separação única e aleatória dos dados, em conjuntos de treino, validação e teste, de 70%, 20% e 10%, respectivamente. Por conseguinte, a etapa de pré-treino e downstream task foi realizada sempre com o mesmo conjunto de dados.

B. Transformações

As transformações utilizadas para o treinamento são as mesmas utilizadas no LEWEL: Resized Crop, Color Jitter, Gaussian Blur, Solarize, Grayscale e Horizontal Flip. Em alguns experimentos, as probabilidades das transformações em uma ou nas duas novas amostras geradas são ajustadas, e são mencionadas na Seção IV. O modelo baseline foi treinado com todas essas transformações no conjunto de treinamento.

IV. Experimentos

Todos os experimentos foram feitos utilizando 2 GPUs GTX1080Ti, Pytorch 1.12.1 e CUDA 10.2. Os hiperparâmetros são os mesmos definidos na Seção III. Além disso, todos os experimentos foram repetidos por três vezes para calcular a média e o intervalo de confiança de 90% para os resultados, este último representado pelos valores entre colchetes nas tabelas. Como a amostra é pequena, foi utilizado a distribuição t de Student com $n-1$ graus de liberdade. As métricas avaliadas foram Acc@1 e Acc@5, sendo que a primeira mede a precisão do modelo ao prever a classe alvo

TABELA II
CLASSIFICAÇÃO LINEAR COM ACURÁCIA MÉDIA

Transformação	Classificação Linear	
	Acc@1	Acc@5
Todas transformações	**82.3** [77.6, 87.0]	**95.3** [93.5, 97.2]
Resized Crop	64.4 [63.2, 65.63]	86.4 [85.3, 87.5]
Color Jitter	32.5 [27.2, 37.8]	58.0 [52.6, 63.4]
Solarize	22.7 [17.5, 27.9]	46.8 [40.4, 53.2]
Gaussian Blur	21.7 [18.9, 24.4]	44.7 [41.1, 48.4]
Grayscale	16.9 [15.1, 18.7]	38.1 [35.1, 41.1]
Horizontal Flip	13.8 [7.4, 20.2]	32.4 [21.4, 43.5]

como a mais provável enquanto a última considera se a classe alvo está entre as cinco mais prováveis.

A. Resultados do Pré-Treino

A Tabela I apresenta os resultados do treinamento na etapa de auto-supervisão, do modelo de referência (baseline) em comparação com os treinamentos realizados com apenas uma transformação por vez. Destaca-se que, depois do baseline, a transformação Resized Crop obteve a melhor acurácia média, e entendemos, como mencionado em [4], na qual argumentam que um bom grau de desalinhamento espacial é benéfico para a aprendizagem. Outras transformações, no entanto, apresentaram resultados muito inferiores ao baseline ou mesmo do Resized Crop, indicando que, transformações no espaço de cor, junto com a Horizontal Flip, por si só, não são suficientes para gerar representação robustas durante o treinamento SSL. Ainda mais, observando o intervalo de confiança, todas essas outras transformações são estatisticamente equivalentes.

B. Resultados na Tarefa de Classificação Linear

Os resultados dos modelos pré-treinados na etapa de auto-supervisão, quando aplicados à tarefa de classificação linear, são apresentados na Tabela II. Esses resultados seguem padrões semelhantes aos da etapa de pré-treino, onde o modelo baseline alcançou os melhores resultados, seguido pelo modelo treinado apenas com a transformação Resized Crop. As demais transformações mantiveram desempenhos consistentes, sugerindo que modelos pré-treinados com transformações menos eficazes ou individualmente menos úteis continuam a exibir desempenho inferior na etapa de downstream task. Isso ressalta a importância da qualidade do pré-treinamento na abordagem SSL e sua correlação com o desempenho subsequente.

C. Experimentos adicionais

Seguindo os primeiros experimentos, formulamos a hipótese de juntar algumas transformações para validar se a adição de mais uma transformação em cada pré-treino (Tabela I) influenciaria os resultados na etapa de downstream task. Dessa forma, foi testado três experimentos adicionais, observando tanto os resultados de pré-treino quanto da tarefa de aplicação. O primeiro experimento juntando as duas melhores transformações individuais, o segundo juntando a transformação com o melhor e com o pior resultado e por último juntando as duas transformações com o pior resultado.

TABELA III
PRÉ-TREINO COM ACURÁCIA 1 PARA SELF-SUPERVISED E CLASSIFICAÇÃO LINEAR

Transformação	Self-supervised	Classificação Linear	
		Acc@1	Acc@5
Resized Crop + Grayscale*	**57.5**	**80.3**	**94.8**
Resized Crop + Color Jitter*	42.1	71.9	90.4
Horizontal Flip + Grayscale*	7.8	17.2	39.1

A Tabela III demonstra o resultado desses experimentos, tanto para o pré-treino quanto para a classificação linear. Os resultados sugerem que a junção do Resized Crop com uma transformação que atua no espaço de cor das amostras melhora consideravelmente os resultados. Ainda mais, pode-se observar que a adição da transformação para cor de cinza já é capaz de ter resultados equiparáveis ao baseline - tanto no pré-treino como na classificação. Isso demonstra que, algumas transformações específicas, trazem grandes saltos nos resultados. Contudo, é importante salientar que esses experimentos foram conduzidos uma única vez.

Adicionalmente, dois conjuntos de experimentos foram realizados para explorar hipóteses adicionais. A partir dos resultados iniciais e as primeiras hipóteses, foi levantado outras duas hipóteses. Uma delas foi com respeito a quantidade de transformações da rede baseline. A quantidade de transformações dessa rede é bem diversa, então buscamos avaliar se a redução da quantidade de transformações impactaria no resultado. Seguindo as definições originais das probabilidades de ocorrer uma transformação para a primeira ou segunda amostra [4], foi reduzido algumas probabilidades pela metade - todas as probabilidades que já não estavam zeradas para a amostra 1 ou amostra 2. Os resultados na Tabela IV demonstram que o efeito dessa alteração indicam o oposto ao esperado. Isto é, esse modelo obteve um resultado um pouco melhor que o baseline original. Devido a limitações, não foi possível analisar com experimentos adicionais as motivações desse resultado. Adicionalmente, é importante citar que, esse modelo só foi repetido por duas vezes. No caso do baseline, os intervalos de confiança de 90% para self-supervised foi de [63.2, 71.1] e para classificação linear foi de [81.9, 87.5] para acurácia 1 e [95.8, 97.0], para acurácia 5.

Também foi testado uma outra hipótese, dessa vez para a transformação de escala em cor de cinza. Existe uma chance de que, durante o treinamento, a rede processar as mesmas imagens sem transformação, o que acabaria deixando o treinamento com menos dado diverso, e consequentemente, podendo piorar o resultado. Dessa forma, sempre vai ocorrer uma transformação para a amostra 1 e nunca vai ocorrer uma transformação em cor de cinza para a amostra 2. Os resultados mostrados na Tabela IV demonstram que, pelo menos para essa transformação, não houve significância estatística relevante. Isto é, reduzindo essa probabilidade ou deixando a probabilidade original (0.2 para as duas amostras, Tabelas I e II), não gerou alteração significativa para essa transformação. No caso do grayscale, os intervalos de confiança de 90% para

TABELA IV
PRÉ-TREINO COM ACURÁCIA 1 PARA SELF-SUPERVISED E CLASSIFICAÇÃO LINEAR

Transformação	Self-supervised	Classificação Linear	
		Acc@1	Acc@5
Baseline*	67.2	84.7	96.4
Grayscale	7.8	16.2	36.7

self-supervised foi de [7.0, 8.6] e para classificação linear foi de [14.2, 18.3] para acurácia 1 e [33.6, 39.8], para acurácia 5.

D. Análise quantitativa

A quantidade de épocas de pré-treino utilizada pelos autores [4] foi de 240 épocas enquanto nesse trabalho foi de 100 épocas. Com a análise empírica feita através de muitos experimentos com o ambiente definido desse trabalho, foi possível afirmar que a redução para 100 épocas foi o suficiente para ter uma estabilização na função de perda. Além disso, que executar por mais épocas não teriam maiores ganhos no resultado, e não afetariam a comparação e análise dos resultados porque existia uma estabilização dos valores das métricas avaliadas.

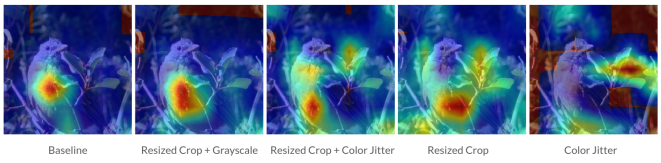

Fig. 2. Mapa de calor para um pássaro, gerado através de Grad-CAM

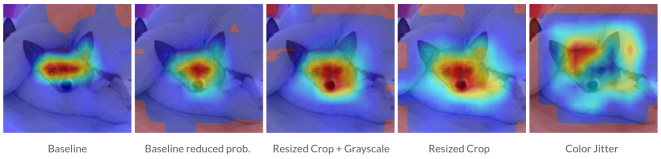

Fig. 3. Mapa de calor para um cachorro, gerado através de Grad-CAM

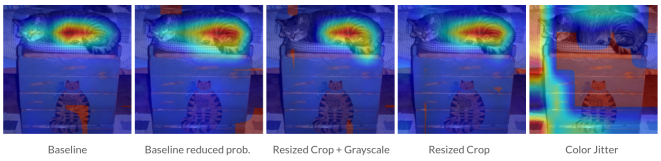

Fig. 4. Mapa de calor para um gato, gerado através de Grad-CAM

E. Análise qualitativa

Na análise qualitativa, buscamos entender como a representação aprendida no pré-treino impacta na avaliação do modelo na etapa de classificação. Para isso foi utilizado mapas de calor do Grad-CAM [8]. Para gerar o mapa de calor, foi considerado os gradientes gerados a partir do bloco de camada 4 da ResNet-50 até a camada de classificação, para a classe específica na qual se quer avaliar. A Figura 2 demonstra que os melhores modelos obtiveram um bom resultado. Por outro lado, um modelo que não apresentou um bom resultado no pré-treino - color jitter, consequentemente não conseguiu uma boa aprendizagem na downstream task, visto que o modelo utiliza muito mais informações irrelevantes para a predição. As Figuras 3, 4 complementam os resultados, mais uma vez reforçando a análise anterior. Nessas figuras, é possível notar que o experimento adicional com o baseline com probabilidades ajustadas manteve bons resultados qualitativos e comparáveis ao baseline. Adicionalmente, a Figura 4 é um exemplo com mais informações na imagem, por apresentar mais objetos na cena, incluindo uma representação de um gato. Ainda assim, os modelos apresentaram um bom resultado.

V. CONCLUSÃO

Este trabalho explorou o impacto das transformações de imagens no contexto de aprendizado auto-supervisionado (SSL), com foco especial nas técnicas de contrastive learning. Diante dos avanços no campo do aprendizado profundo, nossa pesquisa buscou responder algumas perguntas relacionadas à compreensão do papel dessas transformações nas representações aprendidas por modelos, notavelmente exemplificado pelo framework LEWEL. Os resultados obtidos durante os experimentos revelaram alguns detalhes sobre como diferentes transformações afetam as representações aprendidas. A análise quantitativa, incluindo métricas de perda e acurácia, proporcionaram uma compreensão detalhada do desempenho do modelo em diferentes cenários.

REFERENCES

[1] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," 2021.
[2] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020.
[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020.
[4] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Learning where to learn in cross-view self-supervised learning," 2022.
[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.
[6] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020.
[7] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, dec 2022. [Online]. Available: https://doi.org/10.1145\%2F3577925
[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, oct 2019. [Online]. Available: https://doi.org/10.1007%2Fs11263-019-01228-7
[9] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2020.
[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.