Generalization Bounds via Meta-Learned Model Representations: PAC-Bayes and Sample Compression Hypernetworks

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

TL;DR

We design neural network bottleneck architectures that encode the complexity-accuracy trade-off stemming from two statistical learning theories.

- **Sample-Compress theory** → The bottleneck learn the reconstruction function;
- **PAC-Bayesian theory** → The bottleneck encodes the model into latent variables.

Definitions

The general setting

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

The meta-learning 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 $S_i \sim \mathcal{D}_i^m$, containing n datasets;
- 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 reconstruction function

In sample compression, 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 $\mathbf{j}$ chosen from the power set $\mathcal{P}(\cdot)$ of $\mathbf{m}=\{i\}_{i=1}^m$.

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

A general learning pipeline

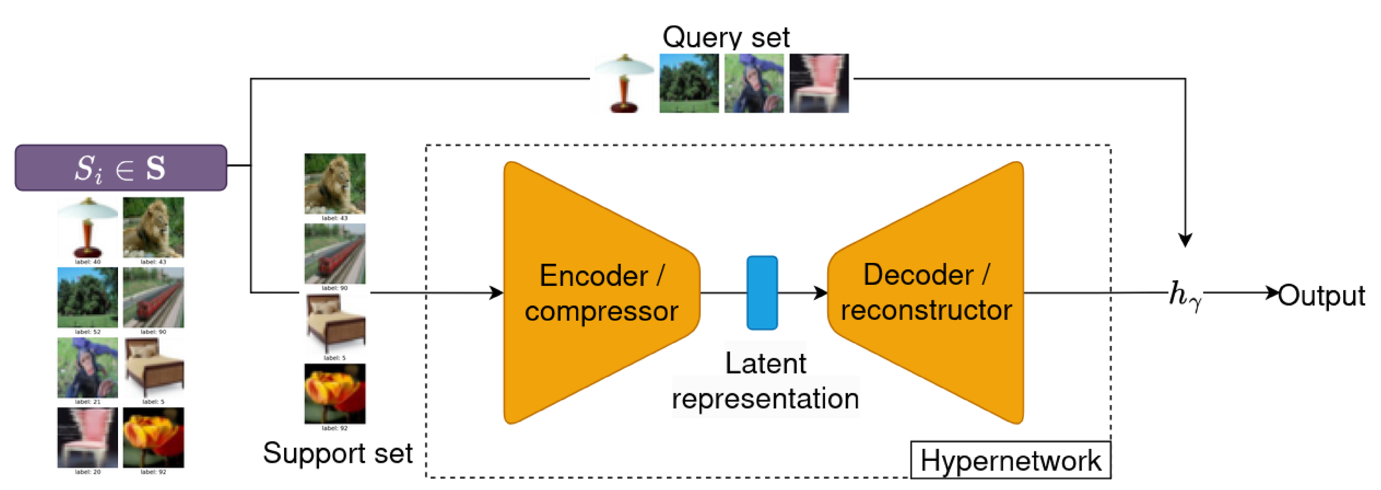
Encoder-Decoder Hypernetworks

We propose learning a hypernetwork $\mathcal{H}$ in the form of an encoder-decoder $\mathcal{H}(\cdot) = \mathcal{R}(\mathcal{C}(\cdot))$, whose output $\gamma \in \mathbb{R}^{|\gamma|}$ is the parameters of a *downstream network*: h_γ.

Objective function: Empirical loss on query sets $\{\hat{T}_i\}_{i=1}^n$ of the downstream predictor h_{γ_i} obtained with support sets $\{\hat{S}_i\}_{i=1}^n$:

$$\min_\theta \left\{ \frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{L}}_{\hat{T}_i}(h_{\gamma_i}) \,\middle|\, \gamma_i = \mathcal{H}_\theta(\hat{S}_i) \right\}.$$

Leading to the following hypernetwork:



The generalization bound

We bound the risk $\mathcal{L}_\mathcal{D}(h_\gamma)$ of the outputted hypothesis h_γ from the empirical loss and the latent representation complexity, seen as a ⟨message, compression set⟩ couple:

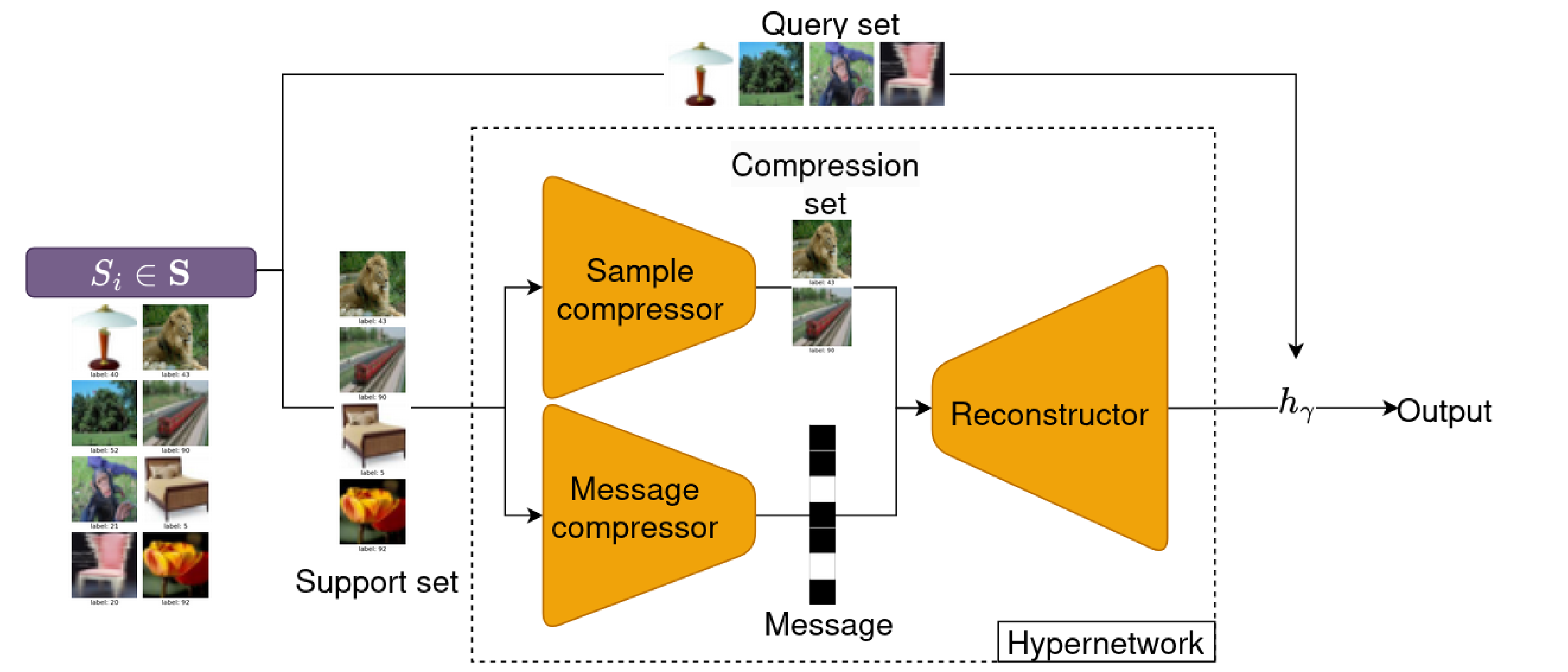
Theorem 1 For any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, distributions P_Σ over messages Σ and P_J over compression sets $\mathcal{P}(\mathbf{m})$, reconstruction function $\mathcal{R}$, $\delta \in (0,1]$, with probability at least $1-\delta$ over the draw of $S \sim \mathcal{D}^m$:

$\forall \mathbf{j} \in J, Q_\Sigma$ over Σ :

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \widehat{\mathcal{L}}_{S_{\bar{\mathbf{j}}}}(\mathcal{R}(S_\mathbf{j}, \boldsymbol{\sigma})), \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \mathcal{L}_\mathcal{D}(\mathcal{R}(S_\mathbf{j}, \boldsymbol{\sigma})) \right) \leq \frac{1}{m - \max_{\mathbf{j}\in J}|\mathbf{j}|}\left[\mathrm{KL}(Q_\Sigma \| P_\Sigma) + \ln\left(\frac{2\sqrt{m - |\mathbf{j}|}}{P_J(\mathbf{j}) \cdot \delta} \right) \right].$$

Sample Compress Hypernetworks

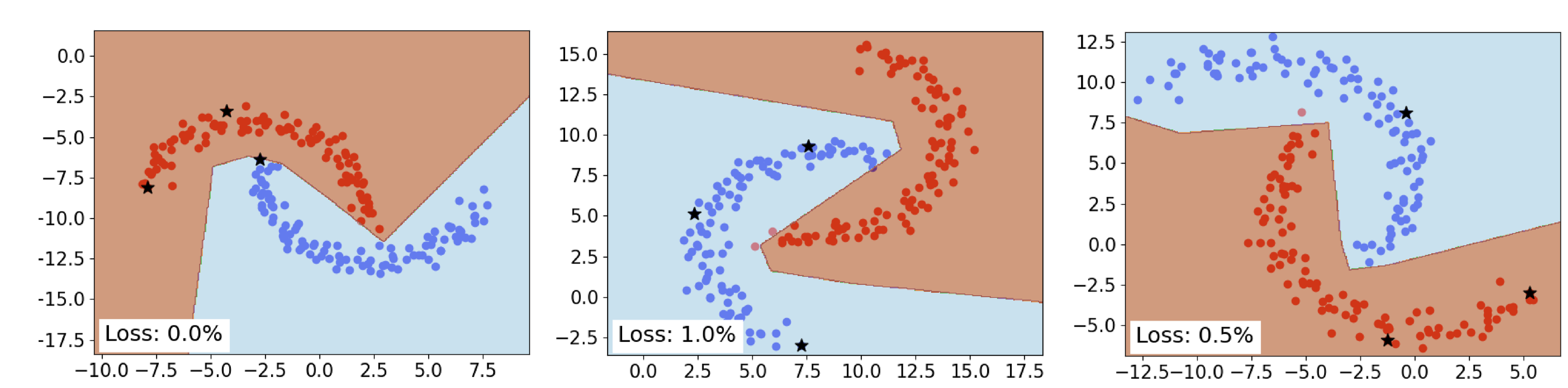
We first create an encoder architecture suited to the sample compression theory:



Given a compression set size b and using priors $P_J(\mathbf{j}) = \binom{m}{|\mathbf{j}|}^{-1}\ \forall \mathbf{j} \in J$ and $P_\Sigma(\boldsymbol{\sigma}) = 2^{-b}\ \forall \boldsymbol{\sigma} \in \{-1,1\}^b$, given $S \sim \mathcal{D}^m$, with probability at least $1-\delta$, we have

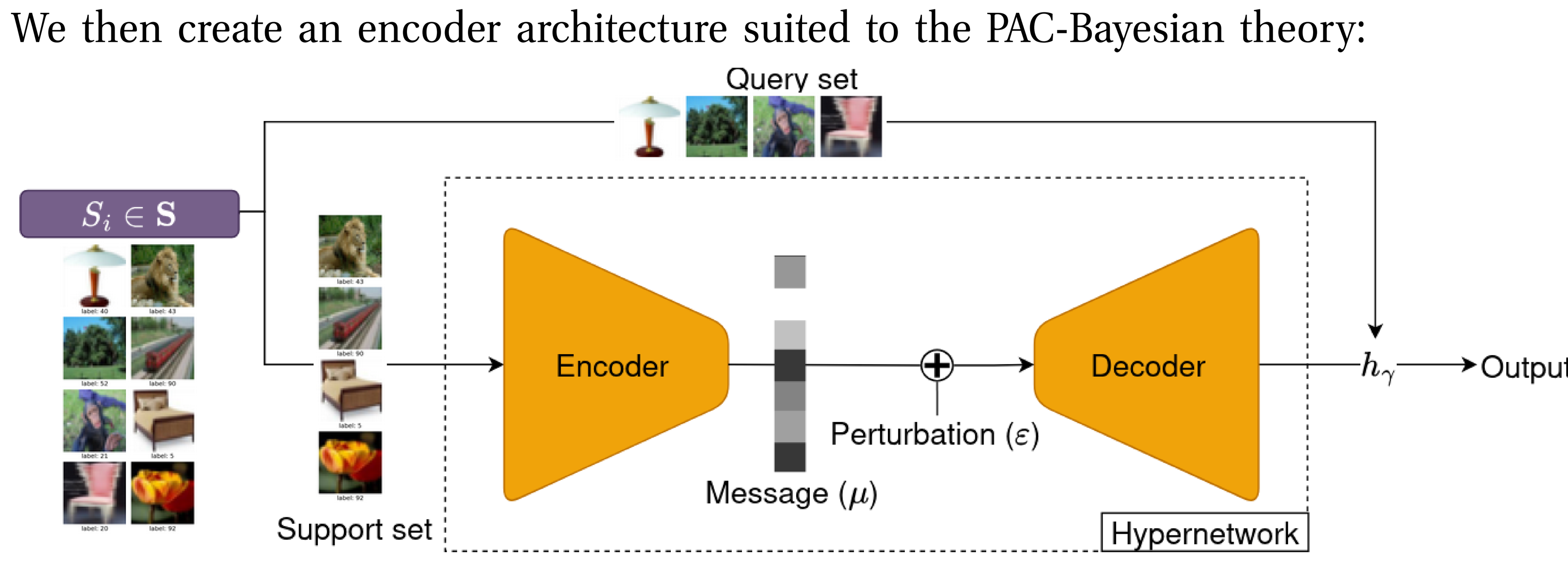
$$\mathcal{L}_\mathcal{D}(h_\gamma) \leq \operatorname*{argsup}_{\tau\in[0,1]} \left\{ \mathrm{kl}\left(\widehat{\mathcal{L}}_{S_{\bar{\mathbf{j}}}}(h_\gamma), \tau\right) \leq \frac{1}{m - |\mathbf{j}|} \ln\left(\binom{m}{|\mathbf{j}|} \frac{2^{b+1}\sqrt{m-|\mathbf{j}|}}{\delta} \right) \right\}.$$

A size-three compression set (and no message) is sufficient to learn small tasks!



PAC-Bayesian Hypernetworks

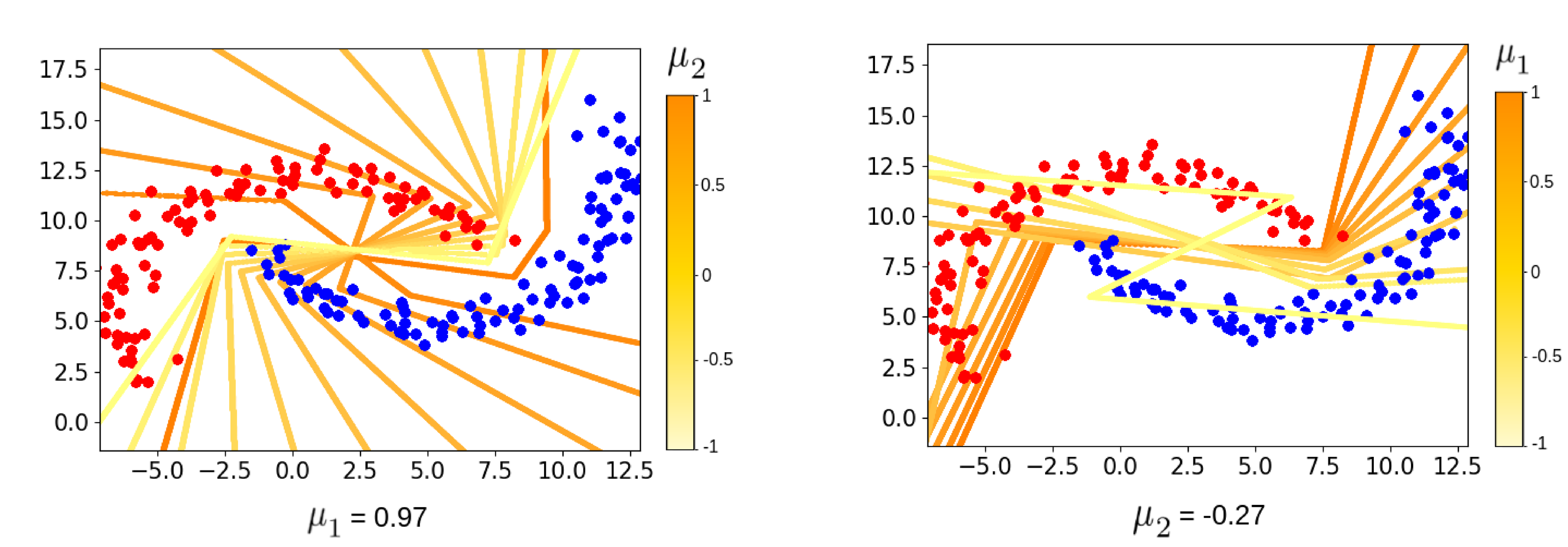
We then create an encoder architecture suited to the PAC-Bayesian theory:



Using a prior $P_\Sigma = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a posterior $Q_\Sigma = \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, given $S \sim \mathcal{D}^m$, with probability at least $1-\delta$, we have

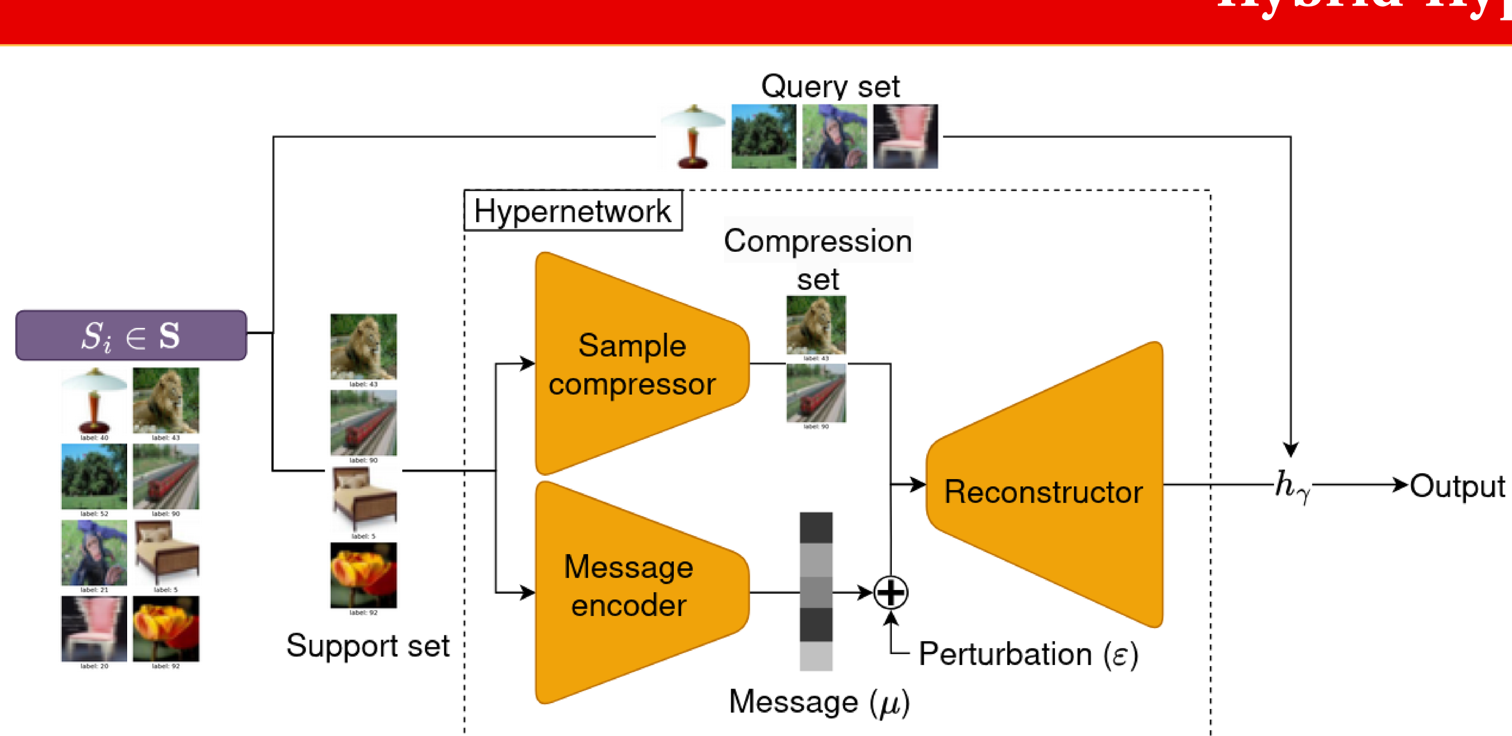
$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \mathcal{L}_\mathcal{D}(h_\gamma) \leq \operatorname*{argsup}_{\tau\in[0,1]} \left\{ \mathrm{kl}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim Q_\Sigma} \widehat{\mathcal{L}}_S(h_\gamma), \tau \right) \leq \frac{1}{m}\left(\frac{1}{2}\|\boldsymbol{\mu}\|^2 + \ln\frac{2\sqrt{m}}{\delta} \right) \right\}.$$

With a size-two message, we can isolate the role of each component!



Hybrid Hypernetworks

Combining both Sample Compress and PAC-Bayes leads to the following encoder architecture:



We conceive a meta-learning environment in which binary tasks are created by sampling two classes from the MNIST (CIFAR100) task. Test tasks contain 2000 (200) examples.

Algorithm	MNIST Bound (↓)	MNIST Test error (↓)	CIFAR100 Bound (↓)	CIFAR100 Test error (↓)
(Pentina & Lampert, 2014)	0.767 ± 0.001	0.369 ± 0.223	0.801 ± 0.001	0.490 ± 0.070
(Amit & Meir, 2018)	1372 ± 23.36	0.351 ± 0.212	950.9 ± 343.1	0.284 ± 0.120
(Guan & Lu, 2022) - kl	0.754 ± 0.003	0.366 ± 0.221	0.802 ± 0.001	0.489 ± 0.073
(Guan & Lu, 2022) - Cat.	1.132 ± 0.021	0.351 ± 0.212	1.577 ± 0.567	0.282 ± 0.122
(Rezazadeh, 2022)	11.43 ± 0.005	0.366 ± 0.221	10.91 ± 0.368	0.334 ± 0.139
(Zakerinia et al., 2024)	0.684 ± 0.021	0.351 ± 0.212	0.953 ± 0.315	**0.281** ± 0.125
Sample compress hypernet.	**0.280** ± 0.148	0.155 ± 0.109	**0.745** ± 0.101	0.305 ± 0.142
PAC-Bayesian hypernet.	0.597 ± 0.107	**0.150** ± 0.114	0.974 ± 0.022	0.295 ± 0.103
Hybrid hypernet.	0.597 ± 0.107	**0.150** ± 0.114	0.974 ± 0.022	0.295 ± 0.103