Overcoming Catastrophic Forgetting by Incremental Moment Matching

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun
Seoul National University

Jung-Woo Ha
Clova AI Research, NAVER Corp.

Byoung-Tak Zhang
Seoul National University & Surromind Robotics

Introduction

Continual Learning Problem: continuously training neural networks for a new task without information trained on the previous tasks. The goal is to make the network perform well for both tasks.

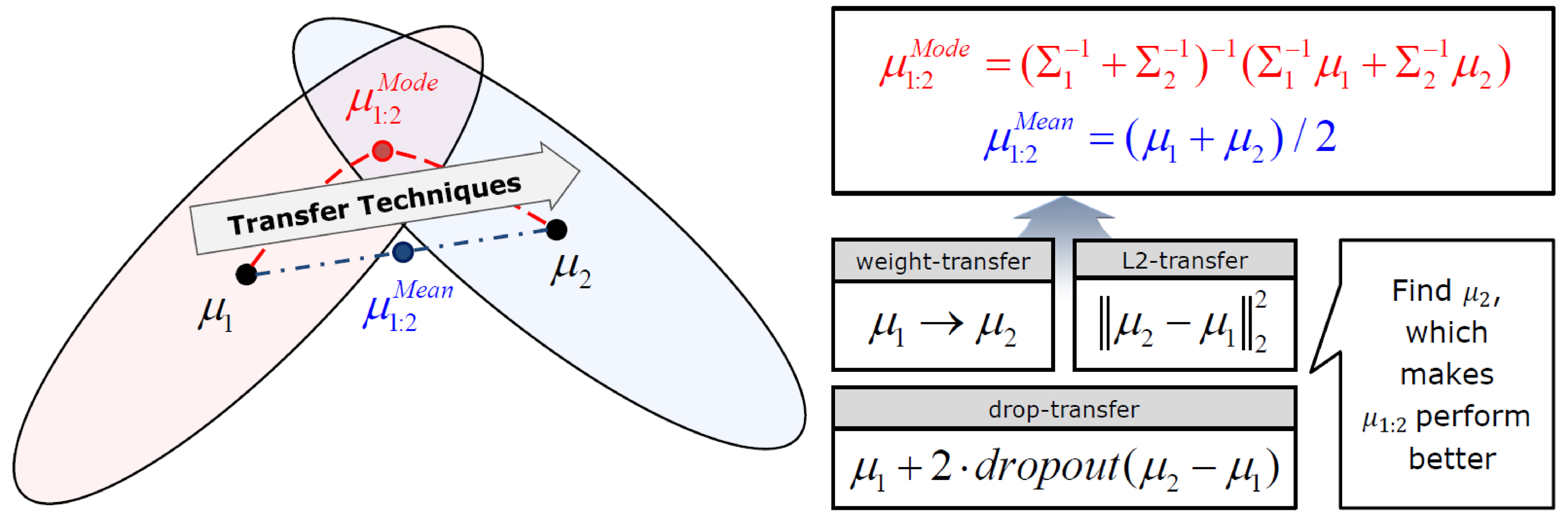
Catastrophic Forgetting: neural networks lose the performance for the previous tasks after training the new task.

Incremental moment matching (IMM): incrementally matching the moment of the posterior distribution of the neural network which is trained on the previous and the new tasks.

Contribution

1. Propose two types of **incremental moment matching (IMM)** methods for overcoming catastrophic forgetting
   - Mean-Incremental Moment Matching (**mean-IMM**)
   - Mode-Incremental Moment Matching (**mode-IMM**)
2. Interpret the IMMs as the **Bayesian** perspectives
3. Propose **drop-transfer** as both a **knowledge transfer method** for IMM and a **continual learning method**
4. **Apply various transfer techniques** in the IMM procedure to make our assumption of Gaussian distribution reasonable

Incremental Moment Matching



Geometric illustration of incremental moment matching (IMM). Mean-IMM simply averages the parameters of two neural networks, whereas mode-IMM tries to find a maximum of the mixture of Gaussian posteriors.

To make IMM be reasonable, the search space of the loss function between two posterior means μ_1 and μ_2 should be reasonably smooth and convex-like. To find a μ_2 which satisfies this condition of a smooth and convex-like path from μ_1, we propose applying various transfer techniques for the IMM procedure.

Merging by Approximating Mixture of Gaussian Posteriors

Mean-IMM: minimize local KL-divergence

$$\mu^*_{1:K}, \Sigma^*_{1:K} = \underset{\mu_{1:K}, \Sigma_{1:K}}{\text{argmin}} \sum_k^K \alpha_k KL(q_k \| q_{1:K})$$

$$\mu^*_{1:K} = \sum_k^K \alpha_k \mu_k$$

$$\Sigma^*_{1:K} = \sum_k^K \alpha_k (\Sigma_k + (\mu_k - \mu^*_{1:K})(\mu_k - \mu^*_{1:K})^T)$$

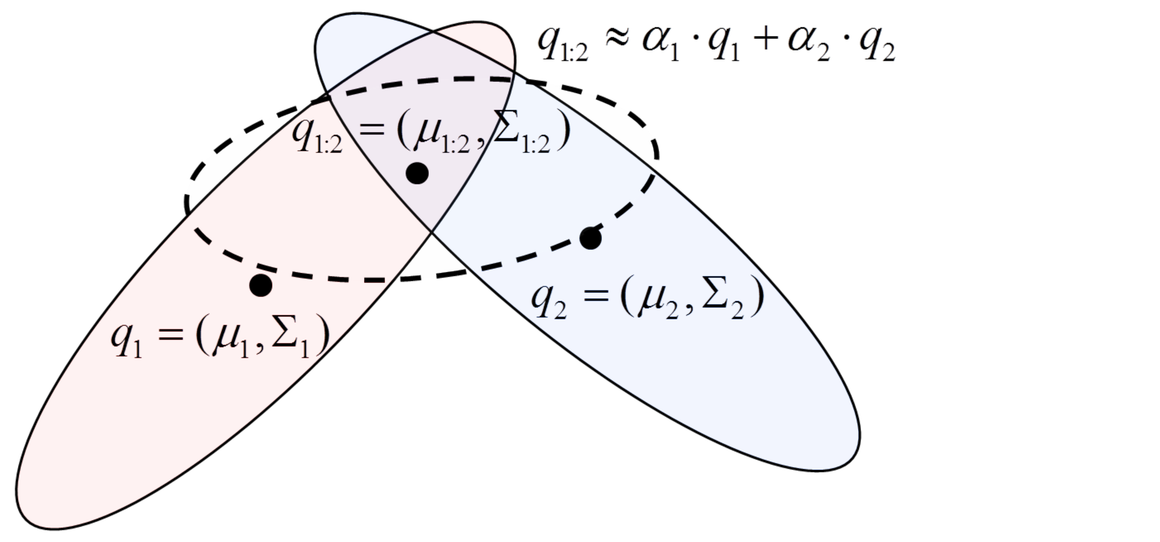
Mode-IMM: find a mode of mixture of local posteriors

$$\mu^*_{1:K}, \Sigma^*_{1:K} = \underset{\mu}{\text{argmax}} \sum_k^K \alpha_k q_k$$

$$\mu^*_{1:K} = \Sigma^*_{1:K} \cdot \sum_k^K \alpha_k \Sigma_k^{-1} \mu_k$$

$$\Sigma^*_{1:K} = (\sum_k^K \alpha_k \Sigma_k^{-1})^{-1}$$

Assume local posterior and approximated global posterior is Gaussian

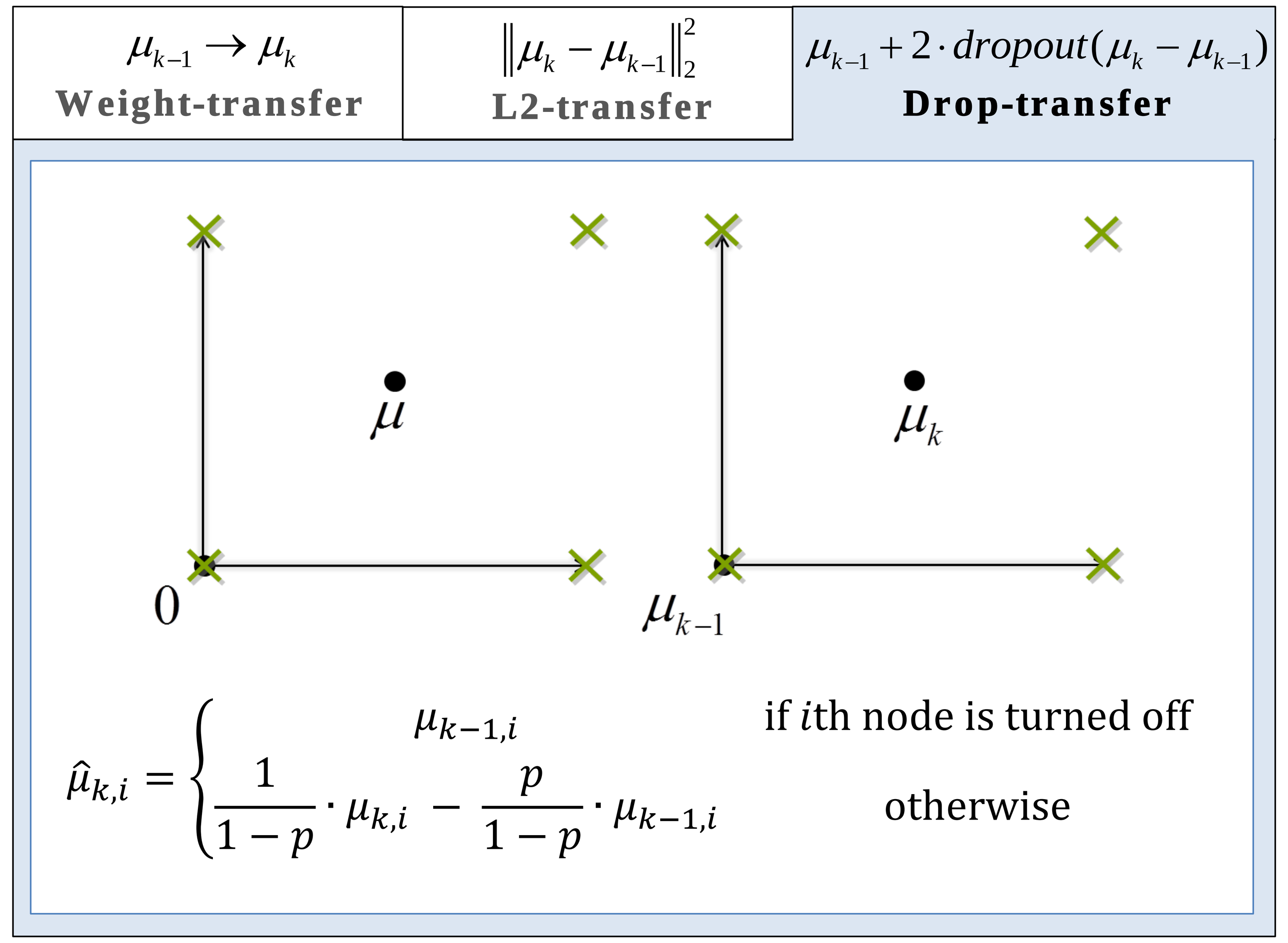


Use inverse Fisher matrix as covariance matrix

$$\Sigma_k^{-1} \approx F_k = E_{x\sim\pi_k, \tilde{y}\sim p(y|x,\mu_k)}\left[\frac{\partial}{\partial \mu_k} \ln p(\tilde{y}|x,\mu_k) \cdot \ln p(\tilde{y}|x,\mu_k)^T\right]$$

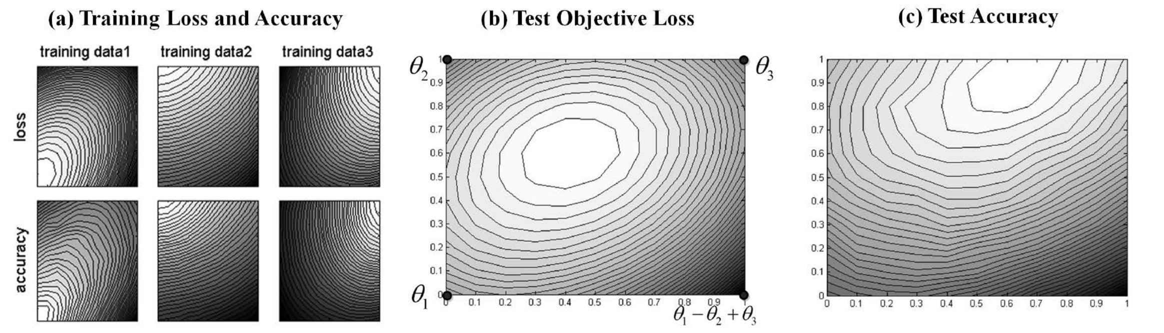
Making Search Spaces Smooth by Transfer Techniques

Transfer Techniques for IMM

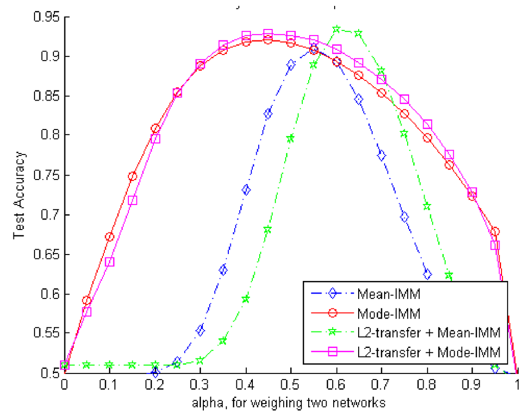


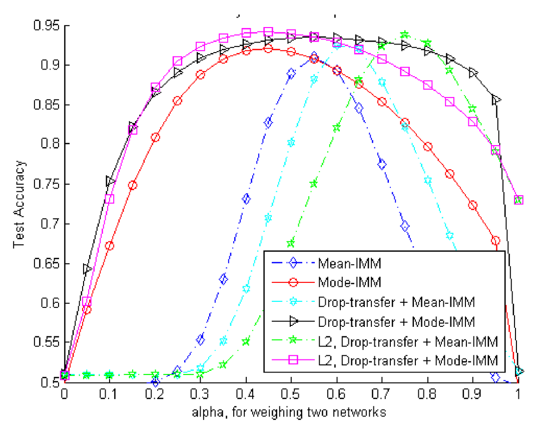
Smooth Search Space

Weight-transfer makes the search space convex-like (CIFAR-10)



Various transfer techniques for IMM makes the search space in the line/curve smooth (Disjoint MNIST)





Experimental Results

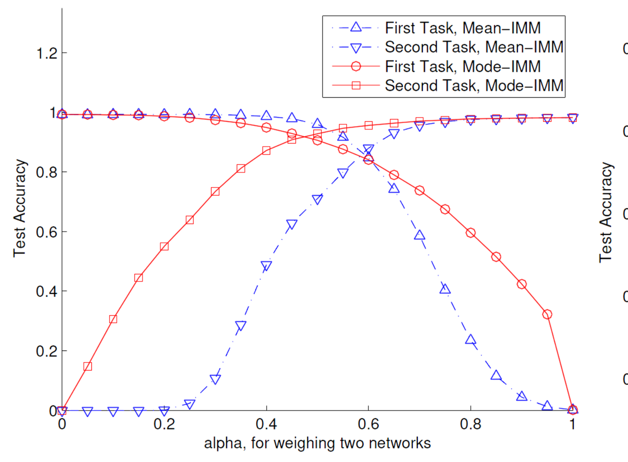
Comparison on Disjoint MNIST and Shuffled MNIST Datasets

Disjoint MNIST Experiment	Explanation of Hyperparam	Untuned Hyperparam	Untuned Accuracy	Tuned Hyperparam	Tuned Accuracy
SGD [3]	epoch per dataset	10	47.72 ($\pm$ 0.11)	0.05	71.32 ($\pm$ 1.54)
L2-transfer [25]	λ in (10)	-	-	0.05	85.81 ($\pm$ 0.52)
Drop-transfer	p in (11)	0.5	51.72 ($\pm$ 0.79)	0.5	51.72 ($\pm$ 0.79)
EWC [8]	λ in (20)	1.0	47.84 ($\pm$ 0.04)	600M	52.72 ($\pm$ 1.36)
Mean-IMM	α_2 in (4)	0.50	90.45 ($\pm$ 2.24)	0.55	91.92 ($\pm$ 0.98)
Mode-IMM	α_2 in (7)	0.50	91.49 ($\pm$ 0.98)	0.45	92.02 ($\pm$ 0.73)
L2-transfer + Mean-IMM	λ / α_2	0.001 / 0.50	78.34 ($\pm$ 1.82)	0.001 / 0.60	92.62 ($\pm$ 0.95)
L2-transfer + Mode-IMM	λ / α_2	0.001 / 0.50	92.52 ($\pm$ 0.41)	0.001 / 0.45	92.73 ($\pm$ 0.35)
Drop-transfer + Mean-IMM	p / α_2	0.5 / 0.50	80.75 ($\pm$ 1.28)	0.5 / 0.60	92.64 ($\pm$ 0.60)
Drop-transfer + Mode-IMM	p / α_2	0.5 / 0.50	93.35 ($\pm$ 0.49)	0.5 / 0.50	93.35 ($\pm$ 0.49)
L2, Drop-transfer + Mean-IMM	$\lambda / p / \alpha_2$	0.001 / 0.5 / 0.50	66.10 ($\pm$ 3.19)	0.001 / 0.5 / 0.75	**93.97** ($\pm$ 0.23)
L2, Drop-transfer + Mode-IMM	$\lambda / p / \alpha_2$	0.001 / 0.5 / 0.50	**93.97** ($\pm$ 0.32)	0.001 / 0.5 / 0.45	**94.12** ($\pm$ 0.27)

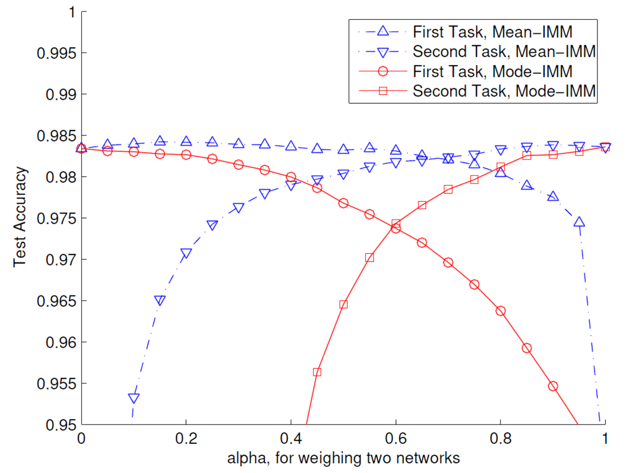
Shuffled MNIST Experiment		Hyperparam	Accuracy	Hyperparam	Accuracy
SGD [3]	epoch per dataset	60	89.15 ($\pm$ 2.34)	-	~95.5 [8]
L2-transfer [25]	λ in (10)	-	-	1e-3	96.37 ($\pm$ 0.62)
Drop-transfer	p in (11)	0.5	94.75 ($\pm$ 0.62)	0.2	96.86 ($\pm$ 0.21)
EWC [8]	λ in (20)	-	-	-	**~98.2** [8]
Mean-IMM	α_3 in (4)	0.33	93.23 ($\pm$ 1.37)	0.55	95.02 ($\pm$ 0.42)
Mode-IMM	α_3 in (7)	0.33	98.02 ($\pm$ 0.05)	0.60	98.08 ($\pm$ 0.08)
L2-transfer + Mean-IMM	λ / α_3	1e-4 / 0.33	90.38 ($\pm$ 1.74)	1e-4 / 0.65	95.93 ($\pm$ 0.31)
L2-transfer + Mode-IMM	λ / α_3	1e-4 / 0.33	**98.16 ($\pm$ 0.08)**	1e-4 / 0.60	**98.30 ($\pm$ 0.08)**
Drop-transfer + Mean-IMM	p / α_3	0.5 / 0.33	90.79 ($\pm$ 1.30)	0.5 / 0.65	96.49 ($\pm$ 0.44)
Drop-transfer + Mode-IMM	p / α_3	0.5 / 0.33	97.80 ($\pm$ 0.07)	0.5 / 0.55	97.95 ($\pm$ 0.08)
L2, Drop-transfer + Mean-IMM	$\lambda / p / \alpha_3$	1e-4 / 0.5 / 0.33	89.51 ($\pm$ 2.85)	1e-4 / 0.5 / 0.90	97.36 ($\pm$ 0.19)
L2, Drop-transfer + Mode-IMM	$\lambda / p / \alpha_3$	1e-4 / 0.5 / 0.33	97.83 ($\pm$ 0.10)	1e-4 / 0.5 / 0.50	97.92 ($\pm$ 0.05)

Test Accuracies with Different Balancing Parameters

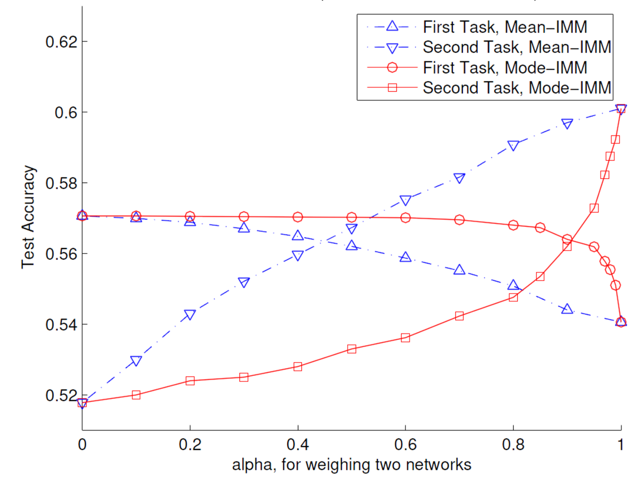
Disjoint MNIST
{0,1,2,3,4}->{5,6,7,8,9}



Shuffled MNIST
Pixels are shuffled with a fixed, random permutation



ImageNet2CUB
General objects (1000 classes) -> Bird (200 classes)



Comparison on Lifelog Dataset

Egocentric Video recorded from Google Glass, 660,000 instances, 3 participants, 46 days

	Location	Sub-location	Activity	A	B	C
Dual memory architecture [12]	**78.11**	72.36	52.92	67.02	58.80	77.57
Mean-IMM	77.60	73.78	52.74	67.03	57.73	**79.35**
Mode-IMM	77.14	**75.76**	**54.07**	**67.97**	**60.12**	78.89

ACKNOWLEDGMENT This work was supported by Naver Corp. and partly by the Korea government (IITP-2017-0-01772-VTT, IITP-R0126-16-1072-SW.StarLab, KEIT-10044009-HRI.MESSI, KEIT-10060086-RISF). JK is supported by 2017 Google PhD Fellowship.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.