Deep N-ary Error Correcting Output Codes

Hao Zhang[1], Joey Tianyi Zhou[1,*], Tianying Wang[1], Ivor W. Tsang[2], Rick Siow Mong Goh[1]

[1] *Institute of High Performance Computing, A*STAR, Singapore*

[2] *AAII, University of Technology Sydney, Australia*

* Corresponding Author

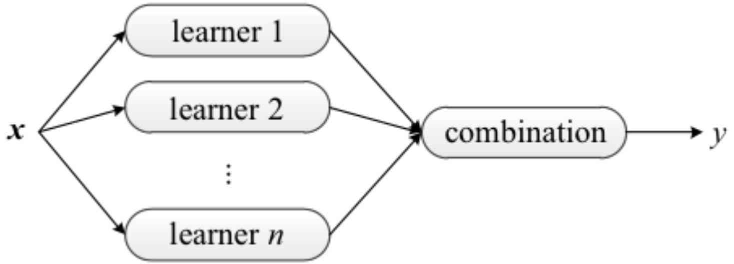
MobiMedia 2020

Ensemble Learning for Multi-class Classification

- Ensemble learning is the process by which multiple models are strategically generated and combined to solve a particular computational intelligence problem.

- An ensemble-based system
  - Combination of diverse models, henceforth classifiers.
  - **Improve** the classification performance and **reduce** the likelihood of an unfortunate selection.

- Ensemble Method:
  - Data-independent ensemble model, e.g. ECOC.



Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

Error Correcting Output Codes (ECOC)

- ECOC
  - An ensemble method designed for multi-class classification problem.
  - A meta method which combines many **binary classifiers**.

- ECOC coding approach aims to construct the ECOC matrix

$$\Lambda \in \{-1,1\}^{N_C \times N_L}$$

Where N_C is the number of classes and N_L is the code length, and its elements are randomly chosen as either −1 or 1.

	M_1	M_2	M_3	M_4	M_5	M_6
C_1	-1	1	-1	1	1	1
C_2	1	1	-1	1	1	1
C_3	1	1	1	1	-1	-1
C_4	-1	1	1	1	-1	1
C_5	1	-1	1	1	-1	-1
C_6	1	1	1	-1	-1	1
C_7	-1	1	-1	-1	1	-1
C_8	1	1	-1	-1	1	1
C_9	1	1	1	-1	1	-1

An example of 6-bit ECOC for a 9-class problem

Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

N-ary Error Correcting Output Codes (N-ary ECOC)

- *N*-ary ECOC
  - An extension of the traditional ECOC methods.
  - Decompose the original classes into N meta-class, where $3 \leq N \leq N_C$.
  - A meta method which combines many **sub-multiclass classifiers**.

- Advantages:
  - More general.
  - Larger row separation.
  - Lower column correlation.

	M_1	M_2	M_3	M_4	M_5	M_6
C_1	1	1	2	4	1	1
C_2	2	1	1	3	2	1
C_3	3	2	1	2	3	1
C_4	4	3	1	1	4	2
C_5	4	3	2	2	4	3
C_6	4	3	3	3	3	4
C_7	3	4	4	4	2	4
C_8	2	4	3	4	2	3
C_9	3	4	2	3	3	2

An example of 6-bit N-ary ECOC for a 9-class problem

Deep *N*-ary ECOC

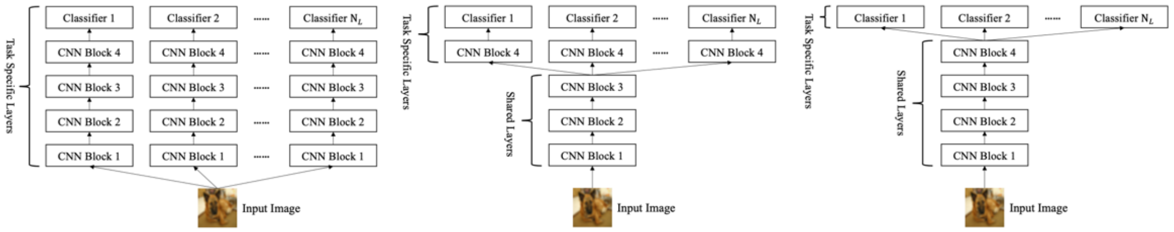
- Traditional ECOC methods:
  - Based on the pre-defined hand-craft features.
  - Focus on how to ensemble the results of base learners on these features.

- Deep N-ary ECOC:
  - Integrate ECOC framework with deep neural networks.

1. Do we necessarily independently train all the deep base learners from scratch for all the situation?
2. Whether the *N*-ary ECOC framework still has advantages over other data-independent ensemble approaches with deep neural network?
3. Any new suggestion on the choice of the meta-class number N and number of base learners N_L?

Deep N-ary ECOC

- Parameter Sharing Strategy
  - No parameter share.
  - Partial parameter share.
  - Full parameter share.
  - The no parameter sharing strategy contains most parameters (N_n), then the partial sharing strategy (N_p) and the full sharing strategy (N_f) is least, say, $N_n > N_p > N_f$.
- For the remaining two questions, we investigate through the experiments.



(a) No Share (b) Partial Share (c) Full Share

Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

Experimental Settings

- Conduct the experiments on 4 image datasets and 2 text datasets
  - Image datasets: MNIST, CIFAR-10, CIFAR-100, FLOWER-102.
  - Text datasets: Text REtrieval Conference (TREC) and Stanford Sentiment Treebank (SST) datasets

Image Dataset					
Dataset	Image Size	# Train Sample	# Dev Sample	# Test Sample	# Classes (N_C)
MNIST	28×28	60,000	N/A	10,000	10
CIFAR-10	32×32	50,000	N/A	10,000	10
CIFAR-100	32×32	50,000	N/A	10,000	100
FLOWER-102	256×256	6,552	818	819	102
Text Dataset					
Dataset	Avg. Sent. Len.	# Train Sample	# Dev Sample	# Test Sample	# Classes (N_C)
TREC	10	5,500	N/A	500	6
SST	18	11,855	N/A	2,210	5

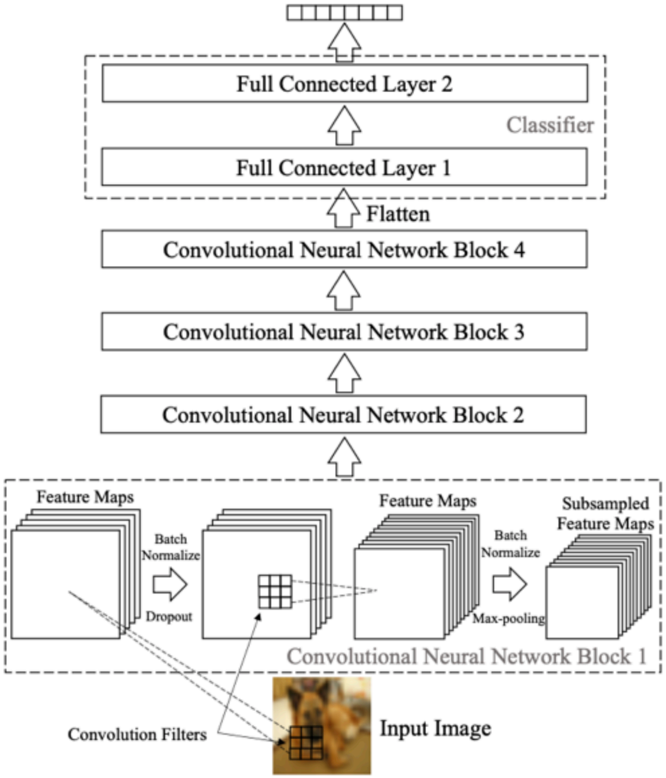
Agency for Science, Technology and Research SINGAPORE

UNIVERSITY OF TECHNOLOGY, SYDNEY

Experimental Settings

❖ Deep Leaning Model for Image Classification

- ➢ LeNet for the MNIST dataset.
- ➢ AlexNet for the FLOWER-102 dataset (pre-trained on ILSVRC dataset).
- ➢ CIFAR-CNN for CIFAR-10/100 datasets.



CIFAR-CNN

Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

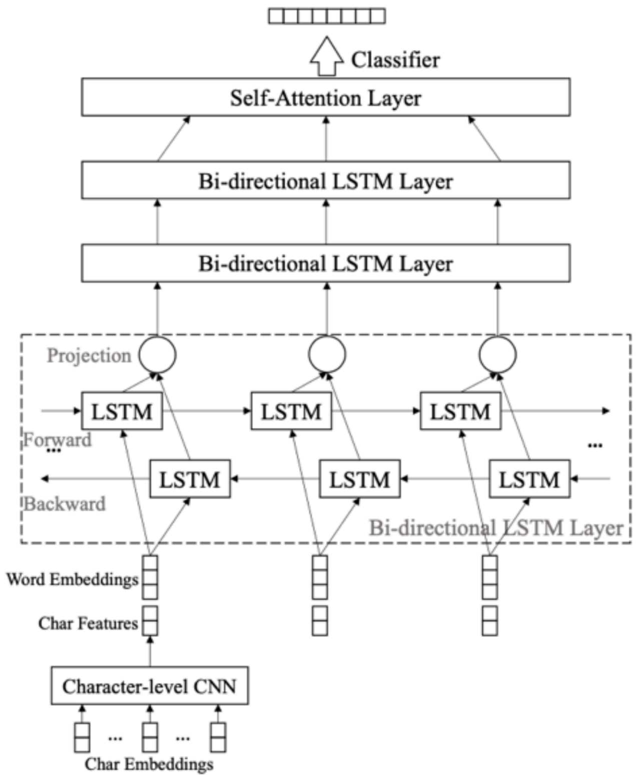
Experimental Settings

❖ Deep Leaning Model for Text Classification

➢ Character-level CNN learned the character features to represent a word from the character sequences of such word.

➢ The word-level Bi-LSTM performs to learn contextual representations .

➢ The self-attention mechanism encodes word feature sequence to a single sentence representation.



Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

Experimental Settings

❖ Summarization of Tested N and N_L for experiments.

Dataset	# Classes (N_C)	Tested # Meta-Class (N)	Tested # Base Learners* (N_L)
MNIST	10	2, 4, 5, 8, 10	60
CIFAR-10	10	2, 4, 5, 8, 10	100
CIFAR-100	100	2, 5, 10, 30, 50, 75, 95, 100	100
FLOWER-102	102	2, 3, 5, 10, 20, 40, 60, 80, 90, 95, 102	60
TREC	6	2, 3, 4, 5, 6	60
SST	5	2, 3, 4, 5	60

* It indicates the maximal number of classifiers is used for training.

Experiments

- Ensemble accuracies of different methods on benchmark datasets.
  - Compared to single model, the improvement ratio of N-ary ECOC is inverse relation with single model performance.
  - The N-ary ECOC scheme outperforms ECOC and ERI ensemble methods on most image and text datasets.

Dataset	Method	Single Model	Ensemble Model*		
			ERI	ECOC	N-ary ECOC (N)
MNIST	LeNet [32]	$98.98 \pm 0.07\%$	$99.11 \pm 0.11\%$	$99.23 \pm 0.08\%$	**99.57** $\pm 0.09\%$
CIFAR-10	CIFAR-CNNs	$87.12 \pm 0.43\%$	$90.54 \pm 0.31\%$	$89.37 \pm 0.54\%$	**91.95** $\pm 0.24\%$
CIFAR-100	CIFAR-CNNs	$61.50 \pm 0.57\%$	$69.57 \pm 0.29\%$	$34.26 \pm 2.42\%$	**69.94** $\pm 0.32\%$
FLOWER-102	AlexNet [15]	$83.12 \pm 0.29\%$	$86.32 \pm 0.60\%$	$77.05 \pm 0.73\%$	**87.94** $\pm 0.28\%$
TREC	Bi-LSTMs	$90.50 \pm 0.12\%$	$94.80 \pm 0.09\%$	$95.80 \pm 0.08\%$	**95.60** $\pm 0.10\%$
SST	Bi-LSTMs	$44.17 \pm 0.92\%$	$48.69 \pm 0.18\%$	$48.91 \pm 0.26\%$	**50.86** $\pm 0.13\%$

* Here N_L are 60, 100, 100, 60, 60 and 60, respectively, for the ensemble models from top to bottom row. While N are 3, 4, 95, 95, 3, 4, respectively, for the N-ary ECOC.

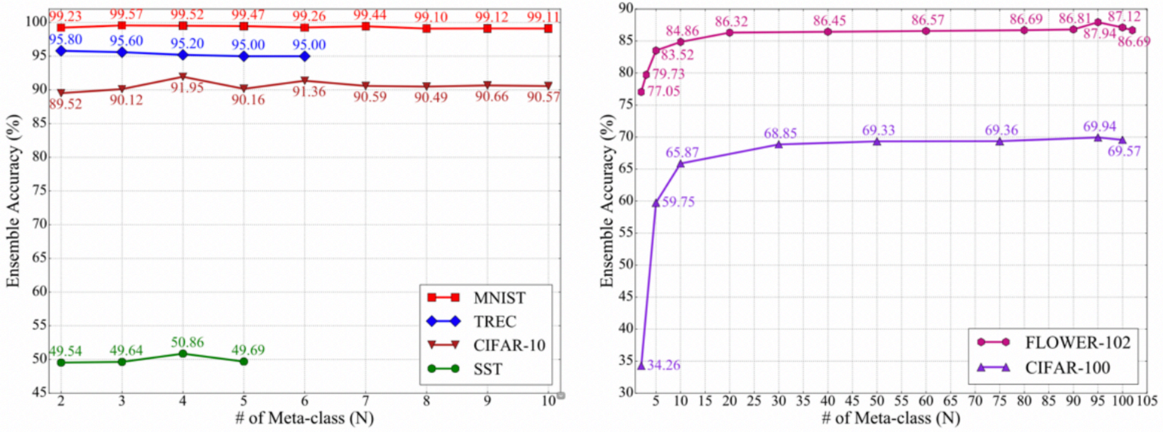
ERI: ensemble of random initialization

Agency for Science, Technology and Research SINGAPORE

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

Experiments

- Evaluation on the Effect of Meta-class Number N.
  - For dataset with small number of N_C, the performances of ensemble models with different N are relatively stable.
  - the performance of ensemble models with different N fluctuates significantly on the datasets with a large value of N_C.



(a) Datasets with small value of N_c

(b) Datasets with large value of N_c

Agency for Science, Technology and Research
SINGAPORE

UTS
UNIVERSITY OF TECHNOLOGY, SYDNEY

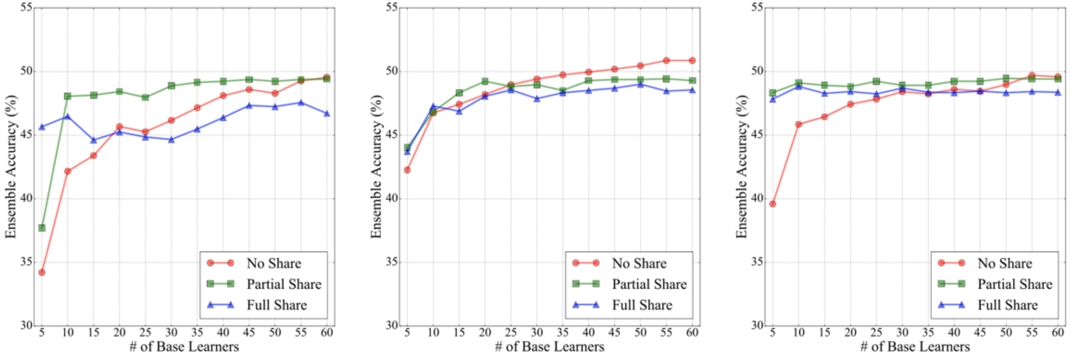
Experiments

- Evaluation on the Effect of Base Learner Number N_L.
  - Smaller number of base learners are required for dataset with small N_C than that of large N_C to reach the optimal ensemble accuracies generally.

Dataset	N	# of Base Learners (N_L) 10	20	30	45	50	60	80	100
MNIST	3	99.14%	99.20%	99.35%	99.48%	99.57%	99.57%	-	-
CIFAR-10	4	87.45%	89.76%	91.78%	91.83%	91.82%	91.92%	**91.95%**	91.93%
CIFAR-100	95	67.94%	69.12%	69.11%	69.33%	69.34%	69.46%	69.67%	**69.94%**
FLOWER-102	95	86.06%	86.45%	86.45%	87.06%	87.16%	**87.94%**	87.46%	87.59%
TREC	3	93.80%	94.00%	95.20%	95.20%	95.60%	**95.60%**	95.50%	**95.60%**
SST	4	46.74%	48.19%	49.41%	50.18%	50.45%	**50.86%**	-	-

Experiments

- Comparison with Three Parameter Sharing Strategies.
  - Take SST dataset as an example.
  - When the number of meta-class *N* is small, both partial and no share models improve significantly with the increase of N_L. The partial share generally outperforms the no and full share except when N_L is less.
  - When the number of meta-class *N* is large, the performance of the three strategies are stable, and the improvement of no share is most significant with the increase of N_L.



(a) ECOC ($N = 2$) (b) N-ary ECOC ($N = 4$) (c) ERI ($N = 5$)

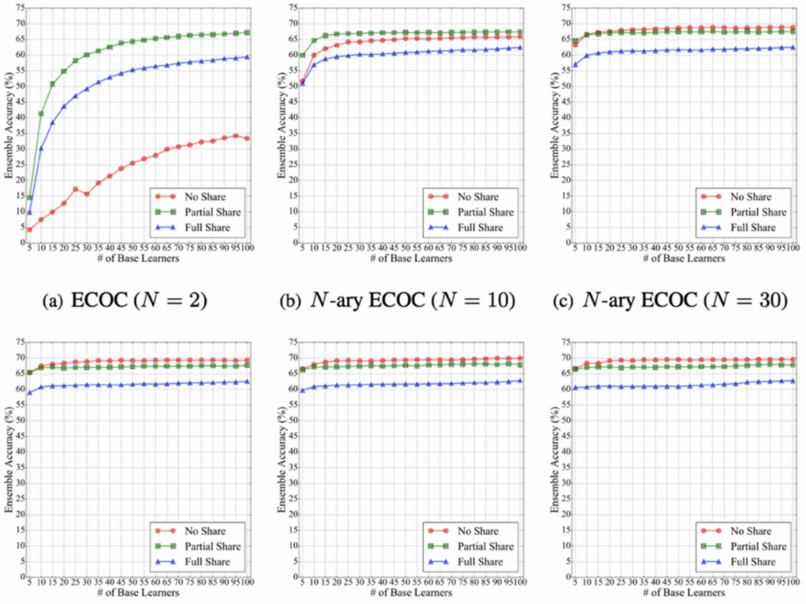
Agency for Science, Technology and Research
SINGAPORE

UTS
UNIVERSITY OF TECHNOLOGY, SYDNEY

Experiments

❖ Comparison with Three Parameter Sharing Strategies.

- Take CIFAR-100 dataset as an example.
- ECOC model with no share strategy fails to achieve satisfactory performance.
- For N-ary ECOC with small N, partial share strategy outperforms no and full share strategies.
- For the ERI model, no share strategy is comparable to partial share when N_L is small. It always performs best when N_L increases, meanwhile, the performance of full share is worst.



(a) ECOC ($N = 2$) (b) N-ary ECOC ($N = 10$) (c) N-ary ECOC ($N = 30$)

(d) N-ary ECOC ($N = 50$) (e) N-ary ECOC ($N = 95$) (f) ERI ($N = 100$)

Conclusion

- For the dataset with small N_C:
  - No share model is better than or equal to the partial share model, thus no share strategy is suggested.
  - When the number of meta-class N is large, these three strategies perform stable.

- For the dataset with large N_C:
  - When the number of meta-class N is small, the performance of partial share model is the best.
  - when the number of meta-class N is large, no share strategy outperforms partial and full share strategies in most cases. Thus no share strategy should be preferred.

- If the number of meta-class is N large, the performance between three sharing strategies is marginal. Then full share could be suggested due to its parameter efficiency.

Agency for Science, Technology and Research
SINGAPORE

UTS
UNIVERSITY OF TECHNOLOGY, SYDNEY

Thank You!