Because Order Counts: Curriculum Learning with Attentions in CNNs

Luay Muhtaseb
luay.muhtaseb1@mail.huji.ac.il

Lihi Shalmon
Lihi.Shalmon@mail.huji.ac.il

Abstract

Attention mechanisms excel in natural language processing and computer vision, often outperforming convolutional neural networks (CNNs). However, their reliance on large datasets limits their use in data-scarce scenarios, where CNNs are more practical. In our work we explore if curriculum learning, a method that sequences training data from simple to complex, could improve efficiency of attention augmented CNNs in comparison to standard CNNs. Using a challenging , small data subset, suited to our limited resources, we compared ResNet-20 with a few common attention-augmented architectures. Our findings show that in this restricted environment ResNet-20 achieves higher test accuracy. Additionally curriculum type and model architecture significantly influencing performance, though their interaction lacks statistical significance. These results highlight curriculum learnings potential to enhance data efficiency, offering practical insights for resource-constrained settings. We also share our code and logs on the training to support further investigation 10.1.

1 introduction

Attention mechanisms have advanced deep learning in NLP and computer vision [1],[2], initially popularized by Transformers but also integrated into CNNs. While these CNNs can capture broader dependencies, they often require large datasets [2] and more computational resources. In contrast, standard CNNs rely on local feature extraction with linear complexity of $O(k \cdot n \cdot d)$, but may underperform when a wider context is needed.

To overcome this problem, we examine whether curriculum learning [3], which orders the data in a way that facilitates the learning process can improve attention augmented CNNs and whether it could reduce the data requirements. Previous work shows that curriculum learning improves CNN training efficiency [4] (e.g., ResNet-20 [5]), with increased gains on more challenging tasks, suggesting it may also benefit attention-augmented models. Using a small, challenging subset of CIFAR-100 [6] , we apply curriculum methods under limited data conditions.

Our study addresses:

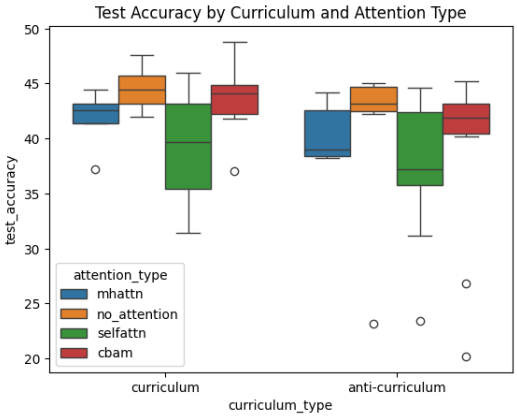

Figure 1: **ResNet-20 Outperforms Attention**
The boxplot shows test accuracy distributions 2, revealing a higher median and lower variance for ResNet-20 compared to attention-based architectures. Additionally, both curriculum and model type had significant main effects, but their interaction wasn't significant.

- Can CNNs with attention train effectively on small datasets using curriculum learning?
- Does this approach outperform standard CNNs?
- Which attention-based model performs best under these constraints?

We selected a challenging, standardized dataset used by [4] demonstrated that curriculum learning can improve CNN performance. Due to our limited computational resources, using the same low-resolution subset allowed us to run more experiments, ultimately conducting eight after tuning 72 models.
Inspired by [7], we adopt ResNet-20 and augment it with three attention-based modules. Following a methodology similar to [8] and [4], we train all models using cross-entropy loss and interpret results according to the accuracy to determine whether curriculum learning assists the training process and reduces data demands on the attention-augmented architectures.

study.

2 Methodology

2.1 Using Attention Augmented CNNs

We use Convolutional Neural Networks (CNNs) [9], specifically ResNet-20 [5], which leverages skip connec-

tions to mitigate vanishing gradients and enable deeper architectures for vision tasks (this is the baseline).

To test the effect of attention, we integrate three attention mechanisms [7]:

1. **Self-Attention (SelfAtt):** Applies 1×1 convolutions to capture spatial dependencies.

2. **Multi-Head Self-Attention (MHA):** Uses 4 heads for diverse long-range dependencies.

3. **Convolutional Block Attention Module (CBAM [10]):** Employs channel and spatial attention to highlight critical features.

These are placed between the feature extractor layers to balance efficiency and performance [7].

2.2 Dataset

For consistency, we follow the data used by Hacohen and Weinshall .We use a subset of CIFAR-100 containing 3,000 images of small mammals at 32×32 resolution to evaluate curriculum learning under limited data conditions. Due to its small size and high intra-class similarity, this dataset presents a challenging task. We apply random cropping, horizontal flipping, and normalization to enhance generalization.

2.3 Curriculum Learning

Curriculum learning organizes training data in a structured manner to optimize convergence and generalization. The curriculum learning method modifies standard training by leveraging prior knowledge about the difficulty of training examples to improve convergence and classification accuracy. Instead of sampling mini-batches uniformly, we use a predefined scoring function to prioritize simpler samples first. Curriculum learning modifies the training process by controlling the sequence and subset of examples presented to the network. This is achieved via three ingredients:

- **Scoring**: This measures difficulty with a score $s(x)$. We use transfer learning with a ResNet-20 model trained for 5 epochs to assign higher scores to harder examples.

- **Order**: This determines how we advance. We tested usage of *curriculum* (easy to hard) strategy and *anti-curriculum* (hard to easy). Due to time limitations, we did not analyze the training of the models without any curriculum.

- **Pacing**: This adjusts the number of samples per epoch, growing the amount of samples in every epoch on the basis of a criterion and order. **The algorithm for curriculum learning can be found in the appendix 10.2.**

2.4 Pacing Functions

The pacing function g_θ controls how the number of training samples increases over time. We explore **exponential pacing-** where the number of samples starts small and grows exponentially:

$$g_\theta(i) = N_0 \times \left(\frac{N}{N_0}\right)^{\frac{i}{M}} \tag{1}$$

This approach emphasizes early-stage learning from a small, easy subset, then rapidly scales up to full complexity.
Previous work by Hacohen and Weinshall demonstrated that curriculum learning (CL) significantly improves accuracy on datasets like CIFAR-100, though they found different pacing functions often converge to similar final performances. In contrast, Wu et al. [11] conducted extensive experiments showing that CL reduces the convergence time while enhancing learning performance, attributing these gains primarily to pacing functions rather than the curriculum alone. These findings motivate our exploration of exponential pacing to optimize efficiency across architectures.

2.5 Loss Function and Parameter Search

We employed the cross-entropy loss function, which is commonly used for CNN classification tasks, and optimized the models based on the loss of the test set. To optimize our models, we employed Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 5e-4, a standard choice for training convolutional neural networks like ResNet. Learning rates were tailored to each architecture and determined by Weights & Biases' bayesian hyperparameter optimization [12] (details in the appendix 10.2).

2.6 Balancing the classes within training

In curriculum learning, ordering samples by difficulty can unintentionally favor certain classes at different training stages, potentially skewing the learning process. Following Hacohen and Weinshall, we mitigate class dominance by balancing classes within each batch to ensure equal exposure to all categories.

3 Experiments

In this experiment, we evaluated 4 models paired with 2 curricula, yielding 8 unique trials. For each trial, we optimize hyperparameters, such as learning rate and pacing rate, using bayesian optimization. Training progress is monitored using Weights & Biases, while tracking the average difficulty of each epoch, the loss across training and test sets, and the accuracy.

3.1 Initial Hypothesis

We hypothesize that curriculum learning will enhance model performance, while anti-curriculum will diminish it. We expect attention-based models (CBAM, MHA, self-attention) to outperform the baseline CNN

(ResNet-20) under standard training, though their data-intensive nature may reduce their sensitivity to curriculum type. Among the attention mechanisms, we predict that CBAM will excel due to its channel and spatial focus, as previous work has demonstrated its effectiveness [10]. Curriculum learning establishes the upper performance bound, with anti-curriculum as the lower bound.

To rigorously validate our hypotheses, we employ classical statistical methods:

- **Two-way ANOVA**: Checks the effects of curriculum type, attention mechanism, and their interaction on performance.
- **T-test**: Compares the means of curriculum vs. anti-curriculum under different conditions
- **Tukeys HSD**: Tests which attention mechanisms differ significantly after multiple-comparison corrections.
- **Confusion Matrix**: Displays correct and incorrect classifications by class.
- **Saliency Maps**: Highlight image regions most influential to model predictions.

4 Hyperparameter Sensitivity

We used Weights & Biases (W&B)'s feature importance analysis to quantify the impact of hyperparameters on model performance. W&B estimates feature importance by training a **random forest regressor**, where hyperparameters serve as inputs and the performance metric (test loss) is the target. The feature importance values are derived from the random forest model.

Key findings:

- **Learning rate** had the highest contribution to model performance, reinforcing its role as a critical tuning parameter.
- **Attention mechanisms consistently degraded performance**, contradicting our initial expectations.
- **Curriculum learning remained effective** across hyperparameter variations, reinforcing its impact on generalization.

Our attempts to compute feature importance using OLS regression produced similar results.

5 Observing training progression

Observing step-level accuracy shows that curriculum generally begins with a lower initial accuracy but converges to a higher or more stable level than anti-curriculum (dark line). Additionally, ResNets attention-free architecture and CBAM appear more responsive to curriculum learning. The accuracy values are comparable to those in Hacohen and Weinshall, and the inconsistent advantage of curriculum aligns with their observation that gains on the mammals subset of CIFAR-100 stabilized after about 500 iterations. We did not train beyond this range in the current study, which we acknowledge as a limitation, and recommend exploring longer training regimes in future work.

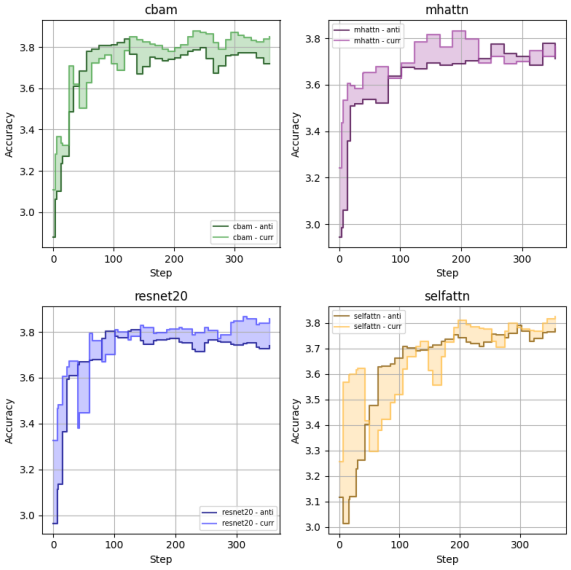

Figure 2: **Accuracy trends align with prior work** The dark line represents the anti-curriculum setting. While any initial advantage of curriculum learning is inconsistent, accuracy tends to stabilize with additional iterations. Both CBAM and ResNet-20 appear more influenced by curriculum, showing noticeable improvement. These trends align closely with prior observations in Figure 7 of Hacohen and Weinshall.

6 Quantitative Evaluation

6.1 Evaluation Framework

Test losses were very similar across the different models. However, test accuracies revealed relatively significant variations in generalization performance. This section presents a statistical analysis of how curriculum type and attention mechanisms affect model performance.

6.2 Curriculum is Effective

ANOVA analysis showed that curriculum learning significantly improved generalization ($p = 0.016$ 1), leading to an average accuracy increase of 2.55% over anti-curriculum ($t = 2.57$, $p = 0.012$). These findings support the idea that structured data presentation enhances model learning, even in data-scarce settings.

6.3 Less is More: Attention Does Not Improve

Attention mechanisms also had a significant impact on accuracy ($p = 0.034$ 1), though their effects varied across different architectures. CBAM and MHA attention mechanisms showed modest performance gains, while self-attention exhibited a decline. This stems from

self-attentions global focus, which tends to amplify the impact of noise and low-resolution artifacts. In contrast, CBAM and MHA concentrate on localized, informative regions, thereby mitigating these issues.

6.4 Attention & Curriculum Do Not Interact

Despite their individual significance, curriculum learning and attention mechanisms showed no significant interaction effect (p = 0.77 1). This indicates that curriculum learning benefits all architectures similarly, regardless of attention type. However, the extent of performance change varied across attention mechanisms:

- CBAM exhibited the largest performance drop (4.2%) when moving from curriculum to anti-curriculum.
- MHA was the least affected, with a smaller drop (1.8%).

6.5 Significant Pairwise differences

To investigate these effects in more detail, we performed Tukeys HSD test 10.3, revealing one significant pairwise difference:

- ResNet-20 without attention vs. ResNet-20 with Self-Attention: 5.04% improvement ($p = 0.017$).
- Other comparisons were non-significant ($p > 0.23$), indicating that most attention variants underperformed the CNN baseline without statistical significance.

7 Error Analysis and Visualizations:

7.1 Curriculum mitigates class confusion

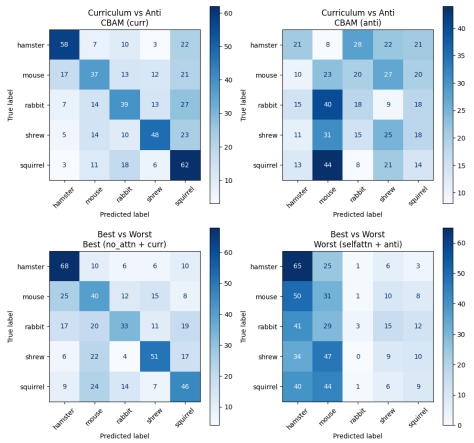

Figure 3: **Curriculum enhances class separation** Confusion matrices for different experimental settings. The Curriculum + No Attention model maintains a strong diagonal structure, while the Self-Attention + Anti-Curriculum model exhibits concentrated misclassifications.

To further examine model performance, we analyzed confusion matrices across different experimental settings. A well-performing model should exhibit a strong diagonal structure, indicating correct classifications, while weaker models display more dispersed misclassifications.

Our results highlight the effect of curriculum learning and attention mechanisms: The best-performing model which used Curriculum without attention exhibited a clear diagonal pattern, suggesting effective class separation.Meanwhile the worst-performing model, using Self-Attention + Anti-Curriculum did not show a diagonal pattern and had a high concentration of misclassifications in the first two columns, which indicates the model confused certain classes very often. Switching from curriculum to anti-curriculum consistently degraded performance, reducing diagonal emphasis. CBAM showed slight improvements in class separation, but attention-based models overall did not significantly alter classification patterns.

7.2 Saliency Shows Resolution Constraints

We analyzed classification using saliency maps to investigate feature focus, inspired by the paper of Woo et al., which revealed distinct patterns across different attention modules. Aware of saliency maps explainability limitations rudin, we treated them as a supplementary tool, not a primary method. **With curriculum learning**, attention models (CBAM, MHA) improved accuracy by focusing on key features, consistent with prior findings [10]. Meanwhile, the no-attention models focus was uniformly distributed, which was somewhat counterintuitive since we initially expected a stronger correlation between localized focus and prediction quality. **Under anti-curriculum**, self-attention fixated on irrelevant areas like backgrounds, aligning with its 23.4% accuracy. CBAM, however, retained some focus on relevant features, better than other models, despite its suboptimal strategy and 20.2% accuracy.

This approach first exposed the datas difficulty: 32x32 pixel images, often nearly dark or blurry, making it hard to distinguish between classes.

8 Discussion

8.1 Key Limitations

This study employed a single training regime for curriculum learning. Running multiple iterations and averaging the outcomes would yield more robust conclusions. Furthermore, evaluation was performed on a single, specific dataset that does not fully represent modern real-world image recognition environments with more complex details and data challenges, thus limiting generalizability.

9 Conclusion

Our experiments demonstrate that curriculum learning consistently improves training efficiency across all architectures, reinforcing its role in structured learning.

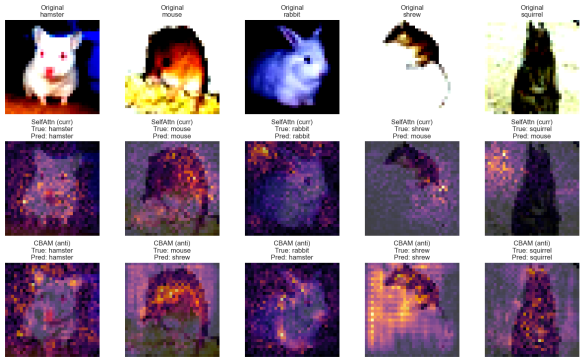

Figure 4: **Models struggle under low resolution** Curriculum learning sharpened the focus on relevant features; anti-curriculum diverted it, especially in self-attention.

However, contrary to expectations, attention-based models did not outperform CNNs. We attribute this to the data-scarce, low-resolution setup, which contrasts with the complex, data-rich environments where attention mechanisms typically excel.

9.1 Future works

Future works could address aspects like the computational complexity of the models, or ensuring robustness of results by increasing the number of iterations per experimental setting. Another promising direction is evaluating curriculum learning on larger, high-resolution datasets, which may reveal its effectiveness on more complex tasks and provide insights into its scalability. Additionally, exploring alternative curriculum strategies, such as boosting-inspired approaches that adaptively prioritize difficult examples, and experimenting with various base models for transfer learning could yield valuable insights into optimizing curriculum learning.

10 Appendix

10.1 Supplementary Code and logs

The source code for this study, including all necessary scripts for replication, is available on GitHub. Detailed training logs, performance metrics, and saliency visualizations are provided in our Weights & Biases project.

10.2 Algorithm: Curriculum Learning

Below is the curriculum learning algorithm:

Algorithm 1 Curriculum Learning Method

Input: Dataset X, scoring function f, pacing function g_θ, number of mini-batches M.

1: Sort dataset X in ascending order according to f (easiest samples first)
2: Initialize an empty list $result$
3: **for** each training step $i = 1$ to M **do**
4: Determine subset size $size = g_\theta(i)$
5: Select top $size$ easiest samples from X
6: Uniformly sample mini-batch B'_i from the selected subset
7: Append B'_i to $result$
8: **end for**
9: **return** $result$ (sequence of mini-batches)

10.3 Statistical Tests

Table 1: Two Way ANOVA Results
Significance of Models and Curricula

Effect	Sum Sq.	df	F	p
Curriculum Type	172.25	1	6.15	0.016*
Attention Type	257.74	3	3.07	0.034*
Curriculum $\times Attention$	31.33	3	0.37	0.773
Residual	1821.18	65	–	–

Table 2: Comparing Model Types with ANOVA

Model	Accuracy (%)
No Attention (Baseline)	36.65 ś 1.74
CBAM	34.90 ś 1.95
MHA	34.79 ś 2.01
Self-Attention	31.44 ś 1.56

Table 3: Tukey HSD
(Multiple Comparisons, FWER = 0.05)

group1	group2	meandiff	p-adj	reject
cbam	mhattn	-0.11	0.99	False
cbam	no_attention	1.89	0.68	False
cbam	selfattn	-3.15	0.23	False
mhattn	no_attention	2.0	0.77	False
mhattn	selfattn	-3.04	0.45	False
no_attention	selfattn	-5.04	0.017	True

References

[1] Irwan Bello et al. "Attention Augmented Convolutional Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019), pp. 3285–3294. DOI: 10.1109/ICCV.2019.00338. URL: https://doi.org/10.1109/ICCV.2019.00338.

[2] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv preprint* (2020). eprint: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[3] Yoshua Bengio et al. "Curriculum Learning". In: *Proceedings of the 26th International Conference on Machine Learning (ICML)* (2009), pp. 41–48. DOI: 10.1145/1553374.1553380. URL: https://doi.org/10.1145/1553374.1553380.

[4] Guy Hacohen and Daphna Weinshall. "On The Power of Curriculum Learning in Training Deep Networks". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)* (2019), pp. 2535–2544. URL: http://proceedings.mlr.press/v97/hacohen19a.html.

[5] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778. DOI: 10.1109/CVPR.2016.90. URL: https://doi.org/10.1109/CVPR.2016.90.

[6] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Tech. rep. University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/cifar.html.

[7] Nikhil Kapila, Julian Glattki, and Tejas Rathi. "CNNtention: Can CNNs do better with Attention?" In: *arXiv preprint* (2024). eprint: 2412.11657. URL: https://arxiv.org/abs/2412.11657.

[8] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2017), pp. 5998–6008. URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[9] Yann LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.1.4.541. URL: https://doi.org/10.1162/neco.1989.1.4.541.

[10] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module". In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1. URL: https://doi.org/10.1007/978-3-030-01234-2_1.

[11] Tianhao Wu et al. "When Does Curriculum Learning Help? A Comprehensive Study". In: *arXiv preprint* (2020). eprint: 2003.08565. URL: https://arxiv.org/abs/2003.08565.

[12] Weights & Biases. *Bayesian Hyperparameter Optimization*. https://docs.wandb.ai/guides/sweeps. Accessed: March 05, 2025. 2023.