Loss Function Discovery for Object Detection via Convergence-Simulation Driven Search

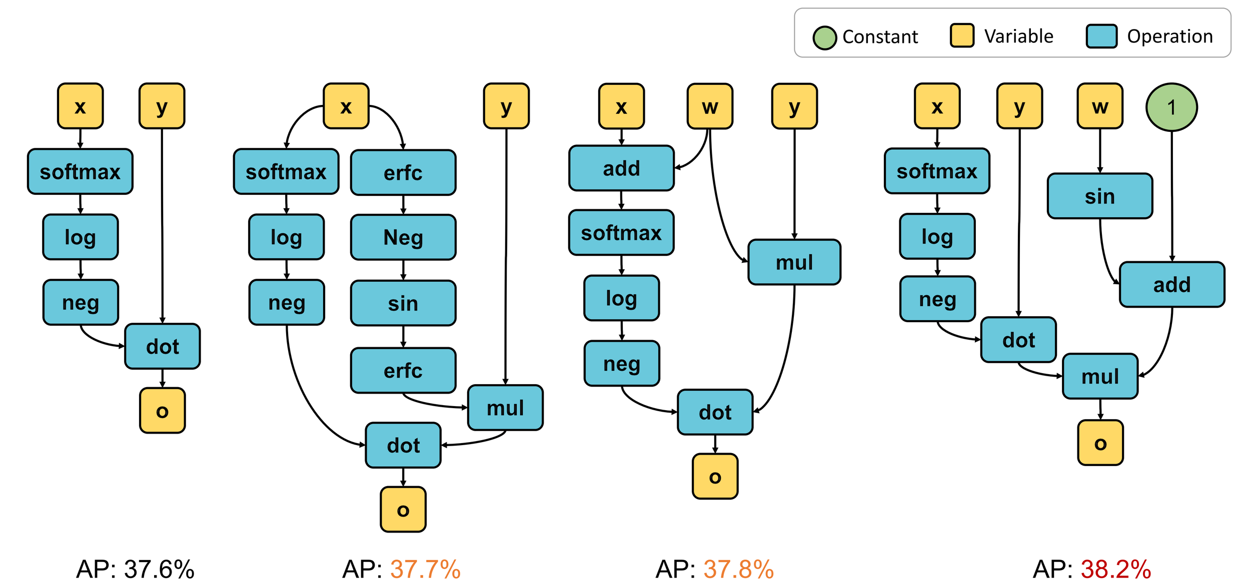
Peidong Liu[1*], Gengwei Zhang[2*], Bochao Wang[3], Hang Xu[3], Xiaodan Liang[2], Yong Jiang[1], Zhenguo Li[3]
[1]Tsinghua University [2]Sun Yat-Sen University [3]Huawei Noah's Ark Lab

Arxiv Link Github Code Homepage: Peidong Liu

Introduction

Inspired by the recent progress in network architecture search, it is interesting to explore the possibility of discovering new loss function formulations via directly searching the primitive operation combinations. So that the learned losses not only fit for diverse object detection challenges to alleviate huge human efforts, but also have better alignment with evaluation metric and good mathematical convergence property. Beyond the previous auto-loss works on face recognition and image classification, our work makes the first attempt to discover new loss functions for the challenging object detection from primitive operation levels and finds the searched losses are insightful. We propose an effective convergence-simulation driven evolutionary search algorithm, called **CSE-Autoloss**, for speeding up the search progress by regularizing the mathematical rationality of loss candidates via convergence simulation modules: convergence property verification and model optimization simulation.

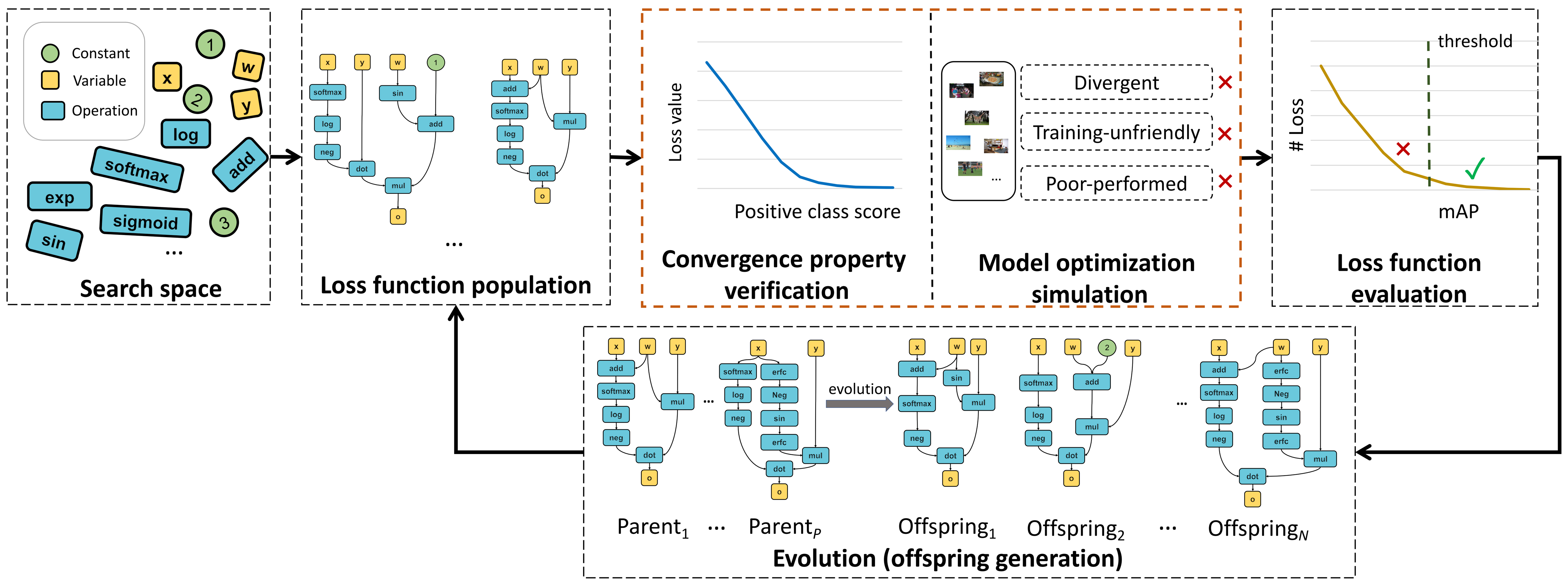
Motivation



AP: 37.6% (a) Cross-entropy loss AP: 37.7% (b) Candidate1 AP: 37.8% (c) Candidate2 AP: 38.2% (d) CSE-Autoloss-A_{cls}

CSE-Autoloss discovers distinct loss formulas with comparable performance with the (a) Cross-Entropy (CE) loss, such as (b) and (c) in the figure above. The best-searched loss (d), named CSE-Autoloss-A_{cls}, outperformed CE loss by a large margin.

Proposed Pipeline: CSE-Autoloss



An overview of the proposed CSE-Autoloss pipeline is illustrated in the figure above. Our CSE-AutoLoss first generates a large amount of candidate loss functions by assembling formulations based on the well-designed search space. Then CSE-Autoloss filters out more than 99% divergent, training-unfriendly, or poor-performed loss candidates by evaluating them with the proposed convergence-simulation modules, i.e. convergence property verification and model optimization simulation, which verifies mathematical property and optimization quality. After that, K top-performing loss functions are selected to evaluate on the proxy task to further obtain top-P loss candidates as parents to derive offspring for the next generation.

Search Space Design

Input Nodes. For classification branch, to better motivate the loss to align with the evaluation metric (i.e. AP), we introduce IoU between groundtruth and prediction into the loss formula. **Primitive Operators.** The primitive operators are listed as below: $-x$, e^x, $\log(x)$, $|x|$, $\sqrt{x}$, $\mathrm{softmax}(x)$, $\mathrm{softplus}(x)$, $\mathrm{sigmoid}(x)$, $\mathrm{gd}(x)$, $\mathrm{alf}(x)$, $\mathrm{erf}(x)$, $\mathrm{erfc}(x)$, $\tanh(x)$, $\mathrm{relu}(x)$, $\sin(x)$, $\cos(x)$, x_1+x_2, x_1-x_2, $x_1\times x_2$, $\frac{x_1}{x_2+\epsilon}$, $\mathrm{dot}(\mathrm{x}_1,\mathrm{x}_2)$. ϵ is a small value to avoid zero division, $\mathrm{softplus}(x)=\ln(1+e^x)$, $\mathrm{sigmoid}(x)=\frac{1}{1+e^{-x}}$ is the sigmoid function. To enlarge the search space, more S-type curves and their variants are included, i.e. $\mathrm{gd}(x)=2\arctan(\tanh(\frac{x}{2}))$, $\mathrm{alf}(x)=\frac{x}{\sqrt{1+x^2}}$, $\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function, $\mathrm{erfc}(x)=1-\mathrm{erf}(x)$.

Convergence Simulation Modules

Convergence property verification. Classification loss should meet basic mathematical properties: Monotonicity and Convergence. For regression branch, consistency of loss value and distance between the predicted and target bounding box, named distance-loss consistency, should be confirmed.
Model Optimization Simulation. We train and test the loss candidates on a small dataset D_{verify}, which is constructed by sampling only one image randomly from each category on benchmark dataset like COCO, to estimate the optimization quality. Then we apply AP performance on D_{verify} of the top loss in the previous generation as a threshold to filter out divergent, training-unfriendly, and poor-performed loss candidates.

Best-discovered Losses

As the following formulations show, we name the best loss combination searched with Faster R-CNN R50 and FCOS R50 as CSE-Autoloss-A and CSE-Autoloss-B respectively, with the subscript *cls* and *reg* indicating classification and regression branch.

$$\text{CSE-Autoloss-A}_{\text{cls}}(x,y,w)=-\mathrm{dot}((1+\sin(w))y,\log(\mathrm{softmax}(x))),$$

$$\text{CSE-Autoloss-A}_{\text{reg}}(i,u,e)=(1-\frac{i}{u})+(1-\frac{i+2}{e}),$$

$$\text{CSE-Autoloss-B}_{\text{cls}}(x,y,w)=-[wy(1+\mathrm{erf}(\sigma(1-y)))\log\sigma(x)+(\mathrm{gd}(x)-wy)(\sigma(x)-wy)\log(1-\sigma(x))],$$

$$\text{CSE-Autoloss-B}_{\text{reg}}(i,u,e)=\frac{3eu+12e+3i+3u+18}{-3eu+iu+u^2-15e+5i+5u}.$$

Generalization of Best-discovered Losses on Architectures and Datasets

Table 1: Detection results on COCO val.

Detector	Loss	AP (%)
Faster R-CNN R50	CE + L1	37.4
	CSE-Autoloss-A	**38.5**$^{+1.1}$
Faster R-CNN R101	CE + L1	39.4
	CSE-Autoloss-A	**40.2**$^{+0.8}$
Cascade R-CNN R50	CE + Smooth L1	40.3
	CSE-Autoloss-A	**40.5**$^{+0.2}$
Mask R-CNN R50	CE + Smooth L1	38.2
	CSE-Autoloss-A	**39.1**$^{+0.9}$
FCOS R50	FL + GIoU	38.8
	CSE-Autoloss-B	**39.6**$^{+0.8}$
ATSS R50	FL + GIoU	40.0
	CSE-Autoloss-B	**40.5**$^{+0.5}$

The quantitative results about the gain brought by the searched loss under the same hyper-parameters for multiple popular two-stage and one-stage detectors are shown in Table 1. Results on Faster R-CNN R50 and FCOS R50 indicate the generalization of CSE-Autoloss across detectors and the best-searched loss combination is capable of stimulating the potential of detectors by a large margin. We respectively apply CSE-Autoloss-A on other two-stage models, and CSE-Autoloss-B on ATSS, which represent the consistent gain brought by the effective searched loss without additional overhead. We further conduct evaluation experiments for Faster R-CNN R50 on VOC and BDD to validate the loss transferability across datasets. Results are displayed in Table 2, which indicate the best-searched loss combination enables the detectors to converge well on datasets with different object distributions and domains.

Table 2: Detection results on PASCAL VOC 2007 test and BDD val.

Loss	VOC mAP (%)	BDD AP (%)
CE + L1	79.5	36.5
CE + GIoU	79.6$^{+0.1}$	36.6$^{+0.1}$
CSE-Autoloss-A	**80.4**$^{+0.9}$	**37.3**$^{+0.8}$

Effectiveness of CSE-Autoloss

The table below shows efficiency improvement with progressive convergence-simulation modules on searching for Faster R-CNN classification branch. The proposed modules filter out 99% loss candidates, which enhances the efficiency to a large extent.

Convergence property verification	Model optimization simulation	#Evaluated loss
		5×10^3
✓		7×10^2
✓	✓	50

We compare the efficiency between different search algorithms including random search, vanilla evolution search, and CSE-Autoloss. Table below shows that the evolution search is much better than random search. But due to the high sparsity of the action space, the vanilla evolution search requires hundreds of wall-clock hours. CSE-Autoloss speeds up the search process by 20x because of the effective convergence-simulation modules and discovers well-performed loss combinations in around 1.5 wall-clock days.

Search algorithm	Search branch	Another branch	#Evaluated loss	Wall-clock hours	AP (%)
Random search	Classification	GIoU	1×10^4	8×10^3	37.8
Evolution search	Classification	GIoU	5×10^3	5×10^2	38.2
CSE-Autoloss	Classification	GIoU	50	26	38.2
CSE-Autoloss	Regression	CE	15	8	37.9

Conclusion

In this work, we propose a convergence-simulation driven evolutionary search pipeline, named CSE-Autoloss, for loss function search on object detection. It speeds up the search process largely by regularizing the mathematical property and optimization quality and discovers well-performed loss functions efficiently without compromising the accuracy. We conduct search experiments on both two-stage and one-stage detectors and validate the best-searched loss functions on different architectures across datasets, which shows the effectiveness and generalization of both CSE-Autoloss and the best-discovered loss functions.