The Augmented Image Prior: Distilling 1000 Classes by Extrapolating from a Single Image

Желтовская Юлия

Plan

- Questions to research
- Problem statement
- Method
- Experiments and results
- Strengths & Weaknesses

Questions to research

- What exactly is required for arriving at semantic visual representations from random weights?
- What neural networks know about the world from their training distribution?

↓

How well neural networks trained from a single datum can extrapolate to semantic classes?

Problem statement

How well neural networks trained from a single datum can extrapolate to semantic classes?

↓

Training student-model via knowledge distillation **without** pretrained teacher's **source dataset**

Method

Distilling 1000 Classes by Extrapolating from a Single Image

Method

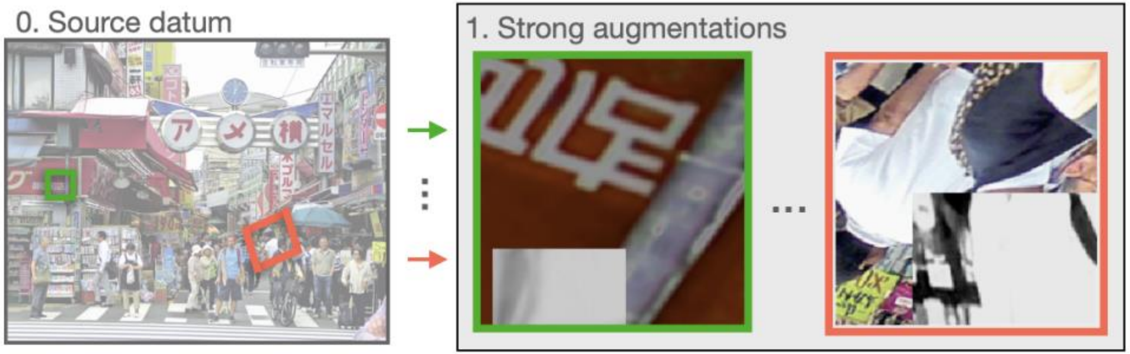
Distilling 1000 Classes by Extrapolating from a Single Image

0. Source datum

1 dense image

Method

Distilling 1000 Classes by Extrapolating from a Single Image

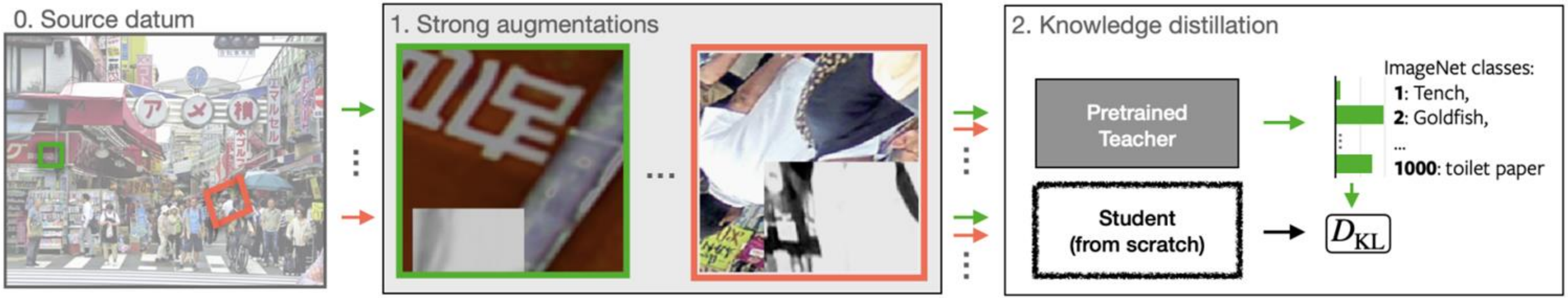


1 dense image

1000 augmented
pathes-images

Method

Distilling 1000 Classes by Extrapolating from a Single Image



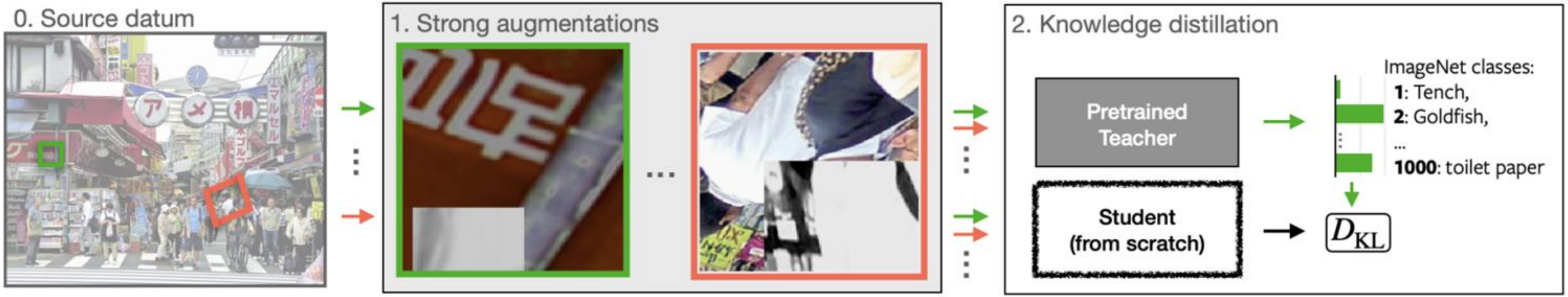
1 dense image

1000 augmented pathes-images

Teacher-student knowledge distillation

Method

Single-image distillation framework:



1 dense image

1000 augmented pathes-images

Teacher-student knowledge distillation

What to conclude?

“Augmented image prior” hypothesis

Within the space of all possible images $\mathcal{I}$, a single real image $x \in \mathcal{I}$ and its augmentations $\mathcal{A}(x)$ can provide sufficient diversity for extrapolating to semantic categories in real images.

Back to method

Single-image distillation framework

Dataset generation

Knowledge distillation

Dataset generation

1. Select good dense single image

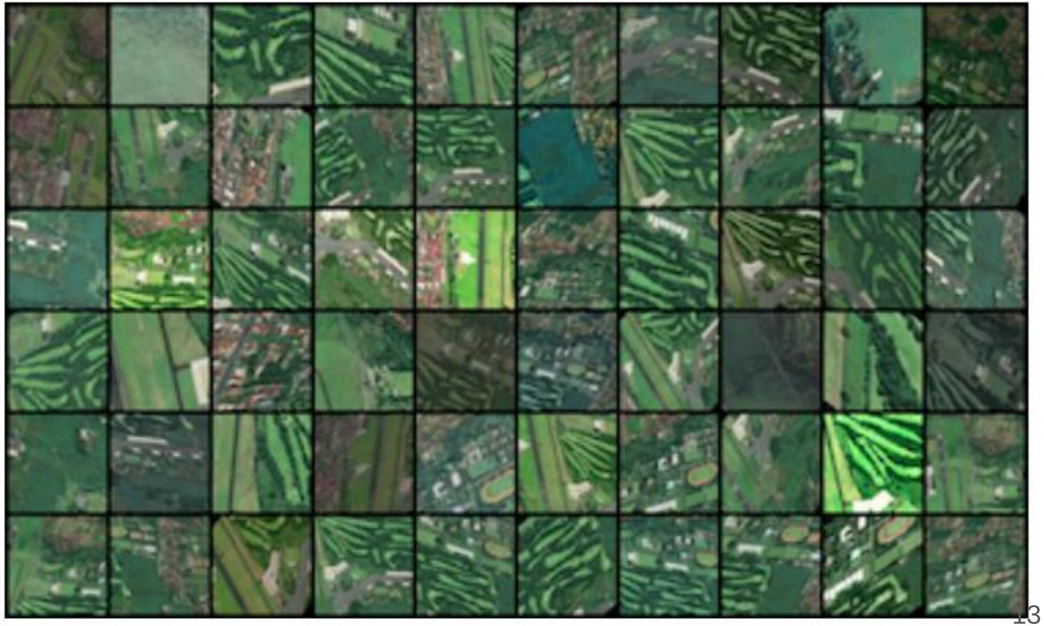
(d) The "City" Image. Size: 2,560x1,920, JPEG: 1.9MB.

(e) The "Animals" Image. Size: 1,300x600, JPEG: 267KB.

Dataset generation

2. “Patchify” a single-image using augmentations

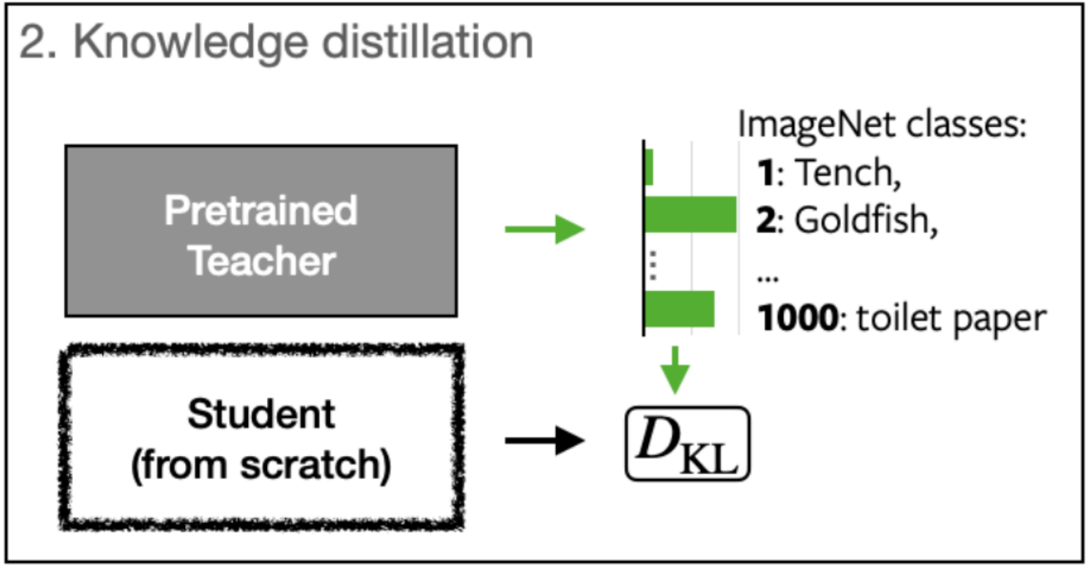
(a) Source image



(b) Patches

Knowledge distillation

Transfer the knowledge of a pretrained teacher to a lower capacity student model



Knowledge distillation

“Distribution-matching” **objective** that aims to mimic the teacher’s output

Knowledge distillation

"Distribution-matching" objective that aims to mimic the teacher's output:

KL divergence between the student output and the teacher's output

Knowledge distillation

"Distribution-matching" objective that aims to mimic the teacher's output: **KL divergence** between the student output and the teacher's output

$$\mathcal{L}_{\mathrm{KL}} = \sum_{c \in \mathcal{C}} -p_c^t \log p_c^s + p_c^t \log p_c^t$$

c are the teachers' classes

student output p^s
teacher's output p^t

Knowledge distillation

"Distribution-matching" objective that aims to mimic the teacher's output:

KL divergence between the student output and the teacher's output

$$\mathcal{L}_{\text{KL}} = \sum_{c \in \mathcal{C}} -p_c^t \log p_c^s + p_c^t \log p_c^t$$

c are the teachers' classes

student output p^s

teacher's output p^t

$$p = \texttt{softmax}(l/\tau)$$

logits l

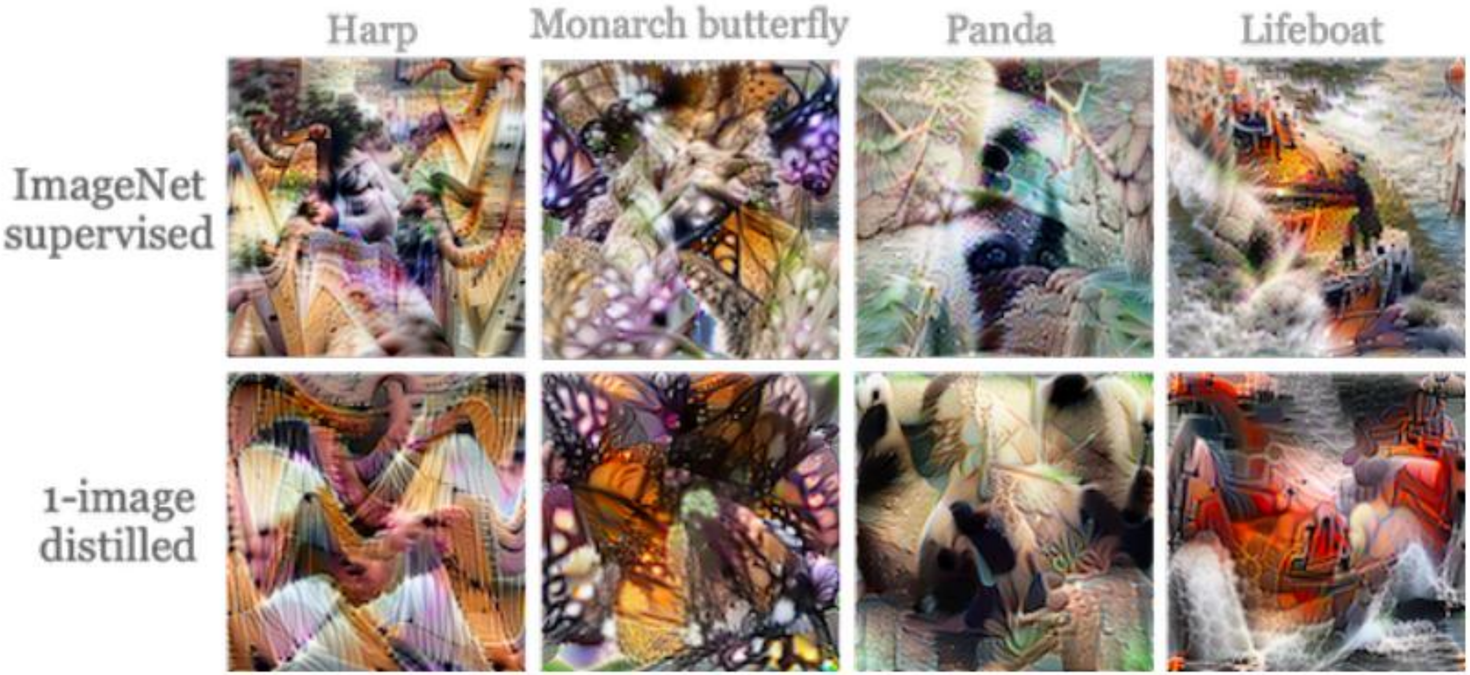
temperature τ

Experiments and results

	Teacher	Acc.	Student	Acc.	Distillation Full	Distillation Ours	$\Delta < 5\%$
CIFAR10	VGG-19	93.28	VGG-16	92.42	92.84	92.14	✓
	ResNet-56	93.77	ResNet-20	92.52	92.29	90.70	✓
	WideR40-4	95.42	WideR16-4	95.20	95.00	93.32	✓
	WideR40-4	95.42	WideR40-4	95.42	94.36	94.14	✓
	WideR16-4	95.20	WideR40-4	95.42	94.30	94.02	✓
CIFAR100	VGG-19	70.79	VGG-16	73.26	71.19	58.66	✗
	ResNet-56	70.99	ResNet-20	65.74	67.04	52.43	✗
	WideR40-4	78.14	WideR16-4	75.56	76.26	68.69	✗
	WideR40-4	78.14	WideR40-4	78.14	75.54	73.80	✓
	WideR16-4	78.14	WideR40-4	75.56	76.29	74.08	✓

Student accuracy when **distilling** with full training set vs our 1-image dataset

Visualizing neurons



Visualizing neurons

Noise for source datum

(a) The "Noise" Image. From uniform noise [0,255].
Size: 2,560x1,920, PNG: 16.3MB.

Distillation dataset		**Accuracy**	
Image	# Pixels	C10	C100
"Noise"	4.9M	69.30	19.50
"Universe"	4.8M	88.18	39.68
"Bridge"	1.1M	92.24	57.87
"City"	4.9M	93.13	64.85
"Animals"	2.8M	93.28	66.12

Choice of source image content is crucial

Distillation on synthetic data

Method outperforms several synthetic datasets

Data	**C10**
CIFAR-10	92.61
Fractals	33.26
StyleGAN	83.42
ZeroSKD	86.60
Ours	89.27

Strengths & Weaknesses

Strengths:

- Interesting and surprising results
- Paper covers a lot of datasets, architectures, and modalities

Weaknesses:

- We still need a “good” teacher which should be trained on a large dataset in this domain
- “Single image” is large high-resolution image with lots of detail, not the 32x32 CIFAR image

Sources

The Augmented Image Prior: Distilling 1000 Classes By Extrapolating From a Single Image: https://arxiv.org/pdf/2112.00725.pdf