Conditional Mutual Information Constrained Deep Learning: Shannon-Turing Dialogs Continue

En-Hui Yang

Dept. of Electrical and Computer Engineering

University of Waterloo

Joint Work with my students

Outline

Part 1: Background

Part 2: Deep Learning（DL） Challenges

Part 3: How Information Theory（IT) Can Help

Part 4: How DNNs Are Evaluated

Part 5: Normalized Condition Mutual Information (NCMI) & Error Rate

Part 6: CMI Constrained DL

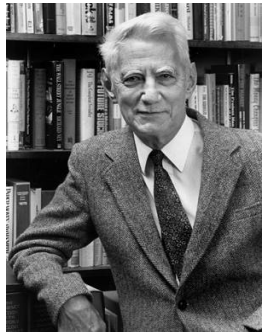
Part 7: Learning Process Visualization Through Lens of CMI and NCMI

Part 8: Experiment Results

Part 1

Background

80 Years Ago: 1943 at Bell Lab



- Wartime research: cryptography
- Planted seeds for our current information age
  - Computation theory---Turing machine (Turing 1936, aged 24)
  - Boolean Algebra (Shannon 1937---Master thesis, aged 21)
  - Intelligent machine (Turing 1948)
  - Information theory (Shannon 1948)

80 Years Later: Deep-Learning Based AI in 2023

熙熙好奇忧虑而来、攘攘满意喜悦而归

- "They come with curiosity and anxiety, and return with satisfaction and joy."

- "Come with curiosity and worry, and go with satisfaction and joy".
- This is a Chinese proverb that expresses the idea of visiting a place with mixed feelings, but leaving with a positive impression. It is often used to describe the experience of tourists or travelers who are attracted by the reputation of a destination, but also have some doubts or fears about it. However, after seeing the beauty and charm of the place, they are happy and content with their choice.

Why Deep-Learning Based AI?

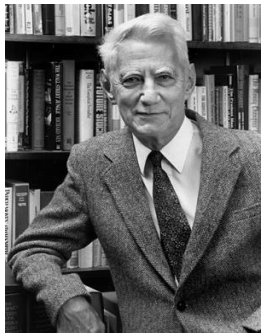
- A deep neural network (DNN) is a high dimension, highly non-linear function.
- Training a DNN is equivalent to solving a huge optimization problem with millions, billions, or trillions of parameters.
- Unlike any other disciplines, it makes little assumption. It just needs data.

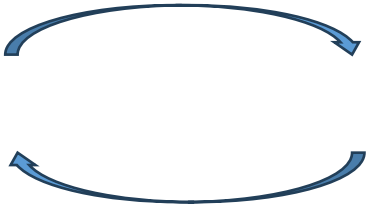
Something never seen in the history

probability theory vs statistics vs deep learning

- Internet ⟹ Mobile (voice) ⟹ Mobile (data) ⟹ Deep-Learning Based AI
- Bigger than Internet.
- It will change our society to a form never experienced or anticipated, including philosophy, education, research, culture, and other human activities.

What Would Shannon and Turing Discuss If They Would Meet Again?





Acquisition

Source Coding

Transmission

Processing

Utilization

Covered partially by IT

CS
AI
Deep Learning (DL)

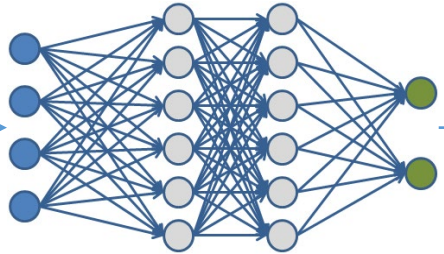
Part 2

DL Challenges

What Is a DNN?

Engineering Perspective

Input x



$$P_{x,\theta} = [P(j|x,\theta)]_{j\in[C]}$$

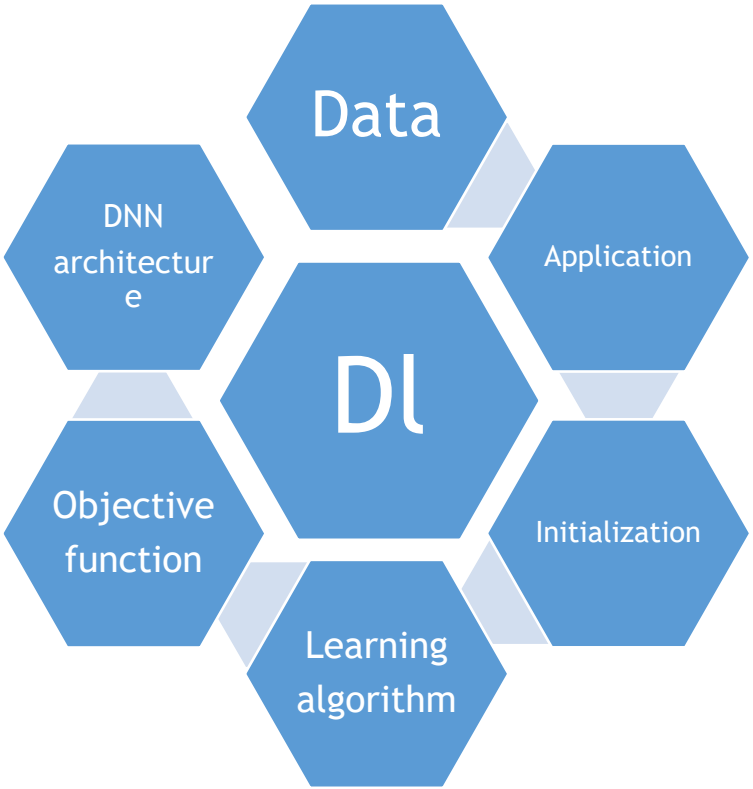
Output Label $c(x) = argmax\ \ P(j|x,\theta)$

Mathematical Perspective

A high dimension, highly nonlinear, complex function:

$$x \quad \rightarrow \quad P_{x,\theta} = [P(j|x,\theta)]_{j\in[C]}$$

How Is a DNN Trained?



$$inf_\theta \; E_{x\sim P}\mathcal{L}_{legacy}(P_{x,\theta}, c_x)$$

e.g.,

$$\mathcal{L}_{legacy}(P_{x,\theta}, c_x) = -ln\, P(c_x|x,\theta)$$

- Huge continuous optimization problem

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, pp. 436-444, 2015.

DL Challenges

1. Very expensive to train a DNN since the dimension of θ is huge (trillions).
2. Very expensive to run a trained DNN as well (millions, billions, or trillions of FLOPS).
3. Lack of understanding of why it works and how it works.
4. Lack of deep theoretic treatment---the fundamental problem of DL has not been touched yet.
5. AI security and robustness are big concerns.

Part 3

How IT Can Help

How IT Can Help: Message for IT Community

IT can help to address Challenges 2 to 5

1. Coding ideas can be embedded into the deep learning process.

2. It is the concept of conditional mutual information (CMI), not mutual information, that is useful to DL.

3. CMI now has not only the usual physical meaning in Shannon sense, but also semantic meaning in deep learning.

4. Putting DNN architectures aside, DL is essentially a new type of IT problem.

How IT Can Help: Message for AI Community

IT can help to address Challenges 2 to 5

1. Information quantities and IT ideas can be used to understand, evaluate, and design DNNs with desired properties.

2. “Dark knowledge” is not dark; it is quite clear---it is manifestations of contextual information of an object, which can be quantified by CMI.

3. The fundamental problem of DL (FPDL) is quite different from DL we know now.

4. The current DL can be regarded as Phase One of the FPDL, and knowledge distillation (KD) with the temperature dropped for the student is an ad hoc way to address FPDL

Part 4

How DNNs Are Evaluated

How DNNs Are Evaluated: Error Rate

- DNN: $x \rightarrow P_{x,\theta} = [P(j|x,\theta)]_{j\in[C]}$
- (X,Y): a pair of random variables, the distribution of which governs a dataset, where X represents the input to the DNN, and Y is the ground truth label of X.
- $P_{Y|X} = [P_{Y|X}(j|X)]_{j\in[C]}$: Conditional Distr. of Y given X.
- $\hat{Y}$: the label randomly predicted by the DNN with the conditional distr. $P_{X,\theta}$ given X.
- $\hat{Y}^*$: the top-1 label predicted by the DNN in response to X.

How DNNs Are Evaluated: Error Rate

- Error rate of the DNN for (X, Y):

$$\varepsilon = \Pr\{\hat{Y} \neq Y\} \qquad and \qquad \varepsilon^* = \Pr\{\hat{Y}^* \neq Y\}$$

Theorem: For any DNN and any (X, Y),

$$\varepsilon \leq \boldsymbol{E}[H(P_{Y|X},\ P_{X,\theta})] \qquad and \qquad \varepsilon^* \leq C\boldsymbol{E}[H(P_{Y|X},\ P_{X,\theta})]$$

where $H(P_{Y|X},\ P_{X,\theta})$ is the cross entropy of $(P_{Y|X},\ P_{X,\theta})$.

Implications on DL ?

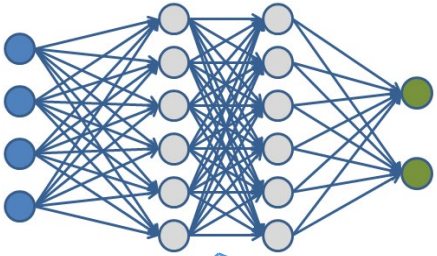
How DNNs Are Evaluated: Beyond Error Rate

Student

Testing scores

- EQ
- Self-discipline
- Other traits

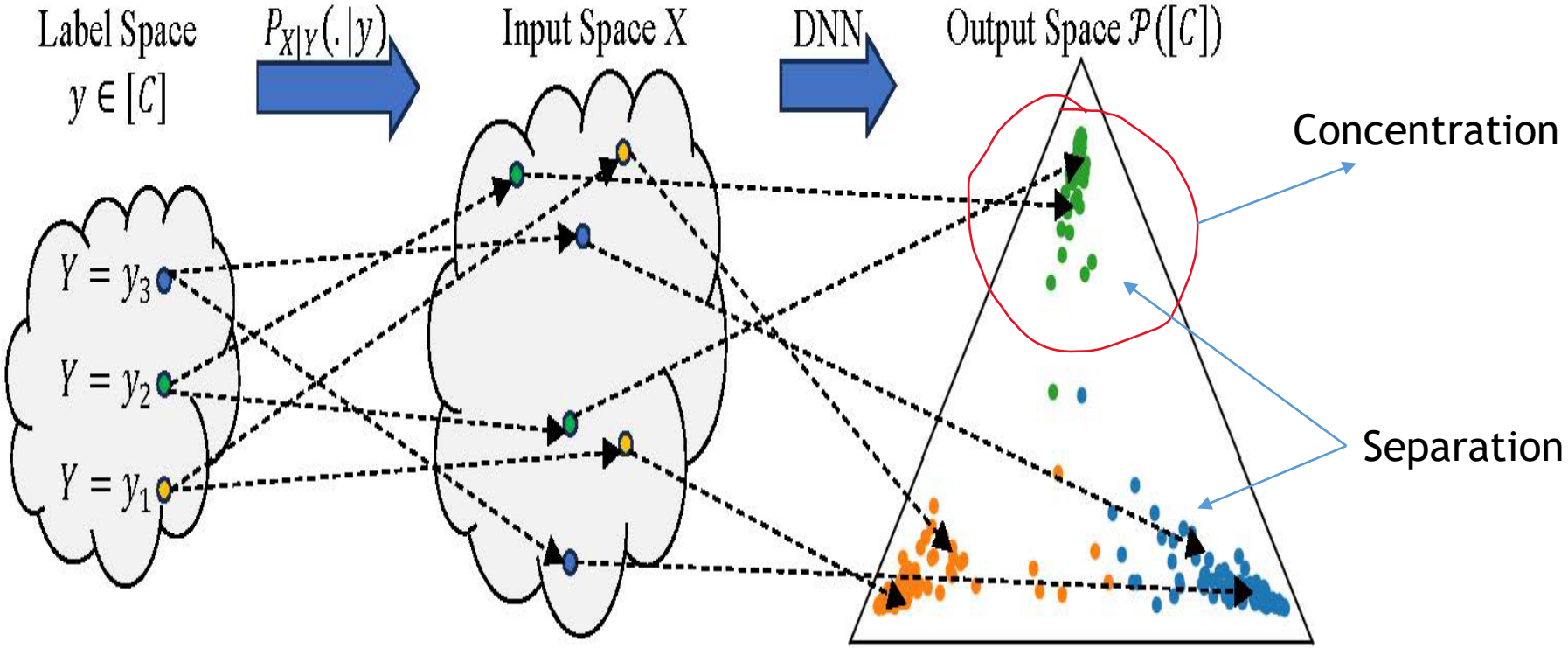
DNN



Error rate

Mapping structure Properties ?

Beyond Error Rate: Concentration and Separation



Concentration: Conditional Mutual Information (CMI)

- $Y \to X \to \hat{Y}$ Forms a Markov chain. Then,

$$
\begin{aligned}
I(X;\hat{Y}|\,Y=y) &= \sum_x P_{X|Y}(x|y)\left[\sum_{i=1}^{C} P(\hat{Y}=i|x)\times\log\frac{P(\hat{Y}=i|x)}{P_{\hat{Y}|y}(\hat{Y}=i|Y=y)}\right] \\
&= \mathbb{E}_{X|Y}\left[\left(\sum_{i=1}^{C} P_X[i]\ln\frac{P_X[i]}{P_{\hat{Y}|y}(\hat{Y}=i|Y=y)}\right)|Y=y\right] \\
&= \mathbb{E}_{X|Y}\left[\mathrm{KL}(P_X||P_{\hat{Y}|y})|Y=y\right].
\end{aligned}
$$

- Average over Y

$$
I(X;\hat{Y}|\,Y) = \sum_{y\in[C]} P_Y(y) I(X;\hat{Y}|y)
$$

- Thus,

$$
\mathrm{CMI}(f) = \mathbb{E}_{(X,Y)}\left[\mathrm{KL}\left(P_X\,||\,Q^Y\right)\right], \qquad \text{with} \quad Q^Y \triangleq \mathbb{E}_{(X|Y)}\left[P_X|\,Y\right].
$$

Separation: Normalized CMI

- The following information quantity is used to quantify the inter-class separation of the DNN:

$$\Gamma = \mathbf{E}\left[I_{\{Y \neq V\}} H(P_X, P_U)\right]$$

where (X, Y) and (U, V) are independent and identically distributed

- Then, the normalized CMI is defined as:

$$\hat{I}(X;\hat{Y}|Y) \triangleq \frac{I(X;\hat{Y}|Y)}{\Gamma}$$

Part 5

NCMI VS ERROR RATE

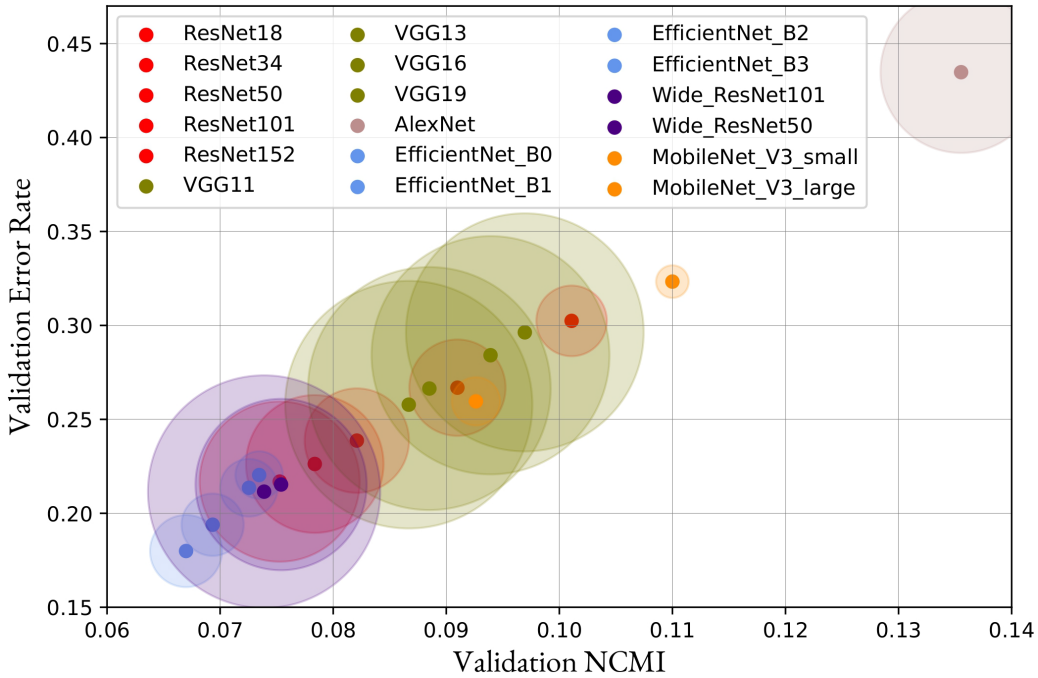
Evaluating DNNs In Terms of CMI and NCMI

- CMI, Γ and NCMI values over the validation set of some pre-trained models on ImageNet along with their error rate ε^* , where the DNNs from the same family are highlighted by the same color.

Models	CMI	Γ	NCMI	Error rate ϵ^*	Models	CMI	Γ	NCMI	Error rate ϵ^*
ResNet18	0.999	9.891	0.101	0.302	AlexNet	1.331	9.830	0.135	0.434
ResNet34	0.902	9.919	0.090	0.266	EfficientNet-B0	0.692	9.433	0.073	0.220
ResNet50	0.815	9.929	0.082	0.238	EfficientNet-B1	0.661	9.114	0.072	0.213
ResNet101	0.779	9.948	0.078	0.226	EfficientNet-B2	0.639	9.224	0.069	0.193
ResNet152	0.749	9.953	0.075	0.216	EfficientNet-B3	0.627	9.365	0.067	0.180
VGG11	0.959	9.899	0.096	0.296	Wide-ResNet50	0.749	9.935	0.075	0.215
VGG13	0.930	9.909	0.094	0.284	Wide-ResNet101	0.734	9.937	0.073	0.211
VGG16	0.878	9.925	0.088	0.266	MobileNet-V3-Small	1.088	9.898	0.110	0.323
VGG19	0.860	9.930	0.086	0.257	MobileNet-V3-Large	0.922	9.956	0.092	0.259

NCMI vs Error Rate

- The error rate vs NCMI value over the validation set of popular pre-trained models on ImageNet. The sizes of the circles represent the sizes of respective models in terms of the number of model parameters; the larger the circle, the larger the model.



Part 6

CMI CONSTRAINED DL

CMIC Deep Learning

- In CMIC-DL, the optimization problem to be solved is as follows:

$$\min_{\theta} \ \mathbf{E}_X \left[H(P_{Y|X}, P_{X,\theta}) \right]$$
$$\text{s.t.} \ \ \hat{I}(X;\hat{Y}|Y) = r,$$

  where r is a positive constant.

- Using the Lagrange multiplier method, the constrained optimization problem becomes

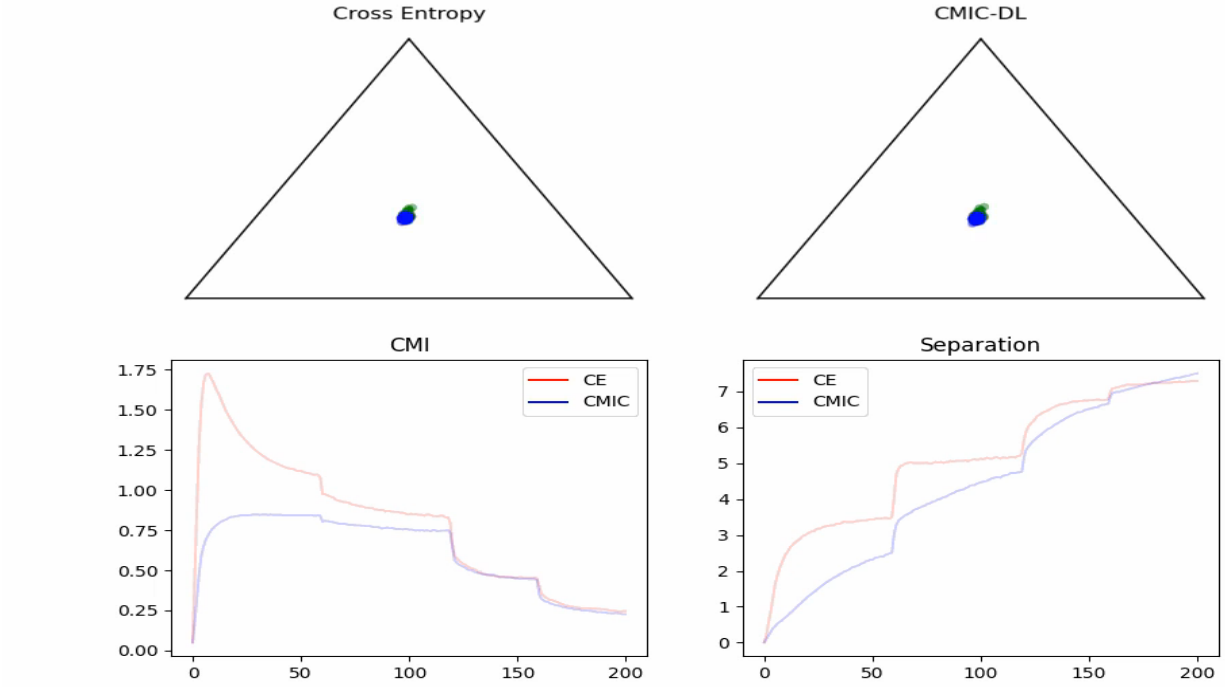
$$\min_{\theta} \ \mathbf{E}_X \left[H(P_{Y|X}, P_{X,\theta}) \right] + \lambda I(X;\hat{Y}|Y) - \beta \mathbf{E} \left[I_{\{Y \neq V\}} H(P_{X,\theta}, P_{U,\theta}) \right]$$

Resemble the rate distortion problem in information theory!

Part 7

LEARNING PROCESS VISUALIZATION

Visualize Learning Processes Through Lens of CMI



Part 8

Experiment Results

Experiments on CIFAR-100

- The validation accuracies (%) of different models trained by CMIC-DL and different benchmark methods over **CIFAR-100**, which are averaged over three different random seeds, and where Bold and underlined values denote the best and second best results, respectively.

Loss	Res32	Res56	Res110	VGG13	WRN-28-10
CL	70.23	72.70	74.20	74.50	80.97
FL	<u>71.62</u>	<u>73.20</u>	74.35	74.53	81.24
LGM	71.50	73.06	<u>74.39</u>	<u>74.57</u>	<u>81.29</u>
OPL	71.03	72.60	73.98	74.11	81.12
CE	70.90	72.40	73.79	73.77	80.93
CMIC	**72.24**	**73.66**	**75.08**	**74.62**	**81.63**

NCMI values over the CIFAR-100 validation set

- The respective NCMI values, measured over the validation set, of the models trained in the previous table via different benchmark methods. The values are averaged over three different runs.

- We use the notation $\hat{I}_{Loss}$ to denote the NCMI value when the underlying DNN is trained using "$Loss$" method.

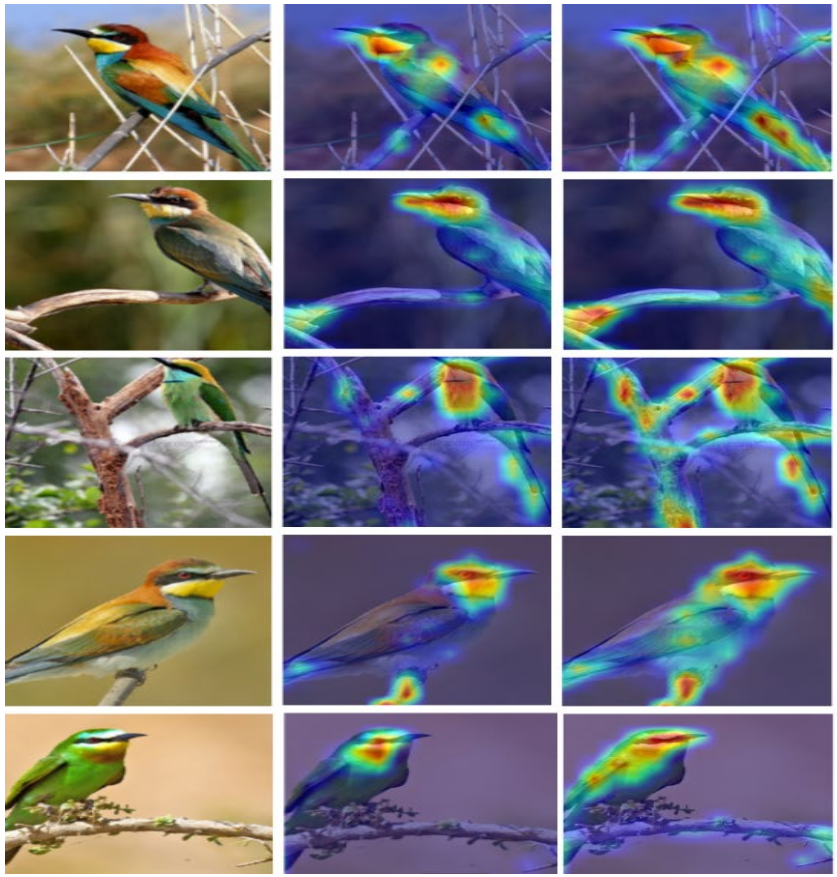
Loss	Res32	Res56	Res110	VGG13	WRN-28-10
$\hat{I}_{\mathrm{CL}}$	0.057	0.045	0.0395	0.0395	0.0309
$\hat{I}_{\mathrm{FL}}$	0.053	0.046	0.0393	0.0399	0.0312
$\hat{I}_{\mathrm{LGM}}$	0.054	0.047	0.0390	0.0398	0.0310
$\hat{I}_{\mathrm{OPL}}$	0.056	0.050	0.0397	0.0402	0.0314
$\hat{I}_{\mathrm{CE}}$	0.057	0.053	0.0402	0.0408	0.0317
$\hat{I}_{\mathrm{CMIC}}$	0.051	0.042	0.0382	0.0392	0.0303

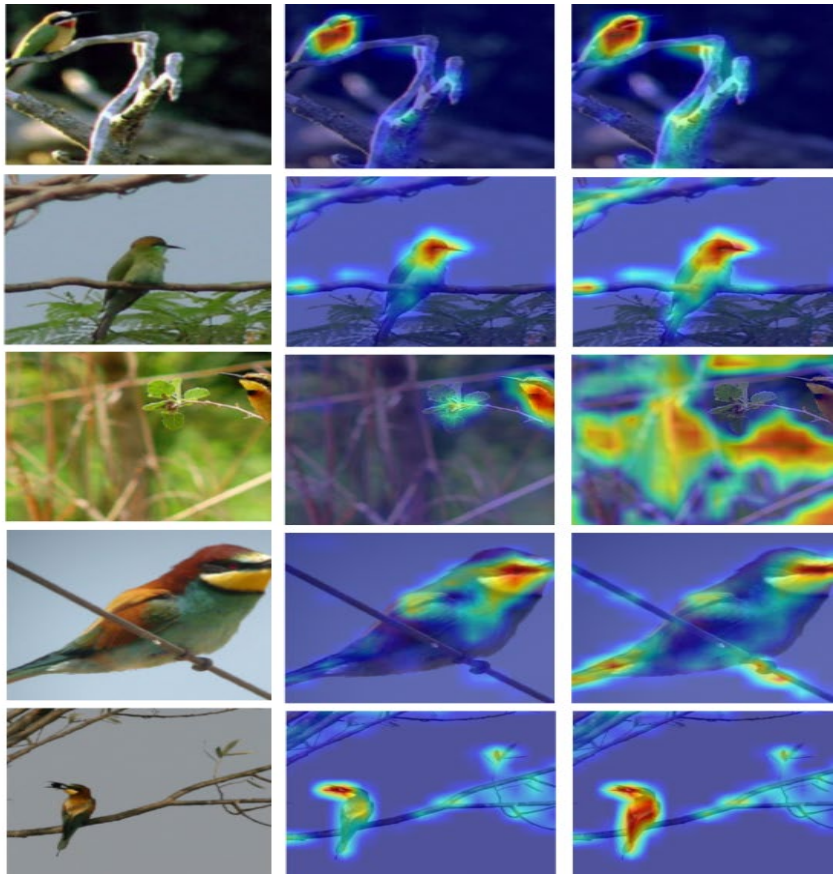
Experiments on ImageNet

- The validation accuracies % of different models trained by CMIC-DL and different benchmark methods on **ImageNet**.

Method	ResNet-18		ResNet-50	
	top-1	top-5	top-1	top-5
CE (Baseline)	69.91	89.08	76.15	92.87
OPL	70.27	89.60	76.32	93.09
CMIC	**70.47**	**89.96**	**76.52**	**93.44**

Semantic Meaning of CMI





Thank You