RandAugment: Practical automated data augmentation with a reduced search space

Abstract

data augmentation
obstacle to a large-scale adoption of these methods is a separate search phase which increases the training complexity and may substantially increase the computational cost
due to the separate search phase, these approaches are unable to adjust the regularization strength based on model or dataset size

Automated augmentation policies are often found by training small models on small datasets and subsequently applied to train larger models.

RandAugment
significantly reduced search space which allows it to be trained on the target task with no need for a separate proxy task
due to the parameterization, the regularization strength may be tailored to different model and dataset sizes.
can be used uniformly across different tasks and datasets and works out of the box, matching or surpassing all previous automated augmentation approaches
due to its interpretable hyperparameter, RandAugment may be used to investigate the role of data augmentation with varying model and dataset size

Introduction

문제
data augmentation methods require expertise
manual work to design policies that capture prior knowledge in each domain
-> difficult to extend existing data augmentation methods to other applications and domains.

neural architecture search
require a separate optimization procedure, which significantly increases the computational cost and complexity of training a machine learning model.

RandAugment
does not require a separate search
dramatically reduce the search space for data augmentation
simple grid search is sufficient to find a data augmentation policy

contributions 세 개
optimal strength of a data augmentation depends on the model size and training set size
-> separate optimization of an augmentation policy on a smaller proxy task may be sub-optimal

a vastly simplified search space for data augmentation containing 2 interpretable hyperparameters

state-ofthe-art results

Related Work

horizontal flips and random cropping or translations
elastic distortions across scale, position, and orientation
randomly erase or add noise to patches
Mixup
Object-centric cropping

combining different operations
Smart Augmentation

Bayesian approach
use transformations
GAN

AutoAugment
choose a sequence of operations as well as their probability of application and magnitude
stochasticity at multiple levels:
1) for every image in every minibatch, a sub-policy is chosen with uniform probability.
2) operations in each sub-policy has an associated probability of application.
3) Some operations have stochasticity over direction.

this work aims to eliminate the search phase on a separate proxy task completely

recent improvements to searching over data augmentation policies
Population Based Augmentation -> have a fixed magnitude schedule
Fast AutoAugment -> trained for density matching leads to improved generalization accuracy

Methods

primary goal - remove the need for a separate search phase on a proxy task.
a separate search phase significantly complicates training and is computationally expensive.
sub-optimal results

fold the parameters for the data augmentation strategy into the hyper-parameters for training a model

- `identity`
- `rotate`
- `posterize`
- `sharpness`
- `translate-x`
- `autoContrast`
- `solarize`
- `contrast`
- `shear-x`
- `translate-y`
- `equalize`
- `color`
- `brightness`
- `shear-y`

we replace the learned policies and probabilities for applying each transformation with a parameter-free procedure of always selecting a transformation with uniform probability 1/K
-> reduce the parameter space but still maintain image diversity

employ the same linear scale for indicating the strength of each transformation.

the learned magnitude for each transformation follows a similar schedule during training
postulate that a single global distortion M may suffice for parameterizing all transformations

four methods for the schedule of M during training

Magnitude Method	Accuracy
Random Magnitude	97.3

Constant Magnitude	97.2
Linearly Increasing Magnitude	97.2
Random Magnitude with Increasing Upper Bound	97.3

```
transforms = [
'Identity', 'AutoContrast', 'Equalize',
'Rotate', 'Solarize', 'Color', 'Posterize',
'Contrast', 'Brightness', 'Sharpness',
'ShearX', 'ShearY', 'TranslateX', 'TranslateY']

def randaugment(N, M):
"""Generate a set of distortions.

  Args:
    N: Number of augmentation transformations to
        apply sequentially.
    M: Magnitude for all the transformations.
"""

  sampled_ops = np.random.choice(transforms, N)
  return [(op, M) for op in sampled_ops]
```

naive grid search is quite effective

Results

Systematic failures of a separate proxy task

A central premise of learned data augmentation is to construct a small, proxy task that may be reflective of a larger task

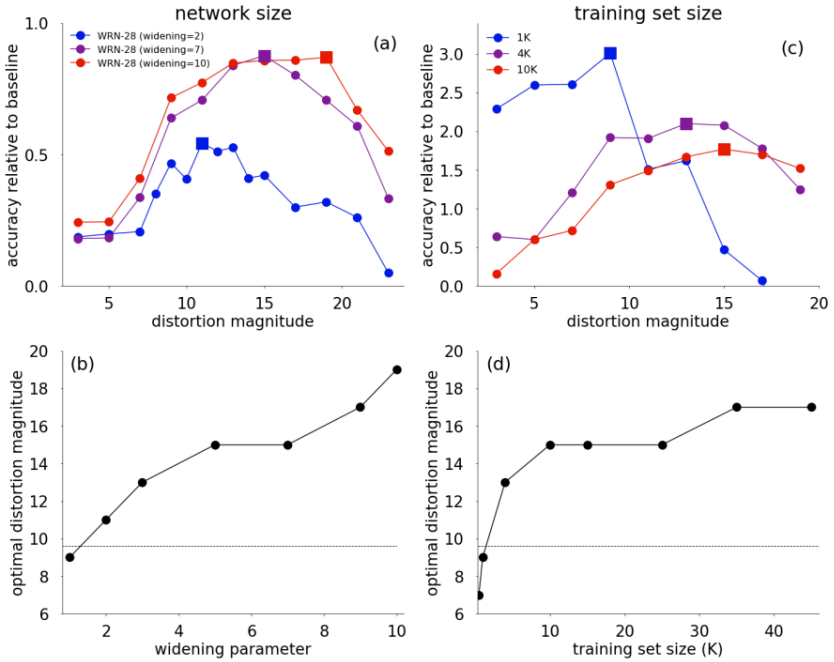
-> unclear if this assumption is overly stringent and may lead to sub-optimal data augmentation policies

small proxy task는 좋지 않다는 가설 증명

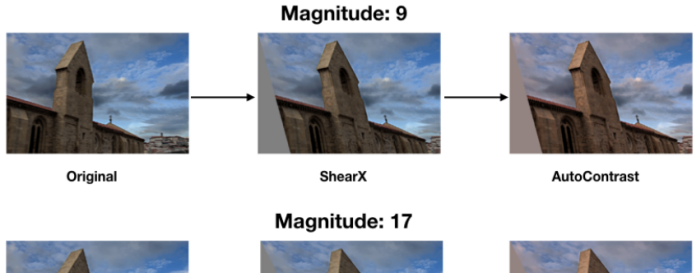
model size and dataset size

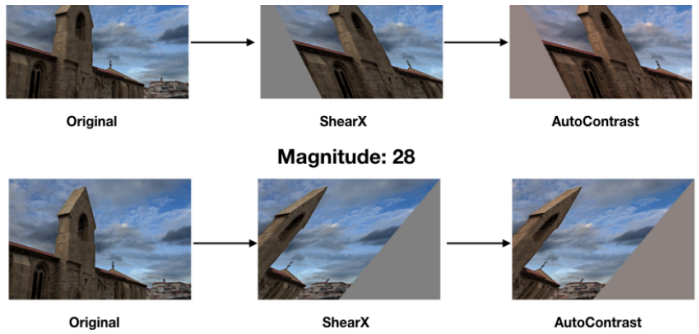
	baseline	PBA	Fast AA	AA	RA
CIFAR-10					
Wide-ResNet-28-2	94.9	-	-	**95.9**	95.8
Wide-ResNet-28-10	96.1	**97.4**	97.3	**97.4**	97.3
Shake-Shake	97.1	**98.0**	**98.0**	**98.0**	**98.0**
PyramidNet	97.3	**98.5**	98.3	**98.5**	**98.5**
CIFAR-100					
Wide-ResNet-28-2	75.4	-	-	**78.5**	78.3
Wide-ResNet-28-10	81.2	**83.3**	82.7	82.9	**83.3**
SVHN (core set)					
Wide-ResNet-28-2	96.7	-	-	98.0	**98.3**

Wide-ResNet-28-2	96.7	-	-	98.0	**98.3**
Wide-ResNet-28-10	96.9	-	-	98.1	**98.3**
SVHN					
Wide-ResNet-28-2	98.2	-	-	**98.7**	**98.7**
Wide-ResNet-28-10	98.5	98.9	98.8	98.9	**99.0**



Namely, larger networks demand larger data distortions for regularization





models trained on smaller training sets may gain more improvement from data augmentation
optimal distortion magnitude is larger for models that are trained on larger datasets
may disagree with the expectation that smaller datasets require stronger regularization.

the optimal distortion magnitude increases monotonically with training set size
One hypothesis for this counter-intuitive behavior is that aggressive data augmentation leads to a low signal-to-noise ratio in small datasets
-> highlights the need for increasing the strength of data augmentation on larger datasets
the shortcomings of optimizing learned augmentation policies on a proxy task comprised of a subset of the training data.

a small proxy task may provide a sub-optimal indicator of performance on a larger task

unified optimization of the model weights and data augmentation policy
merely searching for a shared distortion magnitude M across all transformations may provide sufficient gains that exceed learned optimization methods
merely sampling a few distortion magnitudes is sufficient to achieve good results

CIFAR

SVHN

ImageNet
CIFAR SVHN에는 통하지만 ImageNet에는 통하지 않는 경우도 있음
AutoAugment도 실패
하지만 ImageNet은 거의 모델 하나를 새로 짜는 정도의 개선 보임
without incurring additional computational cost at inference time

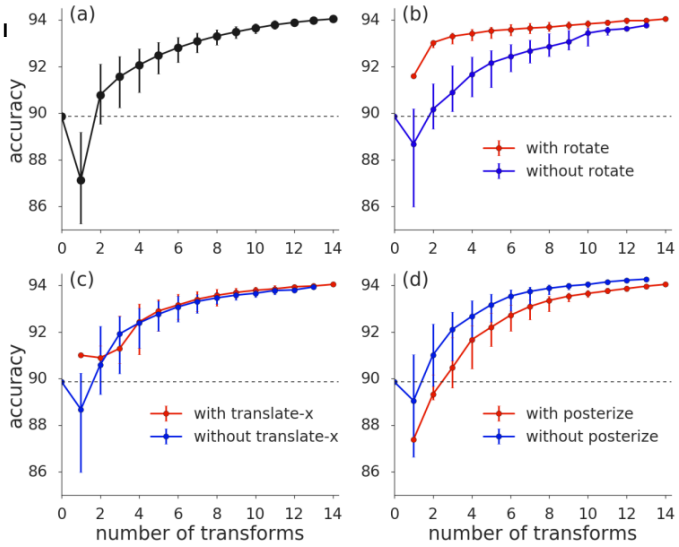
COCO

model	augmentation	mAP	search space
	Baseline	38.8	0

ResNet-101	AutoAugment	**40.4**	10^{34}
	RandAugment	40.1	10^2
	Baseline	39.9	0
ResNet-200	AutoAugment	**42.1**	10^{34}
	RandAugment	41.9	10^2

AutoAugment expended ~15K GPU hours for search, where as RandAugment was tuned by on merely 6 values of the hyperparameters

future - expand the transformation library to include bounding box specific transformation to potentially improve



average improvement in validation accuracy for each transformation when they are added to a random subset of transformations

transformation	Δ (%)	transformation	Δ (%)
rotate	**1.3**	shear-x	0.9
shear-y	0.9	translate-y	0.4
translate-x	0.4	autoContrast	0.1
sharpness	0.1	identity	0.1
contrast	0.0	color	0.0
brightness	0.0	equalize	-0.0
solarize	-0.1	posterize	-0.3

The transformation rotate is most helpful on average

Learning the probabilities for selecting image transformations

equal probability

learning probability?

	baseline	AA	RA	+ 1st
Reduced CIFAR-10				
Wide-ResNet-28-2	82.0	**85.6**	85.3	85.5
Wide-ResNet-28-10	83.5	**87.7**	86.8	87.4
CIFAR-10				
Wide-ResNet-28-2	94.9	95.9	95.8	**96.1**
Wide-ResNet-28-10	96.1	**97.4**	97.3	**97.4**

we take these results to indicate that learning the probabilities through density matching may improve the performance on small-scale tasks and reserve explorations to larger-scale tasks for the future

Discussion

previous methods of learned augmentation suffers from systematic drawbacks

we propose a simple parameterization for targeting augmentation to particular model and dataset sizes

scaling learned data augmentation to larger dataset and models have been a notable obstacle

-> rand는 좋다

open question remains how this method may improve model robustness or semi-supervised learning