Lossy compression for lossless prediction

EECS Seminar: Advanced Topics in Machine Learning

Romain Graux

March 15, 2022

Motivation

10^{21} - 10^{23} bytes data collected per year

Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Motivation

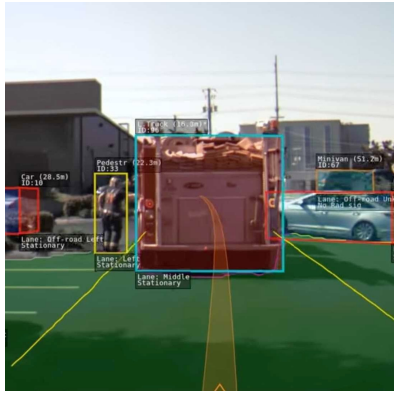
10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Motivation

10^{21} - 10^{23} bytes data collected per year

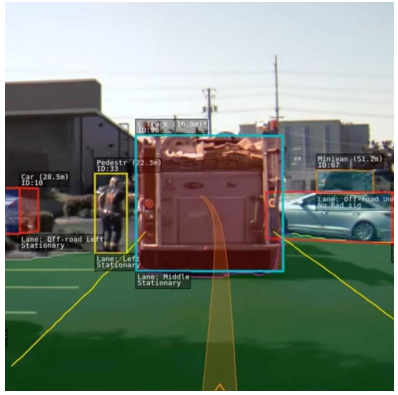
$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.



Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.



Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

- Stores too much not needed information
- Does not ensure good task performance

Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

- Stores too much not needed information
- Does not ensure good task performance

Source

Motivation

10^{21} - 10^{23} bytes data collected per year

$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

- Stores too much not needed information
- Does not ensure good task performance

Source

High bitrate

Motivation

10^{21} - 10^{23} bytes data collected per year

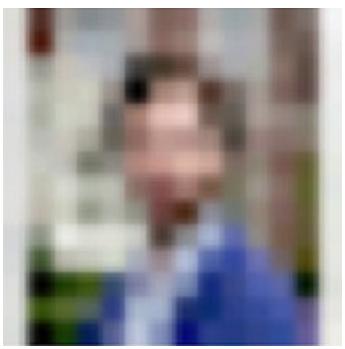
$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

- Stores too much not needed information
- Does not ensure good task performance

Source

High bitrate



Low bitrate

Motivation

10^{21} - 10^{23} bytes data collected per year

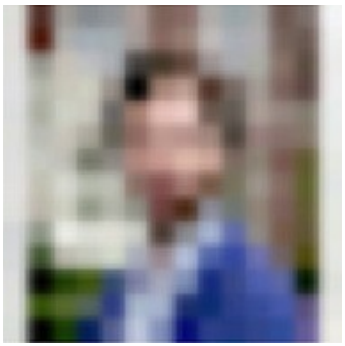
$\rightarrow$ But most data is processed by algorithms performing **downstream tasks**.

Yet current compressors optimize high **perceptual** fidelity

- Stores too much not needed information
- Does not ensure good task performance

Source

High bitrate



Low bitrate

Desired

What they designed

They designed a **task-centric** distortion that ensures good downstream performance

What they designed

They designed a **task-centric** distortion that ensures good downstream performance

- Characterize minimum bit-rate to ensure high performance on desired tasks;

What they designed

They designed a **task-centric** distortion that ensures good downstream performance

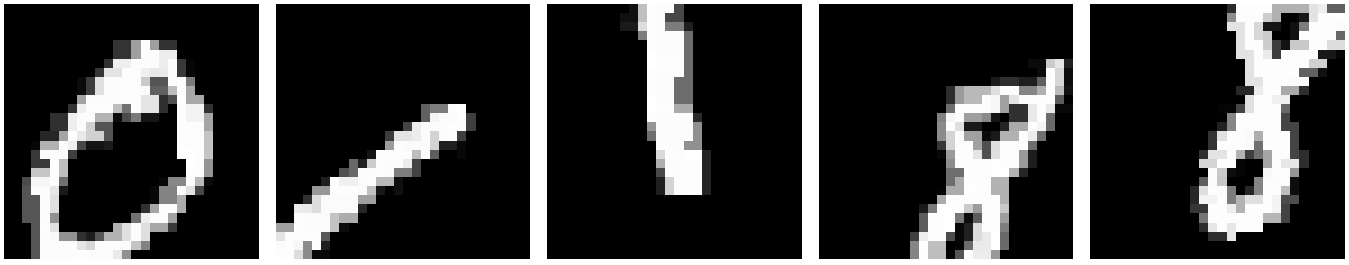
- Characterize minimum bit-rate to ensure high performance on desired tasks;
- Derive unsupervised objectives for training **task-centric** compressors;

What they designed

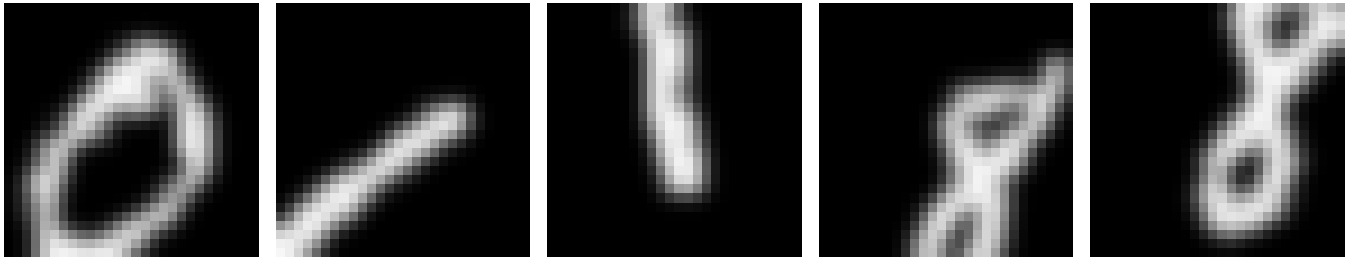
They designed a **task-centric** distortion that ensures good downstream performance

- Characterize minimum bit-rate to ensure high performance on desired tasks;
- Derive unsupervised objectives for training **task-centric** compressors;
- $> 1000x$ compression gains on Imagenet compared to JPEG (see Slide 5).

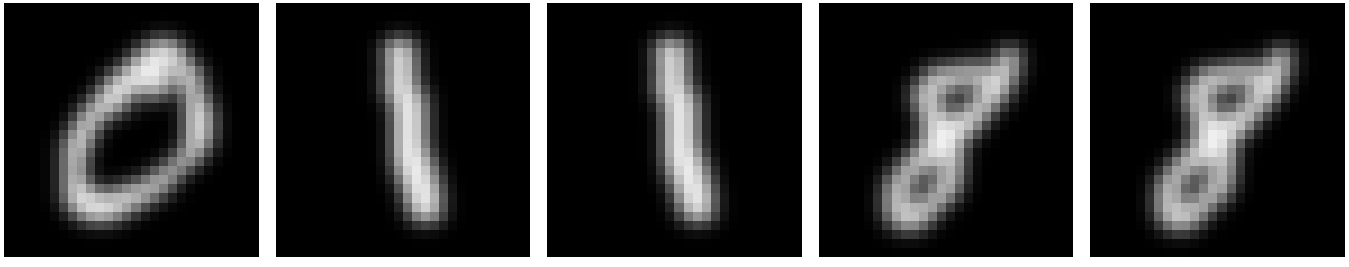
Intuition: Augmented MNIST



Source: Augmented MNIST

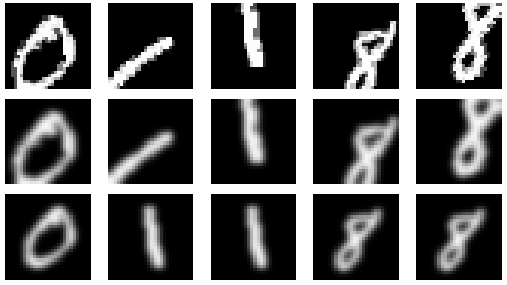


Standard neural compressor: 130 bit-rate



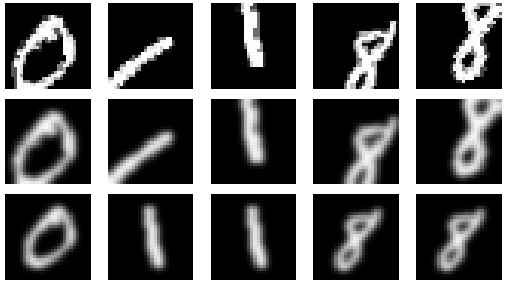
Their neural compressor: 48 bit-rate

Intuition: Augmented MNIST



Prototypical digit ensures → high downstream performance
→ good compression rate

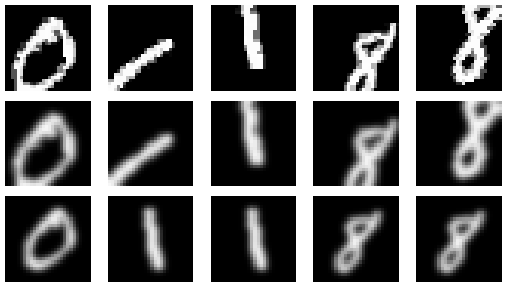
Intuition: Augmented MNIST



Prototypical digit ensures
- $\rightarrow$ high downstream performance
- $\rightarrow$ good compression rate

Why not sending directly the labels?

Intuition: Augmented MNIST

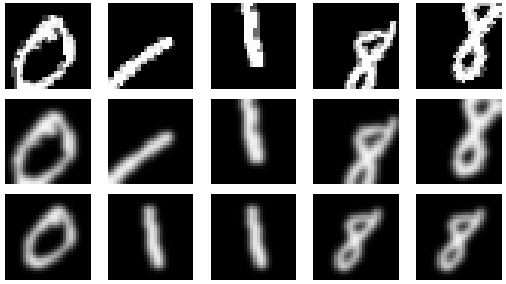


Prototypical digit ensures $\rightarrow$ high downstream performance
$\rightarrow$ good compression rate

Why not sending directly the labels?

- Might be interested in multiple downstream tasks;

Intuition: Augmented MNIST

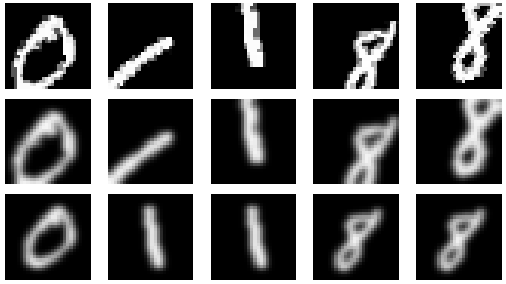


Prototypical digit ensures $\rightarrow$ high downstream performance
$\rightarrow$ good compression rate

Why not sending directly the labels?

- Might be interested in multiple downstream tasks;
- Would require knowing tasks of interest at compression time.

Intuition: Augmented MNIST



Prototypical digit ensures → high downstream performance
→ good compression rate

Why not sending directly the labels?

- Might be interested in multiple downstream tasks;
- Would require knowing tasks of interest at compression time.

→ The objective is **unsupervised**

Performance