Sample Compression Hypernetworks: From Generalization Bounds to Meta-Learning

Benjamin Leblanc[1], Mathieu Bazinet[1], Nathaniel D'Amours[1], Alexandre Drouin[1,2], Pascal Germain[1]
[1]Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[2]ServiceNow Research, Montréal, Canada

Contributions

1. We generalize sample compression bounds to continuous messages;
2. We present an original approach to sample compression by learning the reconstruction function, seeing it as a hypernetwork;
3. We propose an algorithm to meta-learn the generation of predictors with sample compression generalization guarantees.

Sample compression

The setting

- A data-generating distribution $\mathcal{D}$ over an instance space $\mathcal{X} \times \mathcal{Y}$;
- A dataset $S = \{(\mathbf{x}_j, y_j)\}_{j=1}^m \sim \mathcal{D}^m$;
- A predictor $h : \mathcal{X} \to \mathcal{Y}$ and a learning algorithm $A(S) \mapsto h$;
- The empirical loss $\widehat{\mathcal{L}}_S(h) = \frac{1}{m}\sum_{j=1}^m \ell(h(\mathbf{x}_j), y_j)$, with $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$;
- The generalization loss $\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell(h(\mathbf{x}), y)\right]$.

The reconstruction function

In the sample compression framework, a learned predictor $A(S)$ can be fully defined by a reconstruction function $\mathcal{R}$, a *compression set* $S_{\mathbf{j}}$ and a *message* $\boldsymbol{\sigma}$, such that $\mathcal{R}(S_{\mathbf{j}}, \boldsymbol{\sigma}) = A(S)$.

Compression set

$S_{\mathbf{j}} \subseteq S$, with train indexes $\mathbf{j} \in \mathcal{P}(\mathbf{m})$, chosen from the power set of $\mathbf{m}=\{i\}_{i=1}^m$

Message

$\boldsymbol{\sigma} \in \Sigma$, where Σ is the set of all possible messages.

First sample compression bound for uncountable sets Σ

The theorem below bounds the generalization loss $\mathcal{L}_{\mathcal{D}}(h)$ from the empirical loss $\widehat{\mathcal{L}}_S(h)$ and two complexity terms: the compression set size $|\mathbf{j}|$ and the KL divergence between a prior P_Σ and a posterior Q_Σ distributions over messages.

Theorem 1 For any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for any set $J \subseteq \mathcal{P}(\mathbf{m})$, for any distribution P_J over J, for any distribution P_Σ over Σ, for any reconstruction function $\mathcal{R}$, for any loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, for any convex function $\Delta : [0,1] \times [0,1] \to \mathbb{R}$ and for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^m$, we have that for all $\mathbf{j} \in J$ and Q_Σ over Σ:

$$\Delta\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \widehat{\mathcal{L}}_{S_{\bar{\mathbf{j}}}}(\mathcal{R}(S_{\mathbf{j}}, \boldsymbol{\sigma})), \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \mathcal{L}_{\mathcal{D}}(\mathcal{R}(S_{\mathbf{j}}, \boldsymbol{\sigma}))\right) \leq \frac{1}{m-|\mathbf{j}|}\left[\mathrm{KL}(Q_\Sigma || P_\Sigma) + \ln\left(\frac{\mathcal{J}_\Delta(m-|\mathbf{j}|)}{P_J(\mathbf{j})\cdot\delta}\right)\right],$$

with

$$\mathcal{J}_\Delta(m-|\mathbf{j}|) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim P_\Sigma}\ \mathop{\mathbb{E}}_{T_{\mathbf{j}}\sim\mathcal{D}^{|\mathbf{j}|}}\ \mathop{\mathbb{E}}_{T_{\bar{\mathbf{j}}}\sim\mathcal{D}^{m-|\mathbf{j}|}} e^{(m-|\mathbf{j}|)\cdot\Delta\left(\widehat{\mathcal{L}}_{T_{\bar{\mathbf{j}}}}(\mathcal{R}(T_{\mathbf{j}},\boldsymbol{\sigma})), \mathcal{L}_{\mathcal{D}}(\mathcal{R}(T_{\mathbf{j}},\boldsymbol{\sigma}))\right)}.$$

Sample compression hypernetworks

2. We propose learning the reconstruction function.

Our **reconstruction hypernetwork** $\gamma = \mathcal{R}_\theta(S_{\mathbf{j}}, \boldsymbol{\sigma})$ takes two inputs:

- A compression set $S_{\mathbf{j}}$ containing a fixed number c examples;
- A message $\boldsymbol{\sigma}$ taking the form of a vector of fixed size b, either real-valued ($\boldsymbol{\sigma} \in [-1,1]^b$), or discrete ($\boldsymbol{\sigma} \in \{-1,1\}^b$).

The output $\gamma \in \mathbb{R}^{|\gamma|}$ is the parameters of a *downstream network*:

$$h_\gamma : \mathcal{X} \to \mathcal{Y}.$$

Objective function: Minimize the empirical loss of the downstream predictor h_γ on the complement set $S \setminus S_{\mathbf{j}}$:

$$\min_{\theta}\left\{\frac{1}{m-|\mathbf{j}|}\sum_{(\mathbf{x},y)\in S\setminus S_{\mathbf{j}}} \ell(h_\gamma(\mathbf{x}), y)\ \middle|\ \gamma = \mathcal{R}_\theta(S_{\mathbf{j}}, \boldsymbol{\sigma})\right\}.$$

From sample compression to meta-learning

The setting

- A meta-distribution $\mathbf{D}$, such that $\mathcal{D}_i \sim \mathbf{D}$;
- A meta-dataset $\mathbf{S} = \{S_i\}_{i=1}^n$, where $S_i \sim \mathcal{D}_i^{m_i}$;
- Each S_i is split into *support set* $\hat{S}_i \subset S_i$ and *query set* $\hat{T}_i = S_i \setminus \hat{S}_i$.

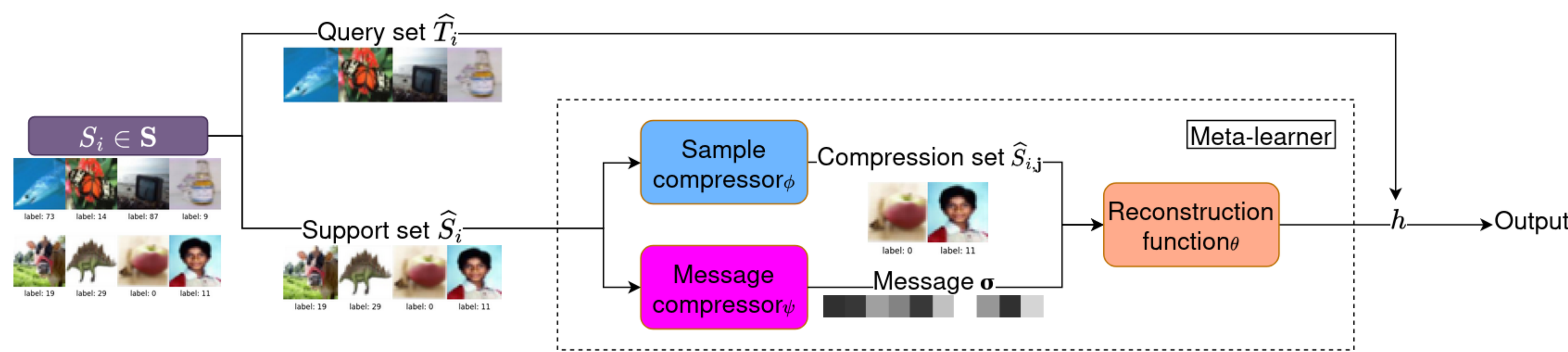
The proposed meta-learning algorithm

3. Let $\mathcal{C}_\phi$ be a *sample compressor* and $\mathcal{M}_\psi$ a *message compressor*. Given a query set $\hat{S}$, these respectively yield the compression set and the message used by the reconstruction hypernetwork to obtain the parameters γ of a downstream predictor:

$$\gamma = \mathcal{R}_\theta\left(\mathcal{C}_\phi(\hat{S}), \mathcal{M}_\psi(\hat{S})\right).$$

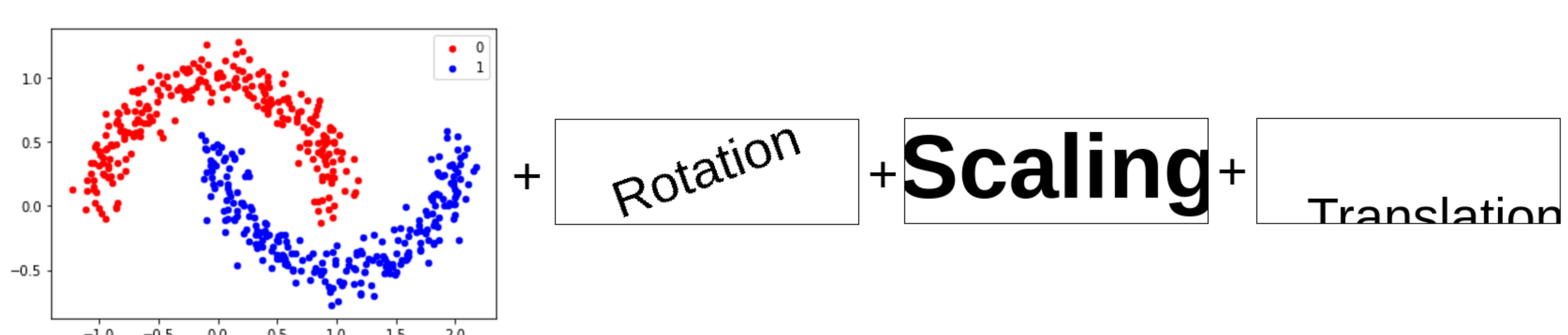
We propose to optimize the following meta-learning objective:

$$\min_{\theta,\phi,\psi}\left\{\frac{1}{n}\sum_{i=1}^n \frac{1}{m_i - |\hat{S}_i|}\sum_{(\mathbf{x},y)\in\hat{T}_i} \ell\left(h_{\gamma_i}(\mathbf{x}), y\right)\ \middle|\ \gamma_i = \mathcal{R}_\theta\left(\mathcal{C}_\phi(\hat{S}_i), \mathcal{M}_\psi(\hat{S}_i)\right)\right\},$$

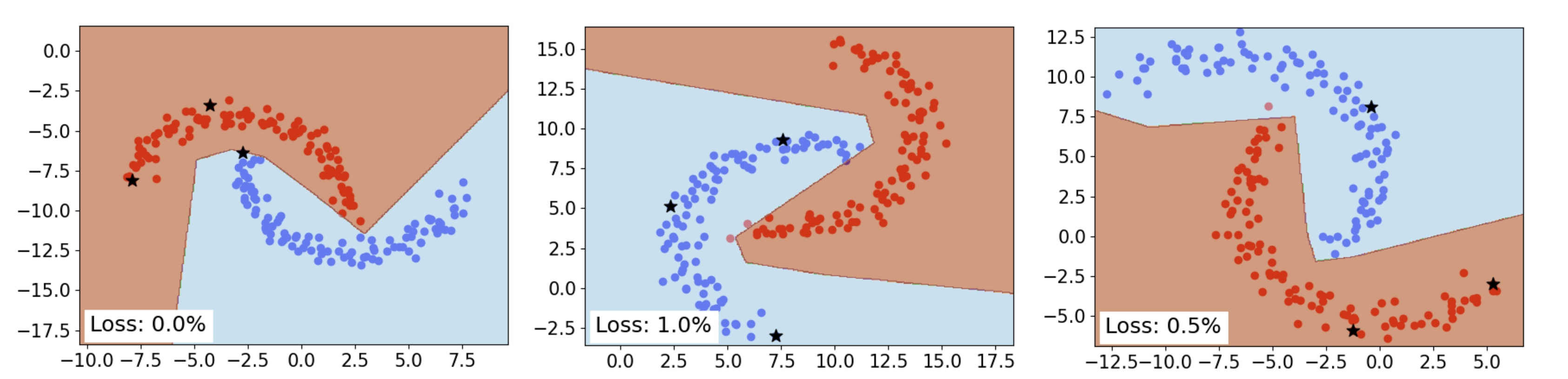


Numerical experiments

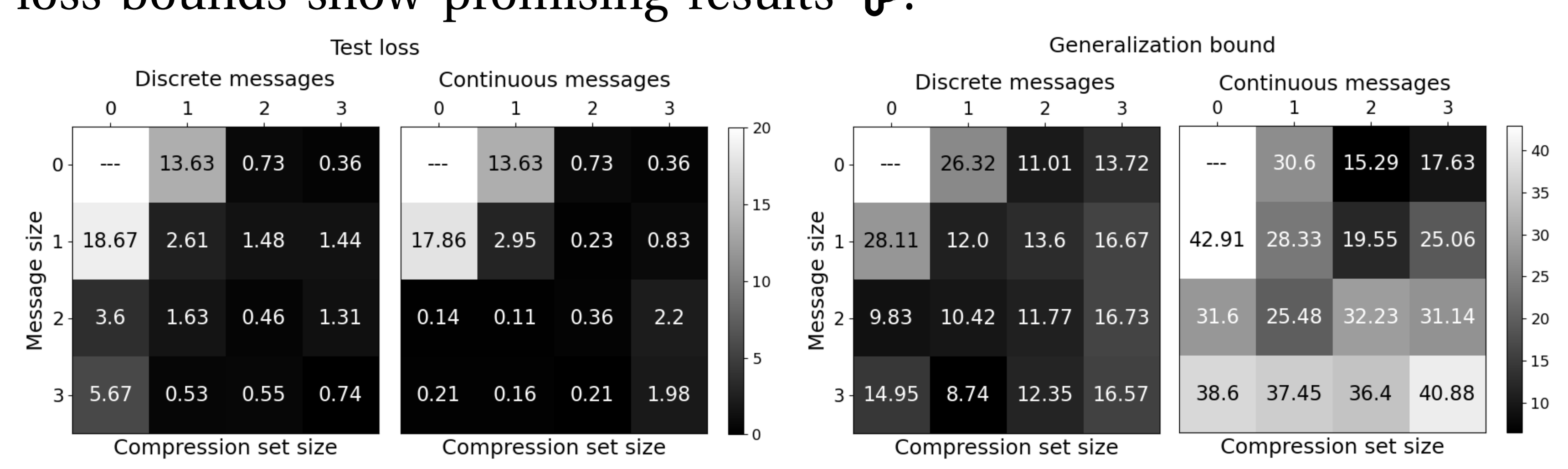
We generated 300 tasks of 200 examples using the following pipeline:



The generated predictor is a single-hidden-layer (of 5 neurons) ReLU MLP. Decision boundaries on three test tasks below (stars ⋆ indicate the compression set examples).



The obtained zero-one loss on 100 test tasks and sample-compressed loss bounds show promising results.



✓ Message size ↑ ⇒ test loss ↓ , generalization bound ↘ ↗

✓ Compression set size ↑ ⇒ test loss ↓ , generalization bound ↘ ↗

Contact: benjamin.leblanc.2@ulaval.ca

ArXiv: