Modulated Sigmoid Contrastive Loss for Vision

Filip Basara

basarafilip@gmail.com

Objectives

Main objective. Explore if sigmoid contrastive loss (SCL) can perform on par with the softmax contrastive loss for self-supervised learning on images.

Side objectives:

- Enhance the sigmoid contrastive loss to amplify learning from harder negative pairs,
- Explore batch efficiency and modify the loss function to optimize the use of negative pairs.
- Solution – **confidence penalty (γ)** that transforms SCL into **Modulated SCL (MSCL)**.

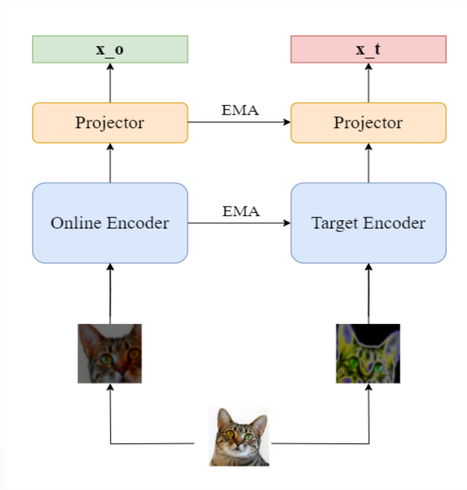

Figure 1 - High level overview of the architecture. Augmented images are processed by online and target encoders to generate 64 dimensional projections ($\mathbf{x_o}$ and $\mathbf{x_t}$), which are used in the loss function (see *Figure 2* on the right).

$$\mathbf{L} = \mathbf{x}_o \mathbf{x}_t^\top \exp(\tau) + b \quad (1)$$

$$\mathbf{Y} = 2 \cdot \mathbf{I}_n - 1 \quad (2)$$

$$\sigma(\mathbf{l}_i) = \sigma(Y_{ij} \cdot L_{ij}) = \frac{1}{1 + \exp(-Y_{ij} \cdot L_{ij})} \quad (3)$$

$$\mathcal{L}_{MSCL} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(1 - \sigma(Y_{ij} \cdot L_{ij})\right)^{\gamma} \log\left(\sigma(Y_{ij} \cdot L_{ij})\right) \quad (4)$$

Figure 2 - MSCL loss function. **L** represents the matrix of pair-wise dot products between normalized projections, with temperature τ and bias b. **Y** encodes labels with $\mathbf{I_n}$ as the identity matrix of size $n \times n$ (n = batch size). Probabilities $\boldsymbol{\sigma(\mathbf{l}_i)}$ are computed using the sigmoid function on the element-wise product of L and Y. When **γ (conf. pen.)** > 0 we are using MSCL; for γ=0 we default to the original SCL.

Approach & Results

- Sigmoid contrastive loss processes every pair independently, turning the learning problem into the binary classification on the dataset of all pair combinations.
- **Architecture -** an online and a target network, which share an identical structure but differ in parameter update mechanisms, ResNet-50 encoder, projector with output size of 64.
- **Multicrop Augmentation**. We augment each image to generate multiple views - two global views and four local views [3][4]. Global views are processed by both online and target networks, local views are processed only by the online network.
- **Experiments**. We use an identical architecture and augmentations for all runs. Quality of the representations is evaluated by training a linear classifier on frozen representations obtained from eval or test images.

Loss	Dataset	Epochs	Batch size	Pairs seen	Conf. Pen. (γ)	Acc. (%)
SCL	STL10	100	256	2941M	0	84.8
MSCL	STL10	100	256	2941M	1.0	85.6
MSCL (cos. sch.)	STL10	100	256	2941M	$1.0 \rightarrow 0.0$	**86.1**
MSCL (filtering)	STL10	100	256	**194M**	1.0	85.2
ReLIC [5]	STL10	100	256	/	/	85.4
MSCL	ImageNet-1k	50	128	8200M	1.0	59.2

Figure 3 - Results using variations of the (M)SCL. **MSCL (filtering)** is a version where we filtered all elements in the confidence penalty matrix below 0.05. **MSCL(cos. sch.)** denotes cosine scheduling of gamma from 1.0 to 0.0 over first 20,000 steps. Finally, we show the results on ImageNet-1k, where the model was trained during a period of one week on a single RTX4090 GPU with a reduced image size of 192. ReLIC[5] results from our reproduction are included for comparison.

Confidence Penalty - Amplfying Learning From Hard Negatives

- The confidence penalty serves as a modulation coefficient and encourages the model to focus more on examples for which it is less certain and prioritize hard examples.
- Improves the training stability and reduces the need to initialize training with optimal values for temperature and bias parameters. [2]
- Sparsity discussion - most negative pairs contribute very little to the loss.

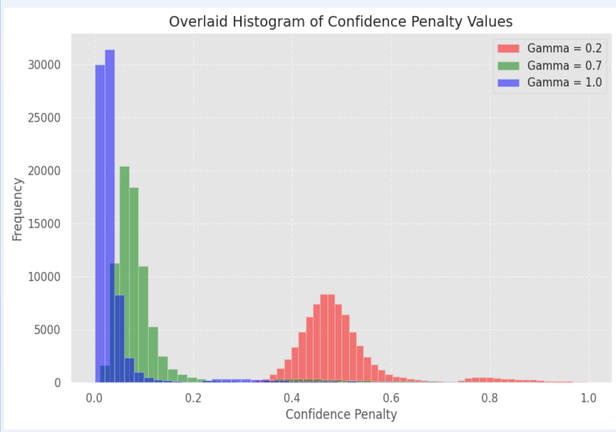

Figure 4 - Overlaid histogram of average confidence penalty values on STL-10. When γ = 1, majority of confidence penalty values are very close to zero. Using smaller γ values gradually moves the confidence penalty distribution towards 1. When γ is set to 0, confidence penalty has no effect and loss defaults to its original formulation from the SigLIP[1] paper.

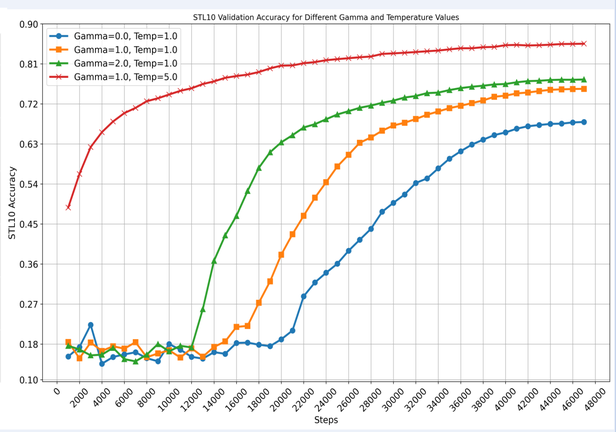

Figure 5 – Impact of γ and initial temp. on STL-10 accuracy. Higher γ values have a positive impact on training stability and convergence, especially when using low initial temperature - log(1). Confidence penalty stabilizes early training stages and makes the sigmoid contrastive loss more robust to temperature scaling.

Confidence Penalty Filtering & Exploring Batch Efficiency

- **Warmup**. Train using all pairs for only 2000 steps and use Modulated SCL (γ=1.0).
- **Confidence threshold**. Filter out all negative pairs with confidence penalty below 0.05. Always keep all positive pairs.
- Apply standard sigmoid contrastive loss (γ=0.0) **only on selected pairs.**
- Larger batch sizes - better results when applying hard masking.

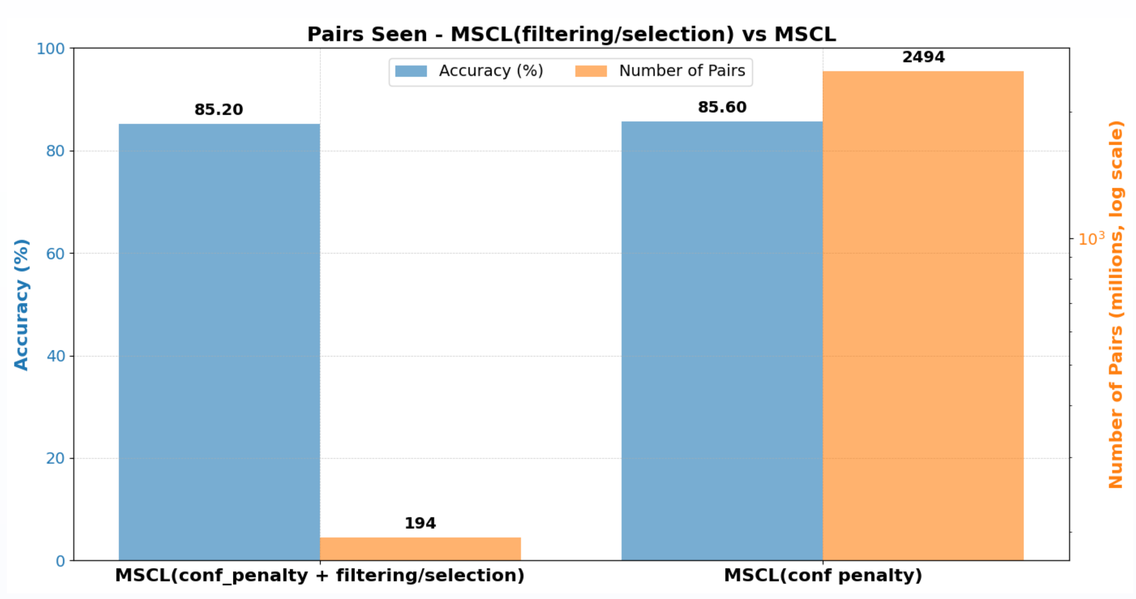

Figure 6 - Bar chart comparison that outlines a significant difference in number of pairs seen, while resulting in slightly lower accuracy. On the **left -** we can see the performance of the MSCL(filtering) version of the loss, where we filter out all negative pairs with confidence penalty below 0.05. On the **right** – we can see the performance of the standard (M)SCL loss, that takes into account all pairs. The MSCL(filtering) version sees more than **15 times fewer pairs**, but results in only 0.4% lower accuracy.

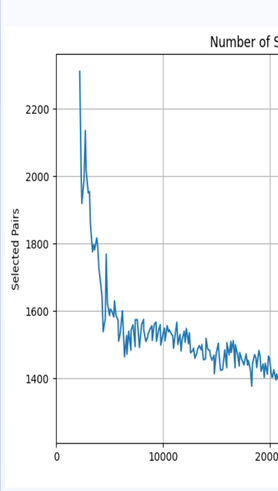

Figure 7 - Number of selected negatives across time steps. The plot doesn't show the number of selected negatives for first 1700 warmup steps, during which it uses all 65536 pairs. Number of selected negatives follow a decreasing trend, as model learns and becomes more confident. On average, the batch utilization is **less than 5%**.

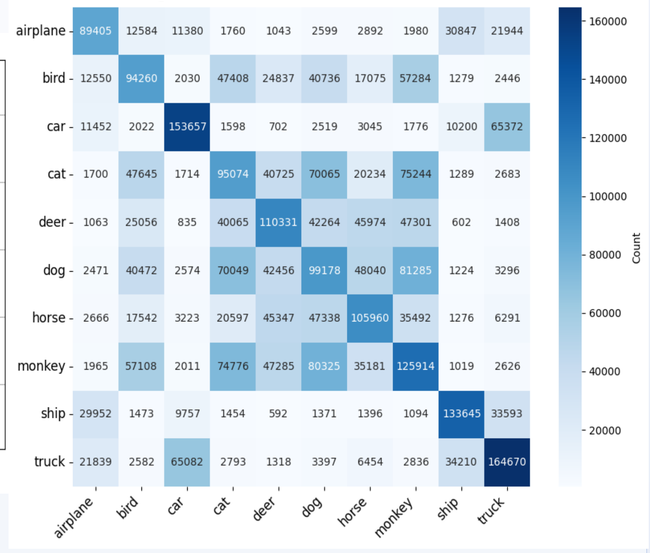

Figure 8 – Label correlation matrix for selected negative pairs. Intuitively, most impactful negatives are sampled from either the same or most similar classes. Also, some classes are easier to distinguish and are sampled less frequently, while others have higher inter and intra-class similarities and are sampled more.

Conclusions & Future Work

- (M)SCL works on par with the softmax contrastive loss – signals potentially wider adoption of sigmoid contrastive loss for vision SSL tasks.
- Initially utilizing all pairs and afterwards progressively selecting harder pairs could result in significantly faster convergence and reduced training costs (curriculum learning).
- Future work:
  - Run larger scale experiments on ImageNet1k using confidence penalty filtering, evaluate performance on other modalities,
  - Explore more efficient batch sampling – sample images to maximize batch utilization,
  - Do a more detailed analysis of augmentation and label impact.

References

[1] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[2] C. Lee, J. Chang, and J. yong Sohn, "Analysis of using sigmoid loss for contrastive learning," 2024.

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2021.

[4] N. Tomasev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic, "Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on Imagenet?," 2022.

[5] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," 2020.

Contact & Download

Email: basarafilip@gmail.com

Code: https://github.com/filipbasara0/sigmoid-contrastive-loss