Capsule Routing via Variational Bayes

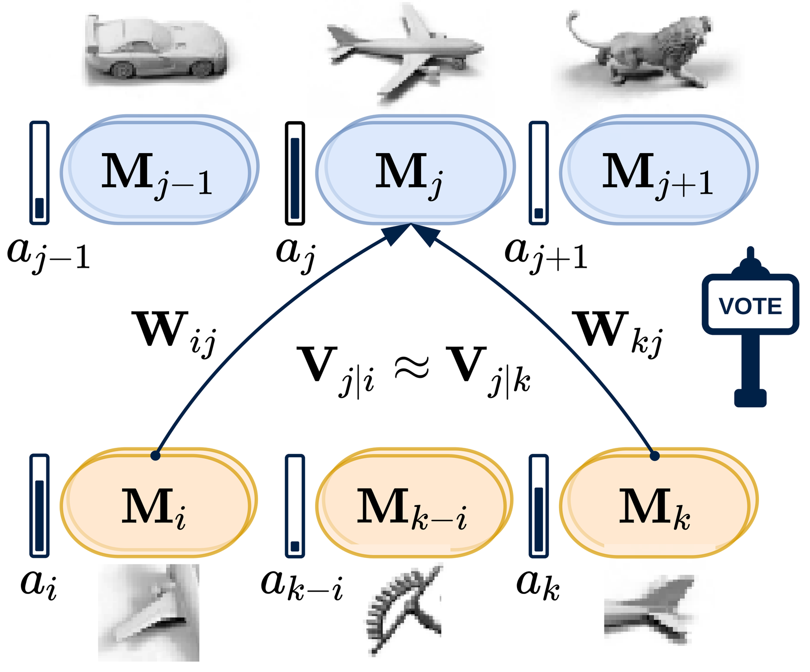
Fabio De Sousa Ribeiro, Georgios Leontidis, Stefanos Kollias
Machine Learning Group, University of Lincoln, UK
{fdesousaribeiro,gleontidis,skollias}@lincoln.ac.uk

UNIVERSITY OF LINCOLN

1. Overview & Contribution

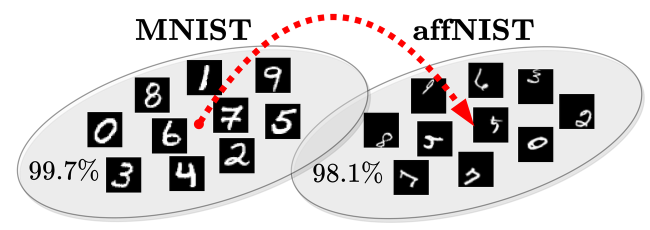
Placing priors on capsule parameters stabilises training and improves performance.

- '**Parts**' vote for '**wholes**' using **viewpoint-invariant** transformations
  $$\mathbf{V}_{j|i} = \mathbf{M}_i \cdot \mathbf{W}_{ij}\ ,\ \mathbf{W}_{ij} \in \mathbb{R}^{4\times 4}.$$
- Previously **unstable** training due to **variance collapse** during routing.
- We propose a robust **Variational Bayes Routing** algorithm.

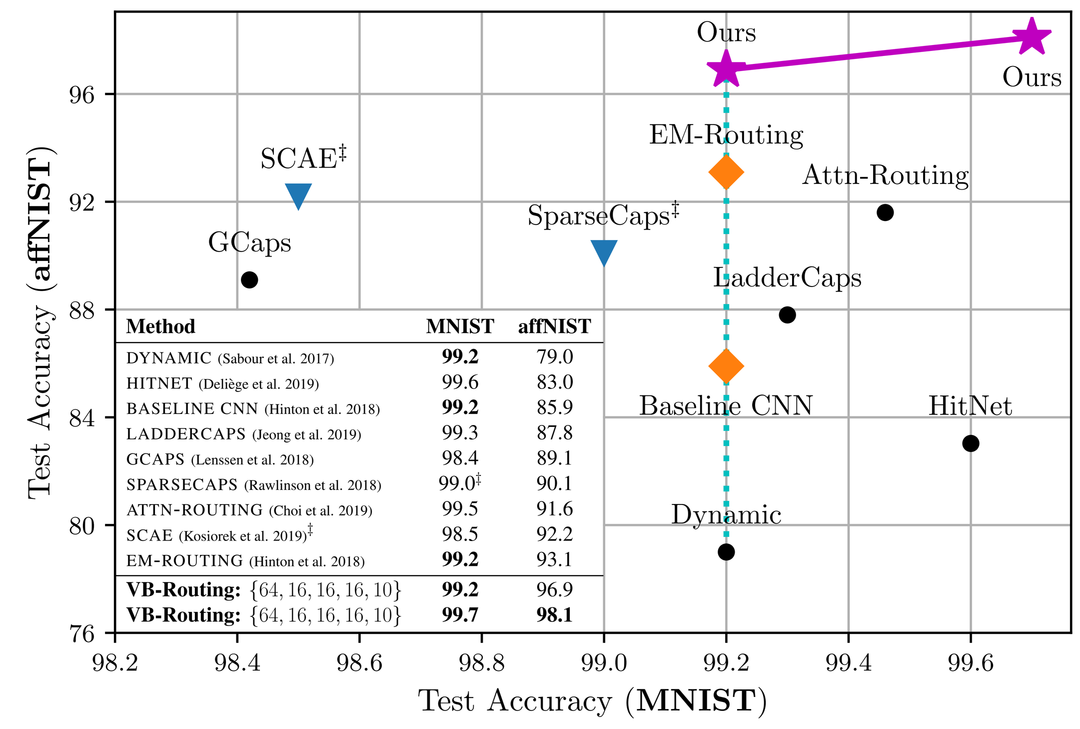


4. Affine Transformation Robustness

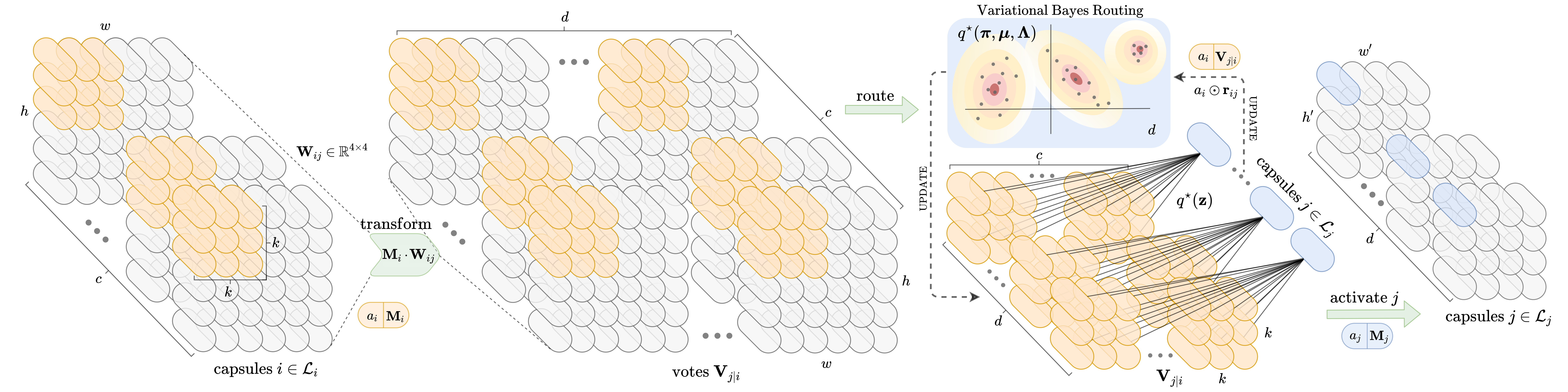
Significant improvement in MNIST to affNIST generalisation.



- **Train** on **MNIST**.
- **Validate** on an **MNIST** 10% training set split.
- **Test** on **affNIST** test set.



Method	MNIST	affNIST
DYNAMIC (Sabour et al. 2017)	**99.2**	79.0
HITNET (Deliège et al. 2019)	99.6	83.0
BASELINE CNN (Hinton et al. 2018)	**99.2**	85.9
LADDERCAPS (Jeong et al. 2019)	99.3	87.8
GCAPS (Lenssen et al. 2018)	98.4	89.1
SPARSECAPS (Rawlinson et al. 2018)	99.0‡	90.1
ATTN-ROUTING (Choi et al. 2019)	99.5	91.6
SCAE (Kosiorek et al. 2019)‡	98.5	92.2
EM-ROUTING (Hinton et al. 2018)	**99.2**	93.1
VB-Routing: $\{64, 16, 16, 16, 10\}$	**99.2**	96.9
VB-Routing: $\{64, 16, 16, 16, 10\}$	**99.7**	**98.1**

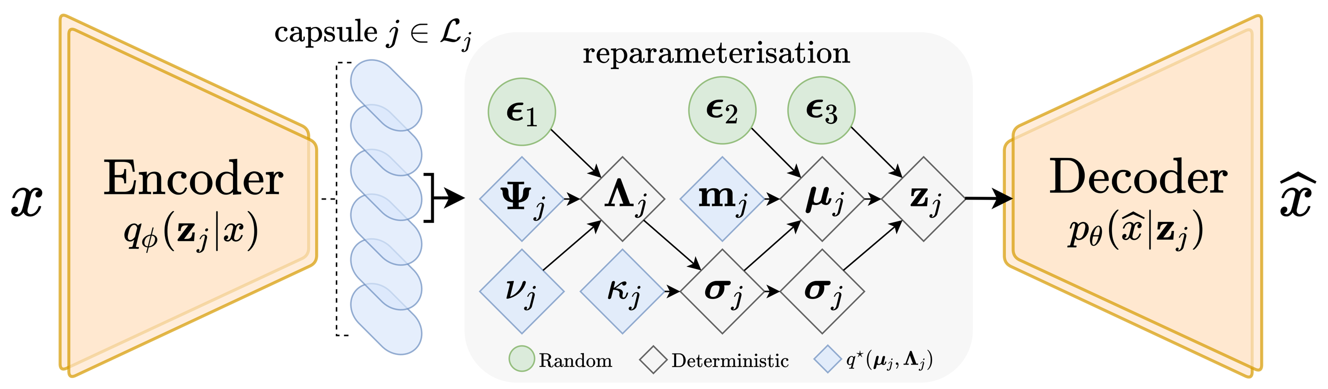


2. Variational Bayes Capsule Routing

- The **posterior** on the latent variables and capsule parameters is approximated with a **factorised variational** distribution
  $$p(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{V}) \approx q(\mathbf{z})q(\boldsymbol{\pi}) \prod_{j\in\mathcal{L}_j} q(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j). \tag{1}$$
- We place **conjugate** priors over $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ as
  $$p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) \prod_{j\in\mathcal{L}_j} \mathcal{N}\big(\boldsymbol{\mu}_j|\mathbf{m}_0, (\kappa_0\boldsymbol{\Lambda}_j)^{-1}\big)\mathrm{Wi}(\boldsymbol{\Lambda}_j|\boldsymbol{\Psi}_0, \nu_0). \tag{2}$$
- This yields **closed-form** iterative updates for $q^\star(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ and $q^\star(\mathbf{z})$.
- We propose to measure **agreement** after the iterations as
  $$\mathbb{H}[q^\star(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)] \approx \mathbb{E}[\ln\det(\boldsymbol{\Lambda}_j)] = \sum_{i=0}^{D-1} \psi\left(\frac{\nu_j - i}{2}\right) + D\ln 2 + \ln\det(\boldsymbol{\Psi}_j), \tag{3}$$
  and **activate** capsules with low **entropy** posteriors $q^\star(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$.

2.1 Capsule-VAE

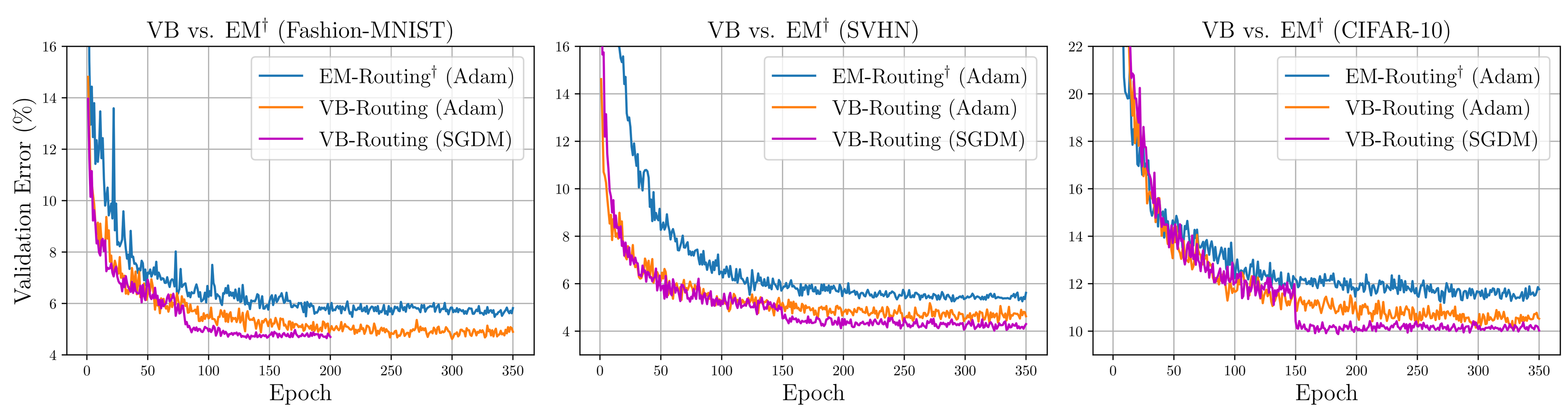
- Sample from capsule j's parameter posterior $q^\star(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$ as
  $$\boldsymbol{\Lambda}_j|\boldsymbol{\Psi}_j, \nu_j \sim \mathrm{Wi}(\boldsymbol{\Psi}_j, \nu_j)\ ,\quad \boldsymbol{\mu}_j|\boldsymbol{\Lambda}_j, \mathbf{m}_j, \kappa_j \sim \mathcal{N}(\mathbf{m}_j, (\kappa_j\boldsymbol{\Lambda}_j)^{-1}). \tag{4}$$



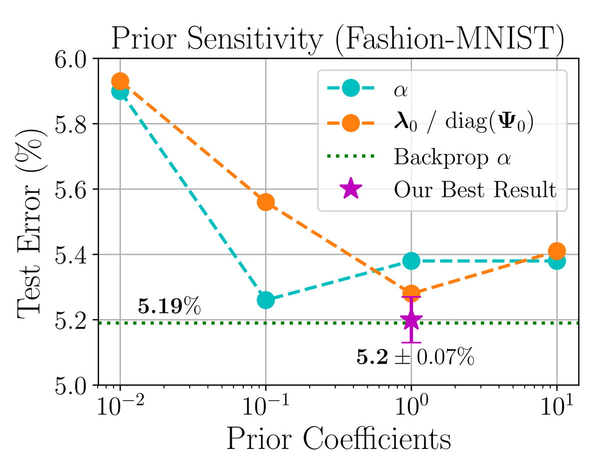
- Use the **reparameterisation trick** to backprop through random nodes.

5. Image Classification

State-of-the-art performance on smallNORB using ≃50% fewer capsules.



- **Sensitivity to Prior Hyperparameters**



- **Runtime of Routing Algorithms**

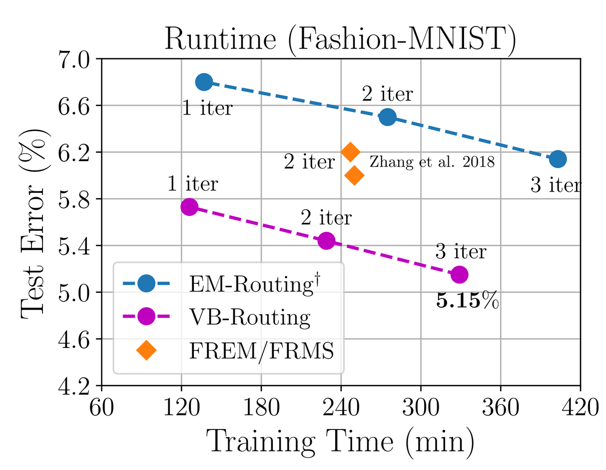
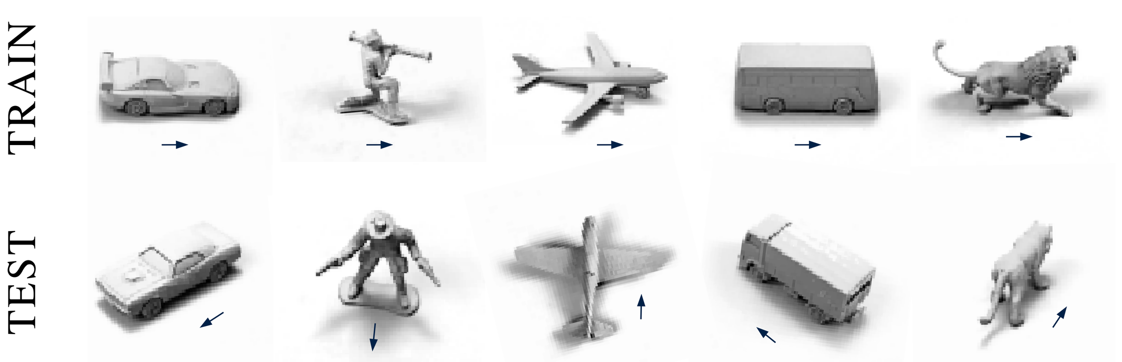


Table 2. $\{\cdot\}$ denotes ensemble size, (†) our EM implementation, and $[\cdot]$ 4× residual block backbone usage.

Method	smallNORB Error (%)	smallNORB Param	FashionMNIST Error (%)	FashionMNIST Param	SVHN Error (%)	SVHN Param	CIFAR10 Error (%)	CIFAR10 Param
DYNAMIC (Sabour et al. 2017)	2.7	8.2M	-	-	4.3	≃1.8M	10.6	8.2M {7}
FREM/FRMS (Zhang et al. 2018)	2.2	1.2M	6.0	1.2M	-	-	14.3	1.2M
EM-ROUTING (Hinton et al. 2018)	1.8 (1.4)	310K	-	-	-	-	11.9	≃460K
VB-Routing: $\{64, 16, 32, 32, d_4\}$	1.8	318K	5.6	323K	**3.87**	**323K**	**11.1** [7.8]	**323K**
vs. EM†: $\{64, 16, 32, 32, d_4\}$	-	-	-	-	5.2	323K	12.3	323K
VB-Routing: $\{64, 16, 16, 16, d_4\}$	**1.29**	**169K**	**5.15**	**172K**	4.2	172K	12.4	172K
vs. EM†: $\{64, 16, 16, 16, d_4\}$	2.0	169K	6.1	172K	-	-	-	-

3. Generalisation to Novel Viewpoints



- **VB-Routing & EM-Routing†**

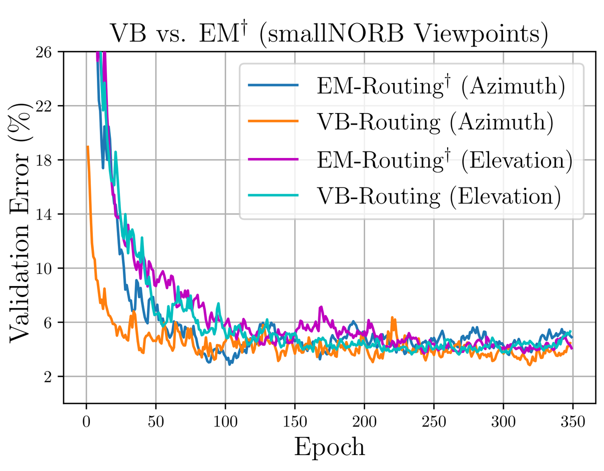


Table 1. (†) denotes our implementation of EM (Hinton et al. 2018).

Viewpoints (Test)	Azimuth (%) CNN	Azimuth (%) **VB**	Azimuth (%) EM†	Azimuth (%) EM	Elevation (%) CNN	Elevation (%) **VB**	Elevation (%) EM†	Elevation (%) EM
NOVEL	20	**11.33**	12.67	13.5	17.8	**11.59**	12.04	12.3
FAMILIAR	3.7	3.71	3.72	3.7	4.3	4.32	4.29	4.3

Full Paper, Poster and Code:

github.com/fabio-deep/Variational-Capsule-Routing

6. Related Work

[1] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International Conference on Learning Representations*, 2018.
[2] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.
[3] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.
[4] Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E Hinton. Stacked capsule autoencoders. In *Advances in Neural Information Processing Systems*, pages 15486–15496, 2019.