Introduction

About this tutorial:

- **Title**: IJCAI 2018 Tutorial — Toward Interpretable Deep Learning via Fuzzy Logic
- **Presenters**: Lixin Fan, Chee Seng Chan, Feiyue Wang
- **Time**: 8:30-12:30, July 13 2018
- **Location**: Room T6, the Stockholm Convention Center, Mässvägen 1, Älvsjö, Sweden
- **Link**: http://web.fsktm.um.edu.my/ cschan/ijcai2018
Outline

Introduction (45 mins)

Historical review of fuzzy logic (45 mins)

Bridge the gap: fuzzy logic and deep learning (45 mins)

References
Who we are

- Prof Feiyue Wang (feiyue@ieee.org),
  Professor and Director of The State Key Laboratory for Management and Control of Complex Systems,
  Institute of Automation, Chinese Academy of Sciences
  Google citation (26453)
- Dr Lixin Fan (lixin.fan01@gmail.com),
  Principal scientist at Nokia Technologies, Tampere, Finland
  Google citation (5635)
- Assoc Prof Chee Seng Chan (cs.chan@um.edu.my),
  Associate Professor, University of Malaya,
  Google citation (892)
Scope of the tutorial

Pre-requisites:
- Basic knowledge about proposition logic
- Basic knowledge about deep learning (convolution neural network)

After this tutorial, you are expected to be familiar with following topics:
- Interpretable Neural Networks: A Personal Journal and Perspective
  - Historical overview
  - Interpretable Neural Network
- What fuzzy logic is
  - definition (fuzzy sets, membership functions, fuzzy inference rules etc.)
  - applications (fuzzy control, fuzzy neural networks etc.)
- Connection between fuzzy logic and deep learning
  - CNN and fuzzy neural networks
  - Fuzzy Exclusive OR (fXOR)
  - Generalized hamming networks
Bridge the gap: fuzzy logic and deep learning

Lixin Fan
Nokia Technologies

IJCAI Tutorial, July 13 2018, Stockholm, Sweden
Bridge the gap: fuzzy logic and deep learning

- Black-box nature and risks of artificial neural networks (ANN)
- What is Fuzzy Logic
- Fuzzy Exclusive-Or (fXOR)
  - Why fuzzy XOR
  - Generalized hamming distance
- Generalized Hamming Network (GHN)
  - Deep Epitome for unravelling GHN
  - Rectified Linear Units re-interpretation
  - Batch Normalization re-interpretation
- Take Home Message
Black-box nature and risks of artificial neural networks (ANN)

Deep Convolutional Neural Network

- Network architecture: stacks of layers (up to hundreds or a thousand)
- Internal representation: millions to billions of networks parameters (weight & bias)
- Network techniques: batch-normalization, ReLU etc. proposed on a trail and error basis

Figure: A. Dosovitskiy et al. “FlowNet: Learning Optical Flow with Convolutional Networks” (ICCV 2015)
Black-box nature and risks of artificial neural networks (ANN)

Intuitive visualization & understanding:

- Activation: layer-wise filter visualization (Alexnet 2012)
- Network dreams: enhance features at different layers for given images (Alex et al 2015)

Figure: A. Krizhevsky et al. “ImageNet Classification with Deep Convolutional Neural Networks” (NIPS 2012)
Black-box nature and risks of artificial neural networks (ANN)

Intuitive visualization & understanding:
- Activation: layer-wise filter visualization (Alexnet 2012)
- Network dreams: enhance features at different layers for given images (Alex et al 2015)

Figure: Left: input image; Right: enhanced feature maps
Black-box nature and risks of artificial neural networks (ANN)

Intuitive visualization & understanding:

- Activation: layer-wise filter visualization (Alexnet 2012)
- Network dreams: enhance features at different layers for given images (Alex et al 2015)

Figure: Mysterious feature maps corresponding to different classes (e.g. peacock)
Black-box nature and risks of artificial neural networks (ANN)

Intuitive visualization & understanding:
- Activation: layer-wise filter visualization (Alexnet 2012)
- Network dreams: enhance features at different layers for given images (Alex et al 2015)

Figure: Mysterious feature maps corresponding to different classes (e.g. peacock)
Black-box nature and risks of artificial neural networks (ANN)

Risk: vulnerable to adversarial attacks

- Deep Neural Networks are easily fooled (Szegdy et al 2014, Nguyen et al 2015)
- Adversarial examples in physical world (Kuraki et al 2016)
- 3D printed systematic adversarial attacks (Athalye et al 2017)

Figure: Left: feature maps; Right: 3D printed adversarial attacks.
Black-box nature and risks of artificial neural networks (ANN)

Risk: over-fitting to random (meaningless) labels (Zhang et al ICLR2017)

<table>
<thead>
<tr>
<th>model</th>
<th># params</th>
<th>random crop</th>
<th>weight decay</th>
<th>train accuracy</th>
<th>test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td>yes</td>
<td>yes</td>
<td>100.0</td>
<td>89.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>no</td>
<td>100.0</td>
<td>89.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>yes</td>
<td>100.0</td>
<td>86.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>85.75</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>9.78</td>
</tr>
<tr>
<td>Inception w/o BatchNorm</td>
<td>1,649,402</td>
<td>no</td>
<td>yes</td>
<td>100.0</td>
<td>83.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>82.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>10.12</td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td>yes</td>
<td>yes</td>
<td>99.90</td>
<td>81.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td>no</td>
<td>99.82</td>
<td>79.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>yes</td>
<td>100.0</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>76.07</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>99.82</td>
<td>9.86</td>
</tr>
<tr>
<td>MLP 3x512</td>
<td>1,735,178</td>
<td>no</td>
<td>yes</td>
<td>100.0</td>
<td>53.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>52.39</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>10.48</td>
</tr>
<tr>
<td>MLP 1x512</td>
<td>1,209,866</td>
<td>no</td>
<td>yes</td>
<td>99.80</td>
<td>50.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no</td>
<td>no</td>
<td>100.0</td>
<td>50.51</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>99.34</td>
<td>10.61</td>
</tr>
</tbody>
</table>

Figure: C. Zhang et al. “UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION” (ICLR 2017)
Black-box nature and risks of artificial neural networks (ANN)

Risk: unpredictable & uninterpretable results

- AlphaGo lost one game to South Korean Go player (Lee SD)
  - The winning move was not logically sound, but AlphaGo mysteriously lost its mind

Figure: Left: Lee SD vs Alpha Go; Right: the winning move 78 was not decisive but led to a devastating sequence of moves 87-97.
Black-box nature and risks of artificial neural networks (ANN)

Risk: unpredictable & uninterpretable results

- The learning process looks so mysterious that engineers/programmers start to worship the AI God, who/which writes new verses of the King James Bible (from god.iv.ai):

![Figure: Left: An AI god will emerge by 2042; Right: AI written Bible.](image)

27[2] And he saw that his master's wife, Behold, my land is before thee: dwell where it pleaseth thee. And he lifted up his eyes, and look from the field, and take also of the earth.

24[63] And Isaac came from the field, that were with him, and that every one after his return from the hand of his father rebuked him, and brought her unto the herd, and fetcht a calf tender and good, and make mention of me unto Pharaoh, For to sojourn in the vale of Siddim was full of days: and his clothes in the fourth river is Gihon: the same became mighty men which came in unto them by an interpreter.
Black-box nature and risks of artificial neural networks (ANN)

Fatal cost:

- Despite equipped with advanced sensors and cameras, the autonomous driving system failed to detect the pedestrian on a shadowed street in a recent accident.
- The reason is still unknown: state-of-art of neural networks are able to detect the pedestrian from video images about half second before hand.
- Yet the cost is a 49-year-old woman’s life being taken away!

Figure: Left: Exterior view; Middle: interior view; Right: State of the arts NN detection results.
Black-box nature and risks of artificial neural networks (ANN)

A long standing criticism:
Black-box nature and risks of artificial neural networks (ANN)

A long standing criticism is still valid today:

  - “We are building systems that govern healthcare and mediate our civic dialogue. We would influence elections. I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge, and not on alchemy,” said Rahimi.
  - “The engineering artifacts have almost always preceded the theoretical understanding,” said LeCun. “Understanding (theoretical or otherwise) is a good thing. It’s the very purpose of many of us in the NIPS community. But another important goal is inventing new methods, new techniques, and yes, new tricks.”
Black-box nature and risks of artificial neural networks (ANN)

Related work in a glance:

- **Visualization**: filters, activation maps, relationship etc.
  - A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks (2016) [1]

- **Distillation**: decision trees, forests etc.
  - Distilling a Neural Network Into a Soft Decision Tree (2017) [3]
Black-box nature and risks of artificial neural networks (ANN)

Related work in a glance:

- **SHAP (SHapley Additive exPlanations) based methods:**
  - A Unified Approach to Interpreting Model Predictions [7]
  - Consistent feature attribution for tree ensembles [8]

- **Logic based methods:**
  - A Logical Calculus of Ideas Immanent in Nervous Activity [9]
  - Harnessing Deep Neural Networks with Logic Rules [10]
Black-box nature and risks of artificial neural networks (ANN)

What Fuzzy Logic can offer:

- Better interpretability in terms of logic inference rules
- To model uncertainty with rigorous mathematical tools
- To furnish transfer learning with non-inductive inference
• Black-box nature and risks of artificial neural networks (ANN)
• What is Fuzzy Logic
• Fuzzy Exclusive-Or (fXOR)
  • Why fuzzy XOR
  • Generalized hamming distance
• Generalized Hamming Network (GHN)
  • Deep Epitome for unravelling GHN
  • Rectified Linear Units re-interpretation
  • Batch Normalization re-interpretation
• Take Home Message
What is Fuzzy Logic

It is a logic:

- **Deductive reasoning with propositional logic** e.g. modus ponens as follows:
  - If somebody is tall, then he/she is good at basketball.
  - John is tall
  - Therefore, John is good at basketball.

- **Inductive reasoning (questioned philosophically though):**
  - Yao Ming is tall and he is good at basketball,
  - Shaquille O’Neal is tall and he is good at basketball,
  - .......
  - Therefore, if somebody is tall then he/she is good at basketball.
What is Fuzzy Logic

It is fuzzy

• It aims to model *vague* notions with *rigorous* mathematical tools
• E.g. the proposition “John is 180cm in height, so he is tall”
• ...
What is Fuzzy Logic

It is fuzzy

- It aims to model **vague** notions with **rigorous** mathematical tools
- E.g. the proposition “John is 180cm in height, so he is tall”
- ...

Is 5'11" (180 cm) tall for guy?

A2A, the other answers here pretty much cover the board.

**It depends on where you are**, and different people have different perceptions.

5'11" is above average in nearly every country, excluding some European countries, namely the Netherlands where it is actually below average.

In many European countries, it is average or slightly above, so probably not considered tall by most.

In the U.S. about 1 in 4 men will reach or surpass this height.

In many Eastern countries this is very tall, sometimes exceptionally tall.
What is Fuzzy Logic

It is fuzzy

- It aims to model *vague* notions with *rigorous* mathematical tools
- E.g. the proposition “John is 180cm in height, so he is tall”
- “John is 178cm in height, so he is tall”
- ...
What is Fuzzy Logic

It is fuzzy

- It aims to model **vague** notions with **rigorous** mathematical tools
- E.g. the proposition “John is 180cm in height, so he is tall”
- “John is 178cm in height, so he is tall”
- “John is 175cm in height, so he is tall”
- ...

Bridge the gap: fuzzy logic and deep learning (45 mins)
What is Fuzzy Logic

It is fuzzy

- It aims to model *vague* notions with *rigorous* mathematical tools
- E.g. the proposition “John is 180cm in height, so he is tall”
- “John is 178cm in height, so he is tall”
- “John is 175cm in height, so he is tall”
- “John is 155cm in height, so he is tall”
- ...

Bridge the gap: fuzzy logic and deep learning (45 mins)
What is Fuzzy Logic

It is fuzzy

- It aims to model vague notions with rigorous mathematical tools
- E.g. the proposition “John is 180cm in height, so he is tall”
- No sharp boundaries for “tallness”: a degree of truth values

Principle of bivalence of classical logic:

- every proposition has exactly two possible truth values, either TRUE or FALSE i.e. 0, 1.

Fuzzy logic rejects the principle of bivalence, by extending truth values to

\[ \{0, \frac{1}{2}, 1\} \] \hspace{2cm} \text{three-valued logic,}

\[ \{0, \frac{1}{N}, \frac{2}{N}, \ldots \frac{N}{N}\} \] \hspace{2cm} \text{finitely many-valued logic,}

\[ [0.0, 1.0] \] \hspace{2cm} \text{infinitely many-valued logic,}

\[ (-\infty, +\infty) \] \hspace{2cm} \text{(infinitely) real-number-valued logic.}
Fuzzy sets (Zadeh 1965) and linguistic variable (Zadeh 1975)

- One linguistic variable “age” with three fuzzy sets (very young, young, old):

L. A. ZADEH
Fuzzy operations:

- Fuzzy intersection:
  \[ I(x, w) = \min(x, w) \]

- Fuzzy union:
  \[ U(x, w) = \max(x, w) \]

- Fuzzy negation:
  \[ N(x) = 1 - x \]
Fuzzy operations:

- Many other forms exist for different fuzzy logic

<table>
<thead>
<tr>
<th>t-norm</th>
<th>s-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\min(a, b)$</td>
<td>$\max(a, b)$</td>
</tr>
<tr>
<td>$BD(a, b) = \max(0, a + b - 1)$</td>
<td>$BS(a, b) = \min(1, a + b)$</td>
</tr>
<tr>
<td>$AP(a, b) = ab$</td>
<td>$AS(a, b) = a + b - ab$</td>
</tr>
<tr>
<td>$HP(a, b) = \frac{ab}{a+b-ab}$</td>
<td>$HS(a, b) = \frac{a+b-2ab}{1-ab}$</td>
</tr>
<tr>
<td>$EP(a, b) = \frac{ab}{2-[a+b-(ab)]}$</td>
<td>$ES(a, b) = \frac{a+b}{1+ab}$</td>
</tr>
</tbody>
</table>
Fuzzy exclusive-or (XOR):

- Why fuzzy exclusive-or
  - Exclusive-or and hamming distance
  - Different version of fXOR
- Generalized hamming distance
  - Definition
  - Properties
  - Distributive property and
Fuzzy exclusive-or (XOR):

- In classical logic, exclusive or (or exclusive disjunction) is a logical operation that outputs true only when inputs differ:

\[ p \oplus q = (p \lor q) \land \neg(p \land q) = (p \land \neg q) \lor (\neg p \land q) \]

where \( \land, \lor \) are logical conjunction and disjunction respectively.
- For two binary strings, the hamming distance computes the bitwise XOR of the two strings and counts the number of 1 bits.
- Normalized hamming distance (w.r.t. the string length) can be used to quantify pairwise dis-similarity between binary patterns as follow:

![Hamming Distance Examples](image)
Fuzzy exclusive-or (XOR):

Formally, a function $E : U^2 \rightarrow U$ is a fuzzy XOR if it satisfies the properties (B.C. Bedregal 09):

- **E1:** $E(x, y) = E(y, x)$ (symmetry);
- **E2:** $E(x, E(y, z)) = E(E(x, y), z)$ (associativity);
- **E3:** $E(0, x) = x$ (0-Identity);
- **E4:** $E(1, 1) = 0$ (boundary condition);

where $U = [0, 1]$ is the unitary interval.

There are different forms of fuzzy extension of exclusive or operations:

- $x \oplus y := (1 - x)y + x(1 - y) = x + y - 2xy$
- $x \oplus y := \max(x - y, y - x)$
- $x \oplus y := \min(2 - x - y, x + y)$
Fuzzy exclusive-or (XOR):

Formally, the \textit{generalized hamming distance} \( h(x, y) := x \oplus y = x + y - 2xy \) is an \textit{abelian group} \( H : (R, \oplus) \) satisfying five axioms:

1. \( x \oplus y = (x + y - 2xy) \in R \) \hspace{1cm} \text{(closure);} \\
2. \( x \oplus y = (x + y - 2xy) = y \oplus \) \hspace{1cm} \text{(commutativity);} \\
3. \( (x \oplus y) \oplus z = (x + y - 2xy) + z - 2(x + y - 2xy)z = \) \\
   \hspace{1cm} \hspace{2cm} = x + (y + z - 2yz) - 2x(y + z - 2yz) = x \oplus (y \oplus z) \hspace{1cm} \text{(associativity)} \\
4. \( \exists e = 0 \in R \text{ s.t. } e \oplus x = x \oplus e = (0 + x - 2x \cdot 0) = x \) \hspace{1cm} \text{(identity element);} \\
5. \( \text{for each } x \in R, \exists x^{-1} = x/(2x - 1) \text{ s.t. } x \oplus x^{-1} = e \) \hspace{1cm} \text{(inverse element);} \\

In addition, \( x \oplus 1 = 1 - x \) complements \( x \), \( 0.5 \) is the fixed point since \( x \oplus 0.5 = 0.5 \).
Fuzzy exclusive-or (XOR):

\[ h(x, w) = x + w - 2xw; \quad x, w \in \mathbb{R} \]

\[
\begin{align*}
    h(0, 0) &= h(1, 1) = 0; & h(0, 1) &= h(1, 0) = 1; \\
    h(x, 0) &= x; & h(x, 1) &= 1 - x; & h(x, 0.5) &= 0.5;
\end{align*}
\]

Figure: (a) \( h(x, w) \) (b) Fuzziness; (c) fuzzy membership; (d) partial derivative.
Fuzzy exclusive-or (XOR):

The partial derivative of GHD: $\frac{\partial h}{\partial x} = 1 - 2w$, $x, w \in R$:

- $w = 0.5$ fixed point, information in $x$ is lost;
- $w = 0$ or 1 boundary points, information maintained;
- $w \in [0, 1]$ fuzzy regions, information suppressed;
- $w \in (-\infty, 0)$ negative confidence region, information enhanced;
- $w \in (1, \infty)$ positive confidence region, information enhanced.

A measurement of fuzziness: $F(a) := a \oplus a, R \to (-\infty, 0.5]$:

- $F(a) \geq 0, \forall a \in [0, 1]; F(a) < 0$ otherwise;
- $[0, 1]$ is the fuzzy region and $F(0.5) = 0.5, F(0) = F(1) = 0$;
- $(-\infty, 0), (1, \infty)$ are the negative/positive confident regions respectively;
Fuzzy exclusive-or (XOR):

*Distributive property* of GHD on two sets of binary patterns:

\[ \bar{X} \oplus \bar{Y} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} x^m \oplus y^n \]

\( \bar{X} : \)

\( \bar{X} \oplus \bar{Y} = 0.224 \)

\( \bar{Y} : \)

\( y^1 \)

\( y^2 \)

\( y^3 \)

\( x^1 \) 0.230 0.233 0.227

\( x^2 \) 0.230 0.254 0.247

\( x^3 \) 0.180 0.217 0.203
Fuzzy exclusive-or (XOR):

\[ h(x, w) = \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2 \sum_{i=1}^{K} x_i w_i \]

\[ = \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2xw \]

Figure: Input image X and fXOR mask W
Fuzzy exclusive-or (XOR):

\[
h(x, w) = \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2 \sum_{i=1}^{K} x_i w_i
\]

\[
= \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2xw
\]

\[
= -2(\boxed{b} + xw)
\]
Fuzzy exclusive-or (XOR):

\[ h(x, w) = \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2 \sum_{i=1}^{K} x_i w_i \]

\[ = \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i - 2xw \]

\[ = -2(b + xw) \]
Generalized hamming network:

- Set bias term analytically, instead of learning it:
  \[ b = -\left( \sum_{i=1}^{K} x_i + \sum_{i=1}^{K} w_i \right)/2 \]

- Bias automatically adapts to \( w \) and \( x \) during learning, so that outputs are bounded and symmetric

Figure: output bounded and symmetric?
Generalized hamming network:

- Neuron outputs interpreted as generalized hamming distance between inputs $x$ and neuron weights $w$:
  - Each neuron evaluates the graded truth value of $x \leftrightarrow w$, where $\leftrightarrow$ denotes a fuzzy equivalence relation;
  - When multiple network layers stacked together, neighbouring neuron outputs from the previous layer are integrated to form composite statements e.g. $((x_1^1 \leftrightarrow w_1^1, \ldots, x_i^1 \leftrightarrow w_i^1) \leftrightarrow w_j^2)$ where superscripts correspond to two layers.
  - Thus stacked layers will form more complex and more powerful statements as the layer depth increases.
Deep epitome:

Why deep epitomes?

- We give a *rigorous analysis* of Generalized Hamming Networks (GHN)
- We disclose an interesting finding i.e., *stacked convolution layers is equivalent to a single yet wide convolution layer*.
- Deep epitomes constructed at each layer provide a visualization that *does not rely on the input data*.
- Deep epitomes allow *direct extraction of features in just one step*.

**Theoretical importance**: the revealed equivalence can be regarded as a constructive manifestation of the *universal approximation theorem*. 
Deep epitome: a pictorial overview

Figure: The hamming convolution of two banks of epitomes. Remarks: a) for the inputs $A, B$ the number of epitomes $M_a$ must be the same as the number of channels $C_b$; and for the output bank $M_d = M_b, C_d = C_a, L_d = (L_a + L_b - 1)$. b) the notation $\oplus^*$ refers to the hamming convolution between two banks of epitomes. The convolution of two single-layered epitomes is treated as a special case with all $M_a, C_a, M_b, C_b = 1$. c) the notation $\cup$ refers to the summation of multiple epitomes of the same length. d) multiple (coloured) epitomes in $D$ correspond to different (coloured) epitomes in $B$; and different (shaded) channels in $D$ correspond to different (shaded) channels of inputs in $A$. 
Deep epitome: merging two epitomes

**Definition**

The *composite convolution* of two banks of epitomes \([M_a A_c^{L_a}]\) and \([M_b B_c^{L_b}]\) with \(M_a = C_b\), is defined as

\[
[M_a A_c^{L_a}] \bigoplus [M_b B_c^{L_b}] := \left\{ m_{a,c}^{A_c^{L_a}} \bigoplus m_{b,c}^{B_c^{L_b}} \right\}_{c_a = 1, \ldots, C_a; m_b = 1, \ldots, M_b}.
\]  

The output of this operation, in turn, is a bank with \(M_b\) of \(C_a\)-channel length-(\(L_a + L_b - 1\)) epitomes denoted as \([M_d D_c^{L_d}]\) with \(M_d = M_b, C_d = C_a, L_d = L_a + L_b - 1\). See figure on the previous slide for an example.
The main result

**Theorem**

A generalized hamming network consisting of multiple convolution layers, is equivalent to a bank of epitome, called deep epitome $[* \mathcal{D} \nabla]$, which can be computed by iteratively applying the composite hamming convolution in equation (1) to individual layer of epitomes:

$$[* \mathcal{D} \nabla] := \bigoplus \big[ M_a A^L_{C_a} \big] \bigotimes \big[ M_b B^L_{C_b} \big] \bigotimes \ldots \bigotimes \big[ M_z Z^L_{C_z} \big],$$

(2)

in which $\nabla = C_a$ is the number of channels in the first bank $A$, $\star = M_z$ is the number of epitomes in the last bank $Z$, and $\diamond = L_a + (L_b - 1) + \ldots + (L_z - 1)$ is the length of composite deep epitome.
Deep epitome: delving deeper

<table>
<thead>
<tr>
<th>$X^3: x_1 \ x_2 \ x_3$</th>
<th>$X^3 \oplus A^2$</th>
<th>$X^3 \oplus^* A^2$</th>
<th>$s_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_2$</td>
<td>${x_1 \oplus a_2}$</td>
<td>$(g_1, \ 1)$</td>
<td></td>
</tr>
<tr>
<td>$A^2: a_1 \ a_2$</td>
<td>${x_1 \oplus a_1, \ x_2 \oplus a_2}$</td>
<td>$(g_2, \ 2)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>${x_2 \oplus a_1, \ x_3 \oplus a_2}$</td>
<td>$(g_3, \ 2)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>${x_3 \oplus a_1}$</td>
<td>$(g_4, \ 1)$</td>
<td></td>
</tr>
<tr>
<td>$g_1 \ g_2 \ g_3 \ g_4$</td>
<td>$X^3 \oplus A^2 \oplus B^2$</td>
<td>$X^3 \oplus^* A^2 \oplus^* B^2$</td>
<td></td>
</tr>
<tr>
<td>$b_2$</td>
<td>${x_1 \oplus a_2 \oplus b_2}$</td>
<td>$(g_1, \ 1)$</td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td>${x_1 \oplus a_2 \oplus b_1, \ x_2 \oplus a_2 \oplus b_2}$</td>
<td>$(g_2, \ 3)$</td>
<td></td>
</tr>
<tr>
<td>$B^2: b_1$</td>
<td>${x_1 \oplus a_1 \oplus b_1, \ x_2 \oplus a_2 \oplus b_1}$</td>
<td>$(g_3, \ 4)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>${x_2 \oplus a_1 \oplus b_1, \ x_3 \oplus a_2 \oplus b_1}$</td>
<td>$(g_4, \ 3)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>${x_3 \oplus a_1 \oplus b_1}$</td>
<td>$(g_5, \ 1)$</td>
<td></td>
</tr>
</tbody>
</table>

Figure: **Left panel**: example tuples $X^3, A^2, B^2$; **Middle**: Hamming outer products $X^3 \oplus A^2, X^3 \oplus^* A^2 \oplus B^2$; **Right**: Hamming convolutions $X^3 \oplus^* A^2, X^3 \oplus^* A^2 \oplus^* B^2$ and corresponding epitomes. Indices $1, 2, ..., 1, 2, ...$ denote subsets $S(1), S(2) ... S(n)$ in which element indices satisfying $k + (L - l) = n$ and $k + (L - l) + (M - m) = n$. 

Bridge the gap: fuzzy logic and deep learning (45 mins)

Lixin Fan,
Visualization of deep epitomes (MNIST):

Figure: Left to right: deep epitomes at layers 1, 2 and 3 for a GHN trained with MNIST classification at iterations 10000. Note oriented strokes as hand written features.

Figure: Corresponding activations for a given MNIST digit image.
Visualization of deep epitomes (CIFAR10):

Figure: Left to right: deep epitomes at layers 1, 2 and 3 for a GHN trained with CIFAR10 classification at iterations 180000. Note different types of salient features e.g. oriented edgelets, textons with associated colours or even rough segmentations.

Figure: Corresponding activations for a given CIFAR10 image.
Rectified Linear Units:

Rectified linear unit in a nutshell:

- Rectified linear unit (ReLU), has been widely used due to its strong biological motivations and mathematical justification (Hahnloser et al. 2000):

\[ f(x) = x^+ = \max(0, x) \]

- Other activation functions like Leaky Rectifier Linear Unit (LReLU) (Maas et al., 2013), Parametric Rectifier Linear Unit (PReLU) (He et al., 2016b) and Exponential Linear Unit (ELU) (Clevert et al., 2015) were proposed to improve ReLU for various considerations.

- There is no consensus about how these proposed nonlinearities compare to ReLU, and more fundamentally, how to interpret contributions of nonlinearities to the successful learning of deep neural networks.
Rectified Linear Units:

ReLU re-interpreted in the lens of fuzzy logic:

- Thresholding of generalized hamming distance
- Thresholding from one side only $\rightarrow \max(0, x)$
- Double-thresholding from both sides demonstrated with improved learning performance (fast convergence rate) empirically

Figure: Double thresholding scheme.
Rectified Linear Units:

ReLU re-interpreted in the lens of fuzzy logic:

- Thresholding of generalized hamming distance
- Thresholding from one side only $\rightarrow \max(0, x)$
- Double-thresholding from both sides demonstrated with improved learning performance for challenging tasks i.e. CIFAR10/100 classification

![Graphs showing accuracy vs log(#mini_batch) for CIFAR10 and CIFAR100 classification](image)

Figure 4: Left: GHN test accuracies of CIFAR10 classification (OPT THRES: parameter $r$ optimized; WO THRES: without nonlinear activation). Right: GHN test accuracies of CIFAR100 classification (OPT THRES: parameter $r$ optimized; WO THRES: without non-linear activation).
Rectified Linear Units:

ReLU re-interpreted in the lens of fuzzy logic:
- Thresholding of generalized hamming distance
- Thresholding from one side only $\rightarrow \max(0, x)$
- Double-thresholding from both sides demonstrated with improved learning performance for challenging tasks i.e. CIFAR10/100 classification
- GHN without non-linear activation (by setting $r = 0$) performs equally well for simple tasks e.g. MNIST classification.

![Graphs showing test accuracies of MNIST classification with GHN.](image)

Figure 3: Test accuracies of MNIST classification with Generalized Hamming Network (GHN). Left: test accuracies without using non-linear activation (by setting $r = 0$). Middle: with $r$ optimized for each layer. Right: with $r$ optimized for each filter. Four learning rates i.e. $\{0.1, 0.05, 0.025, 0.01\}$ are used for each case with the final accuracy reported in brackets. Note that the number of mini-batch are in logarithmic scale along x-axis.
- Batch normalization (BN, Ioffe & Segedy 2015) has been used to stabilize neuron outputs while training, and thus leads to improved learning speed.

- The change in the distributions of internal nodes, referred to as *Internal Covariate Shift*, may cause serious problems during the training stage;

- By normalizing activations throughout the network, it prevents small changes to the parameters from amplifying into larger and suboptimal changes in activations in gradients;

- Batch Normalization also makes training more resilient to the parameter scale.
Batch Normalization:

- Batch normalization (BN, Ioffe & Segedy 2015) has been used to stabilize neuron outputs while training, and thus leads to improved learning speed.

\begin{figure}
\centering
\begin{align*}
\text{Input:} & \quad \text{Values of } x \text{ over a mini-batch: } B = \{x_1 \ldots m\}; \\
& \quad \text{Parameters to be learned: } \gamma, \beta \\
\text{Output:} & \quad \{y_i = \text{BN}_{\gamma,\beta}(x_i)\}
\end{align*}
\begin{align*}
\mu_B & \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i & \quad & \text{// mini-batch mean} \\
\sigma_B^2 & \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 & \quad & \text{// mini-batch variance} \\
\hat{x}_i & \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} & \quad & \text{// normalize} \\
y_i & \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) & \quad & \text{// scale and shift}
\end{align*}
\end{figure}
Batch Normalization:

- Batch normalization is re-interpreted in the lens of fuzzy logic and the fast learning speed is achieved with GHN without using BN:
  
  - No need to learn hyper-parameters of BN
  - Since bias terms adapt to input and weights (and outputs are bounded)

Figure 2: Left to right: mean, max and min of neuron outputs, with/without batch normalized (BN, WO_BN) and generalized hamming distance (XOR). Outputs are averaged over all 64 filters in the first convolution layer and plotted for 30 epochs training of a MNIST network used in our experiment (see Section 4).
Take home messages:

- We set bias terms analytically by following fuzzy XOR inference rule strictly.
- Neuron outputs are interpreted as generalized hamming distance between inputs $x$ and weights $w$.
- A double-thresholding method for ReLU demonstrates fast learning speed for challenging learning problems.
- ReLU plays crucial role in improving the learning performances for challenging tasks.
- Yet it can be skipped in GHN for simple learning problems with negligible performance differences.
- Batch-normalization with its sophisticated learning method may also be skipped for Generalized Hamming Networks, which adopt to input signal by using fuzzy XOR operation.
Q&A
References:

A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks. 


N. Frosst and G. Hinton. 
Distilling a Neural Network Into a Soft Decision Tree. 

M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez. 
Beyond Sparsity: Tree Regularization of Deep Models for Interpretability. 
Q. Zhang, Y. Yang, Y. Nian Wu, and S.-C. Zhu.  
Interpreting CNNs via Decision Trees.  

Q. Zhang and S.-C. Zhu.  

Scott M Lundberg and Su-In Lee.  
A unified approach to interpreting model predictions.  
In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,  
S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information  

Consistent feature attribution for tree ensembles.  

Warren McCulloch and Walter Pitts.  
A logical calculus of ideas immanent in nervous activity.  
Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing.
Harnessing deep neural networks with logic rules.

Lixin Fan.
Revisit fuzzy neural network: Demystifying batch normalization and ReLU with generalized hamming network.

L. Fan.