COSC 522 – Machine Learning

Lecture 16 – From Machine Learning to Deep Learning

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu

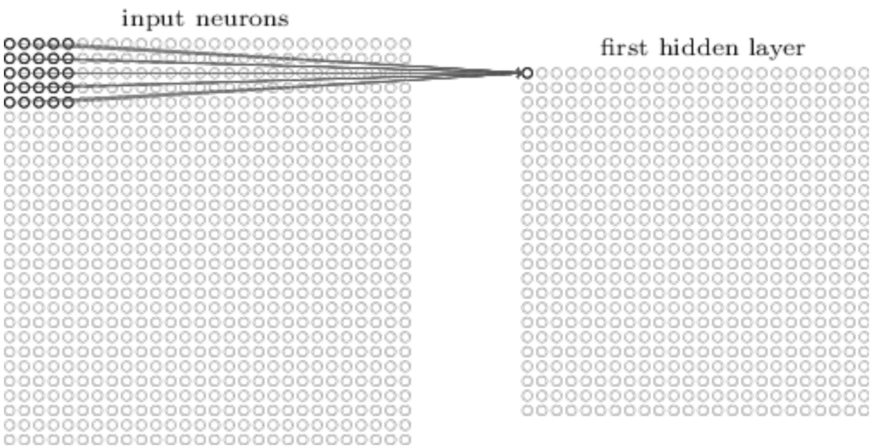
Course Website: http://web.eecs.utk.edu/~hqi/cosc522/

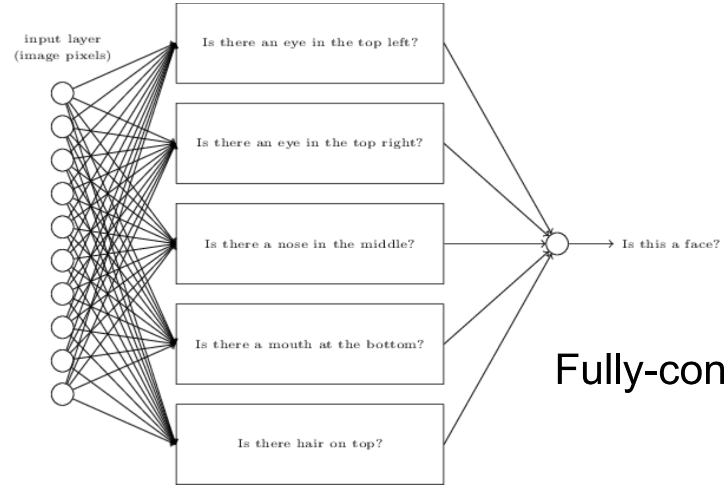
A list of misconceptions

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation

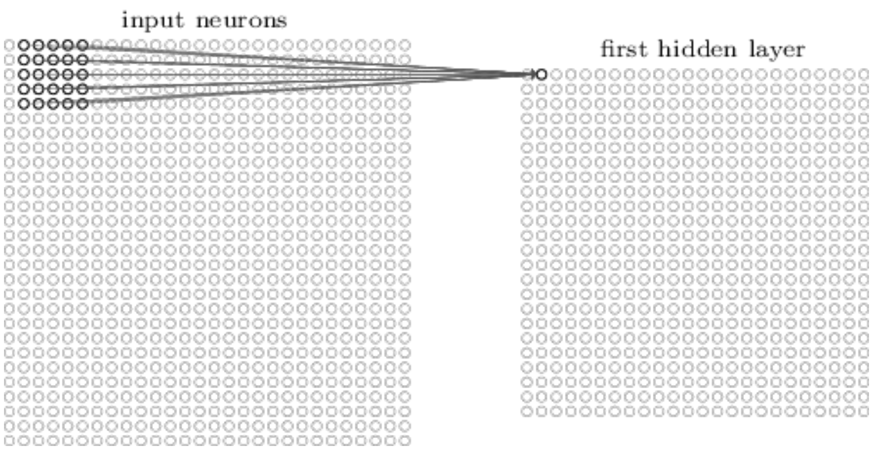
AICIP RESEARCH

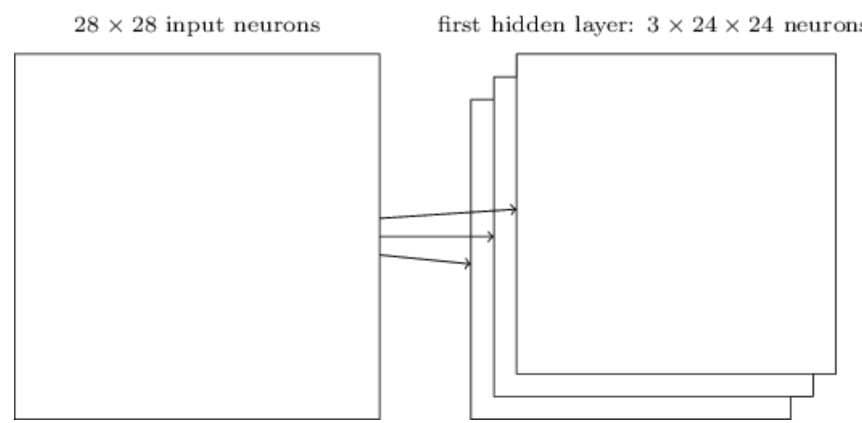
Core idea 1: Receptive field (RF) and shared weight





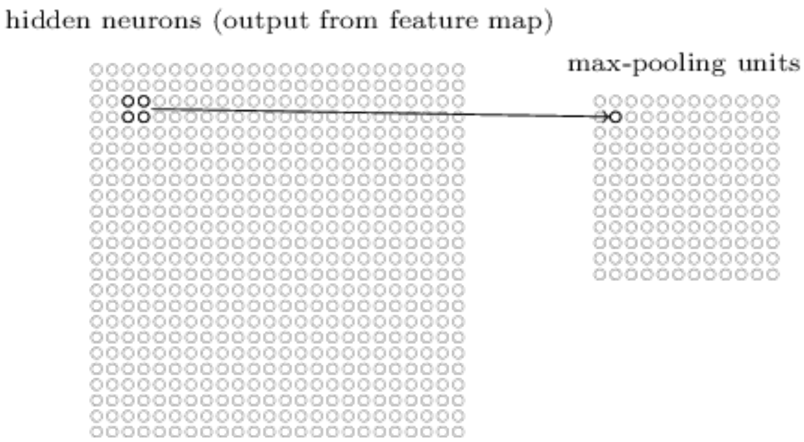
Fully-connected

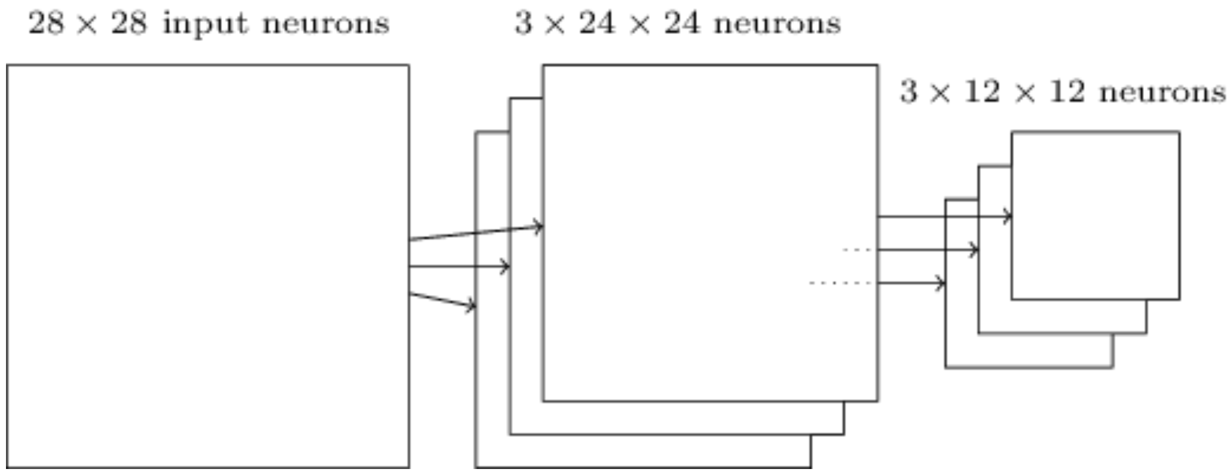




Feature maps

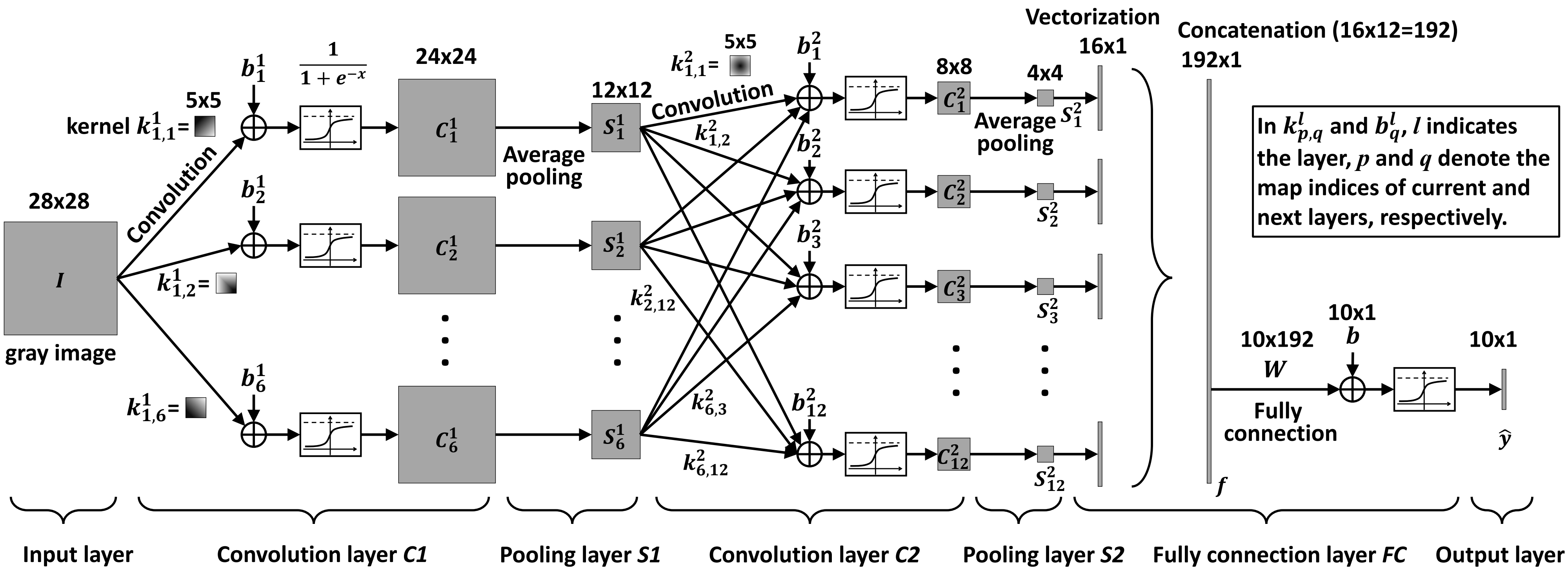
Core idea 2: Hierarchical vision - Max pooling





A simple CNN framework

AICIP RESEARCH

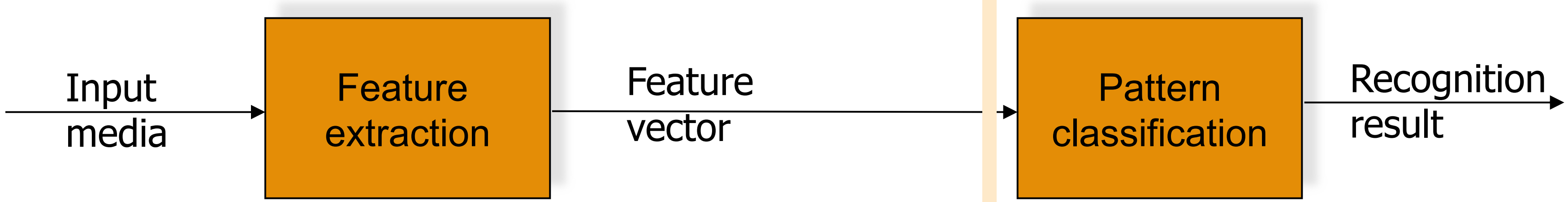


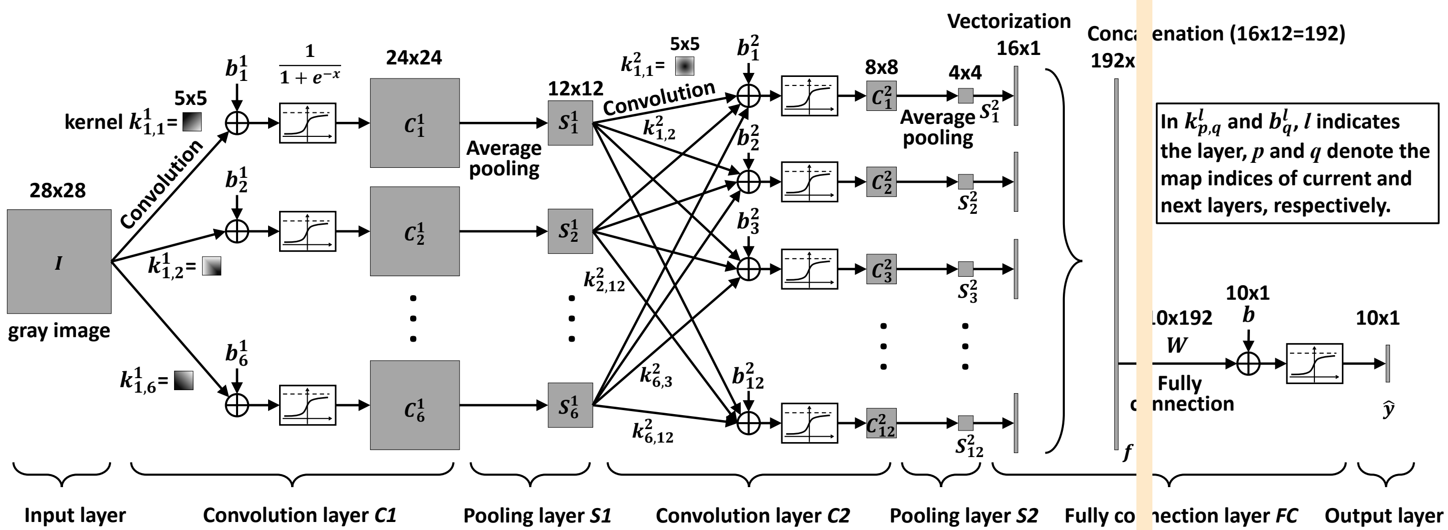
A list of misconceptions

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation

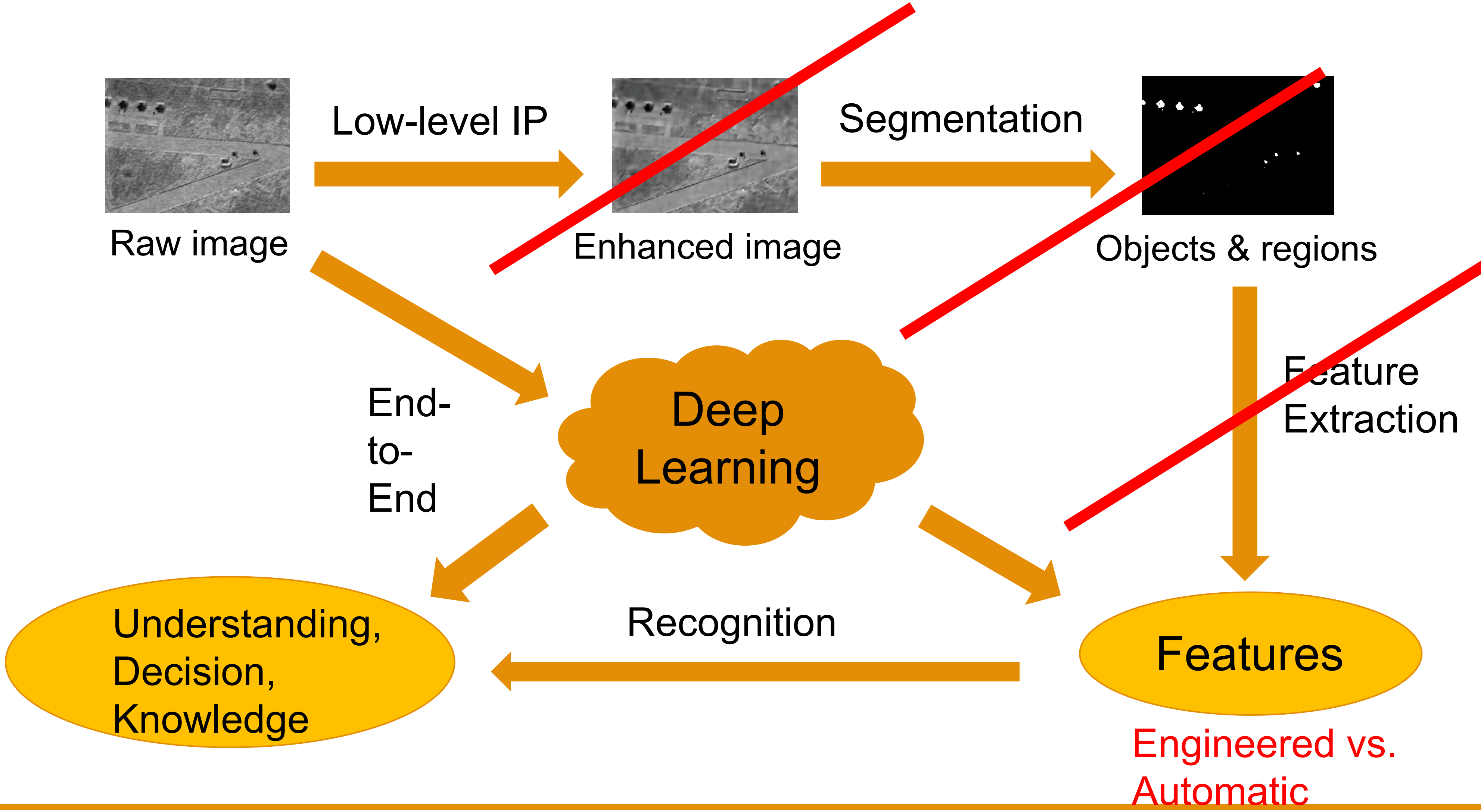
Engineered features vs. automatic features

Need domain knowledge





The flowchart comparison



A list of misconceptions

AICIP RESEARCH

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation

AICIP RESEARCH

Revisit: A bit of history

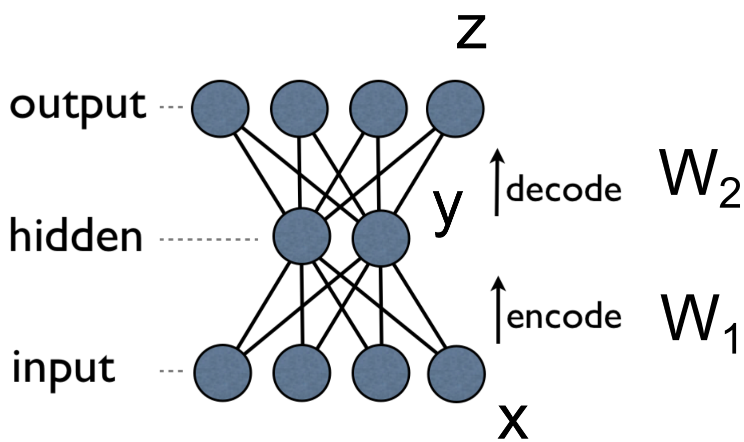
- 1956-1976
  - 1956, The Dartmouth Summer Research Project on Artificial Intelligence, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College ... The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

  - The rise of symbolic methods, systems focused on limited domains, deductive vs. inductive systems
  - 1973, the Lighthill report by James Lighthill, "Artificial Intelligence: A General Survey" - automata, robotics, neural network
  - 1976, the AI Winter
- 1976-2006
  - 1986, BP algorithm
  - ~1995, The Fifth Generation Computer
- 2006-???
  - 2006, Hinton (U. of Toronto), Bingio (U. of Montreal), LeCun (NYU)
  - 2012, ImageNet by Fei-Fei Li (2010-2017) and AlexNet

https://en.wikipedia.org/wiki/Dartmouth_workshop
https://en.wikipedia.org/wiki/Lighthill_report

Unsupervised learning – Autoencoder (AE)



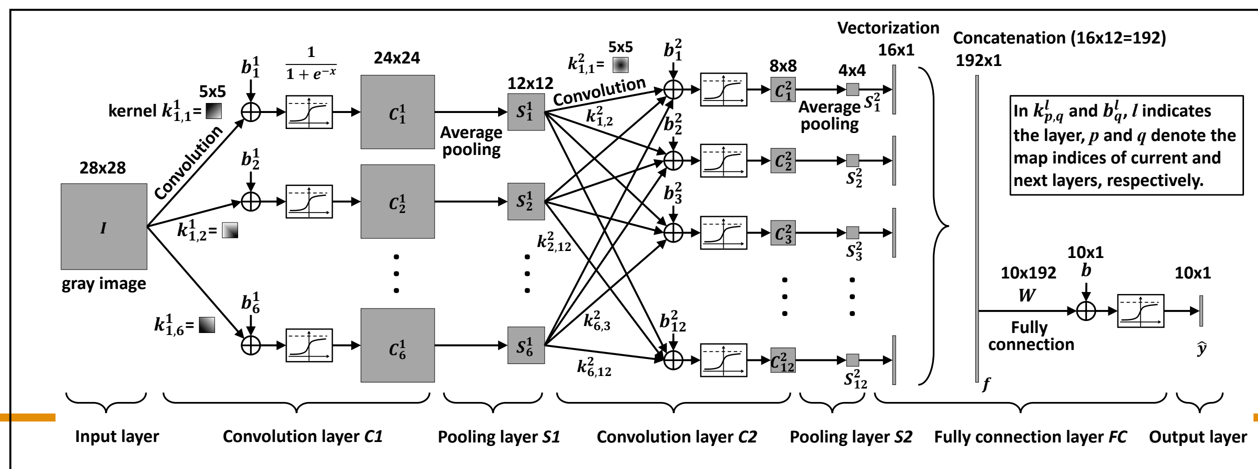
$$\mathbf{y} = f_{\theta_1}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$$

$$\mathbf{z} = g_{\theta_2}(\mathbf{W}_2\mathbf{y} + \mathbf{b}_2)$$

$$\theta_1 = \{\mathbf{W}_1, \mathbf{b}_1\}, \theta_2 = \{\mathbf{W}_2, \mathbf{b}_2\}$$

$$\theta_1^*, \theta_2^* = \arg\min_{\theta_1,\theta_2} \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \arg\min_{\theta_1,\theta_2} \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{x}^{(i)}, g_{\theta_2}(f_{\theta_1}(\mathbf{x}^{(i)})))$$

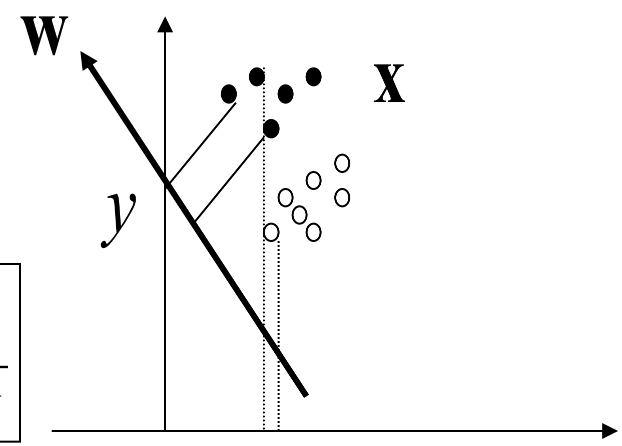
$$L(\mathbf{x}, \mathbf{z}) = \frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2$$



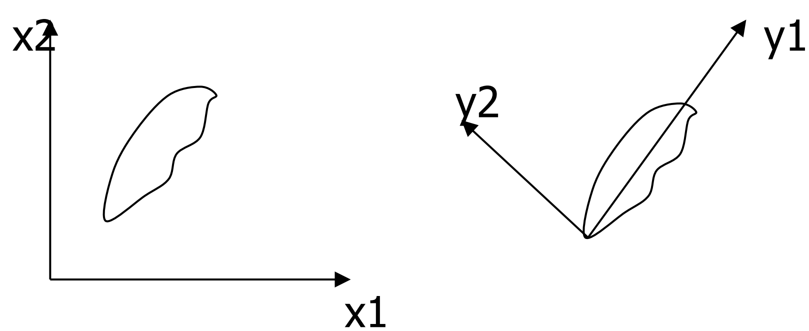
Discrimination vs. Representation of Data

- Best discriminating the data
  - Fisher's linear discriminant (FLD)
  - NN
  - CNN

$$J(\mathbf{w}) = \frac{\left|\tilde{m}_1 - \tilde{m}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\left|\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\right|^2}{\mathbf{w}^T\mathbf{S}_1\mathbf{w} + \mathbf{w}^T\mathbf{S}_2\mathbf{w}} = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$



- Best representing the data
  - Principal component analysis (PCA)



$$\mathbf{x} = \sum_{i=1}^{m} y_i \mathbf{b}_i + \sum_{i=m+1}^{d} y_i \mathbf{b}_i \approx \sum_{i=1}^{m} y_i \mathbf{b}_i + \sum_{i=m+1}^{d} \alpha_i \mathbf{b}_i$$

Error: $$\Delta\mathbf{x} = \sum_{i=m+1}^{d} (y_i - \alpha_i)\mathbf{b}_i$$

PCA as Linear Autoencoder

Raw data (X_{nxd}) → covariance matrix (Σ_X) → eigenvalue decomposition (λ_{dx1} and E_{dxd}) → principal component (P_{dxm}) → $Y_{nxm} = X_{nxd} * P_{dxm}$

The two papers in 2006

- [Hinton:2006a] G.E. Hinton, S. Osindero, Y.W. Teh, “A fast learning algorithm for deep belief nets,” Neural Computation, 18(7):1527-1554, 2006.
- [Hinton:2006b] G.E. Hinton, R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, 313:504-507, July 2006.

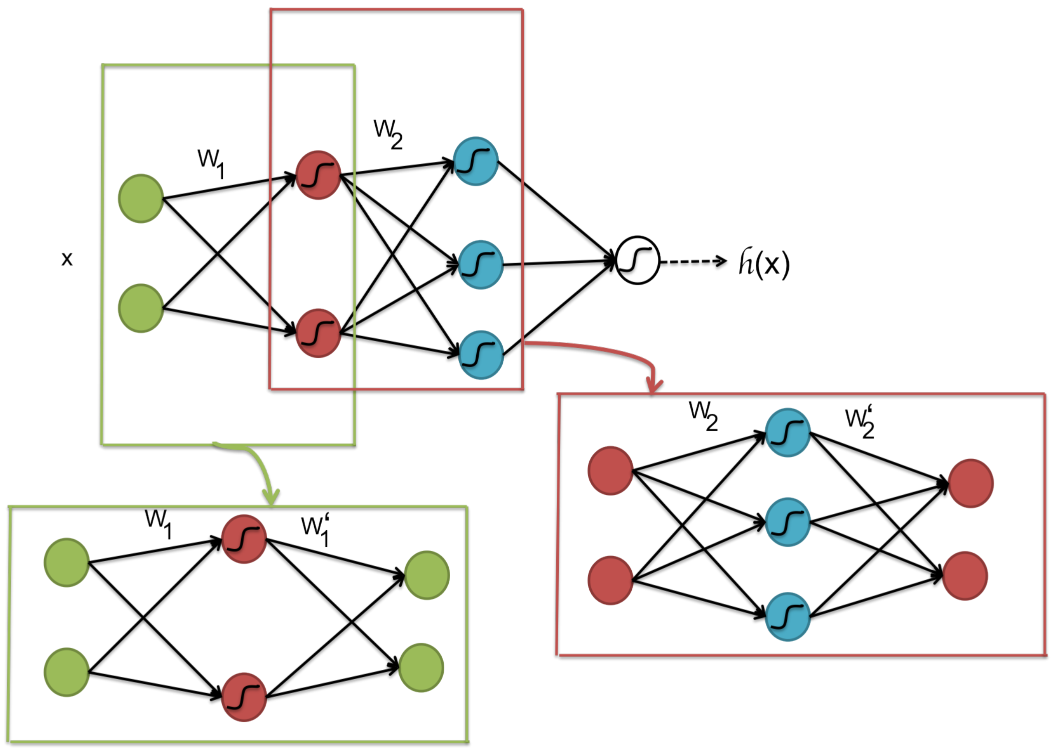
Techniques to avoid overfitting

- Regularization
  - Weight decay or L1/L2 normalization
  - Use dropout
  - Data augmentation
- Use unlabeled data to train a different network and then use the weight to initialize our network
  - Deep belief networks (based on restricted Boltzmann Machine or RBM)
  - Deep autoencoders (based on autoencoder)

AE as pretraining methods

AICIP RESEARCH

- Pretraining step
  - Train a sequence of shallow autoencoders, greedily one layer at a time, using unsupervised data
- Fine-tuning step 1
  - Train the last layer using supervised data
- Fine-tuning step 2
  - Use backpropagation to fine-tune the entire network using supervised data

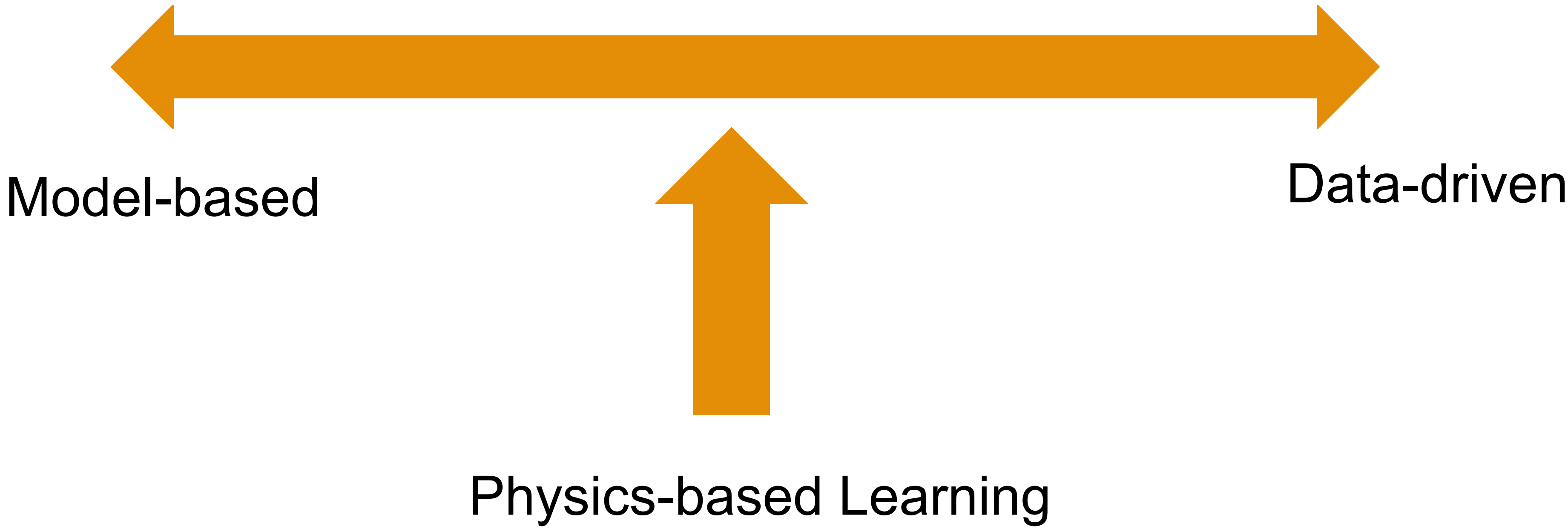


A list of misconceptions

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation

AICIP RESEARCH

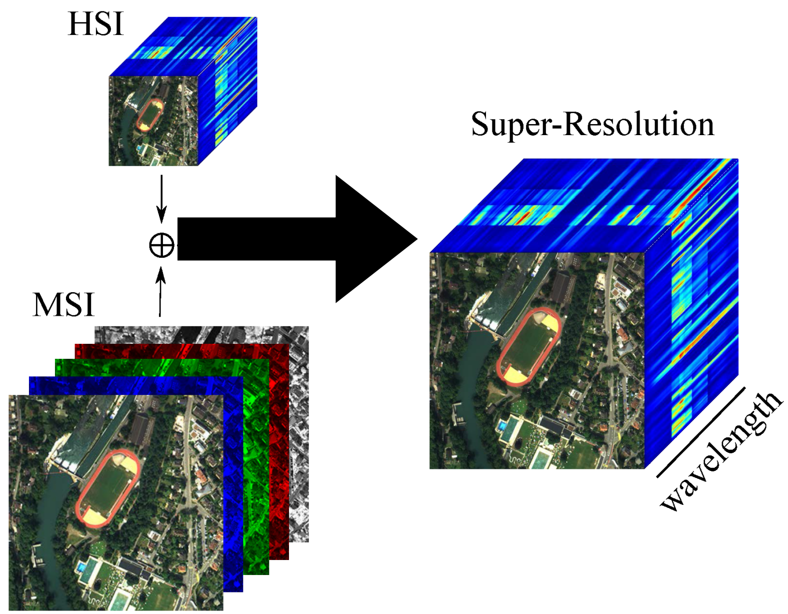
From model-based to data-driven



Case study: Hyperspectral Image (HSI) Super-Resolution (SR)

Hyperspectral images (HSI):
Low spatial but high spectral resolution

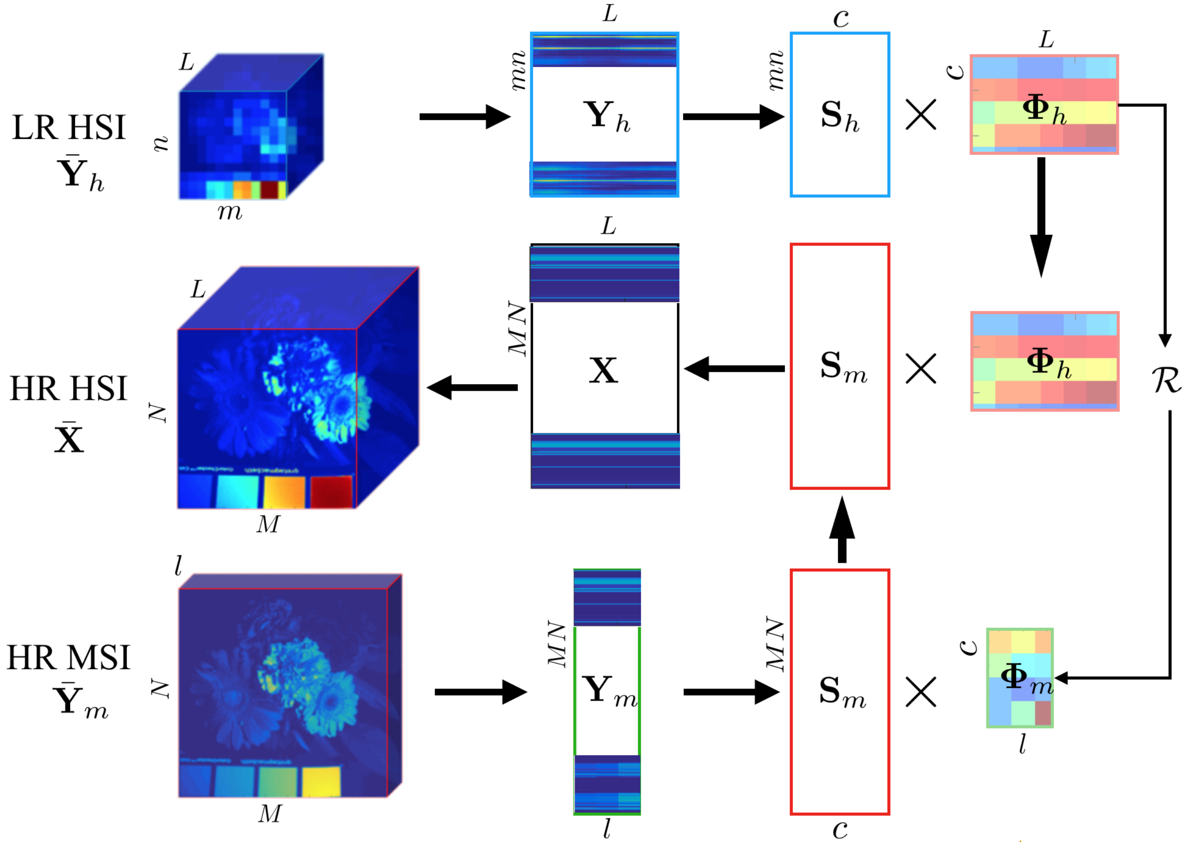
Multispectral images (MSI):
High spatial but low spectral resolution



HSI-SR:
High spatial and High spectral resolution

AICIP RESEARCH

The traditional formulation



Spatial formation

Spectral bases

$$\mathbf{Y}_h = \mathbf{S}_h \mathbf{\Phi}_h,$$

$$\mathbf{X} = \mathbf{S}_m \mathbf{\Phi}_h.$$

$$\mathbf{Y}_m = \mathbf{S}_m \mathbf{\Phi}_m,$$

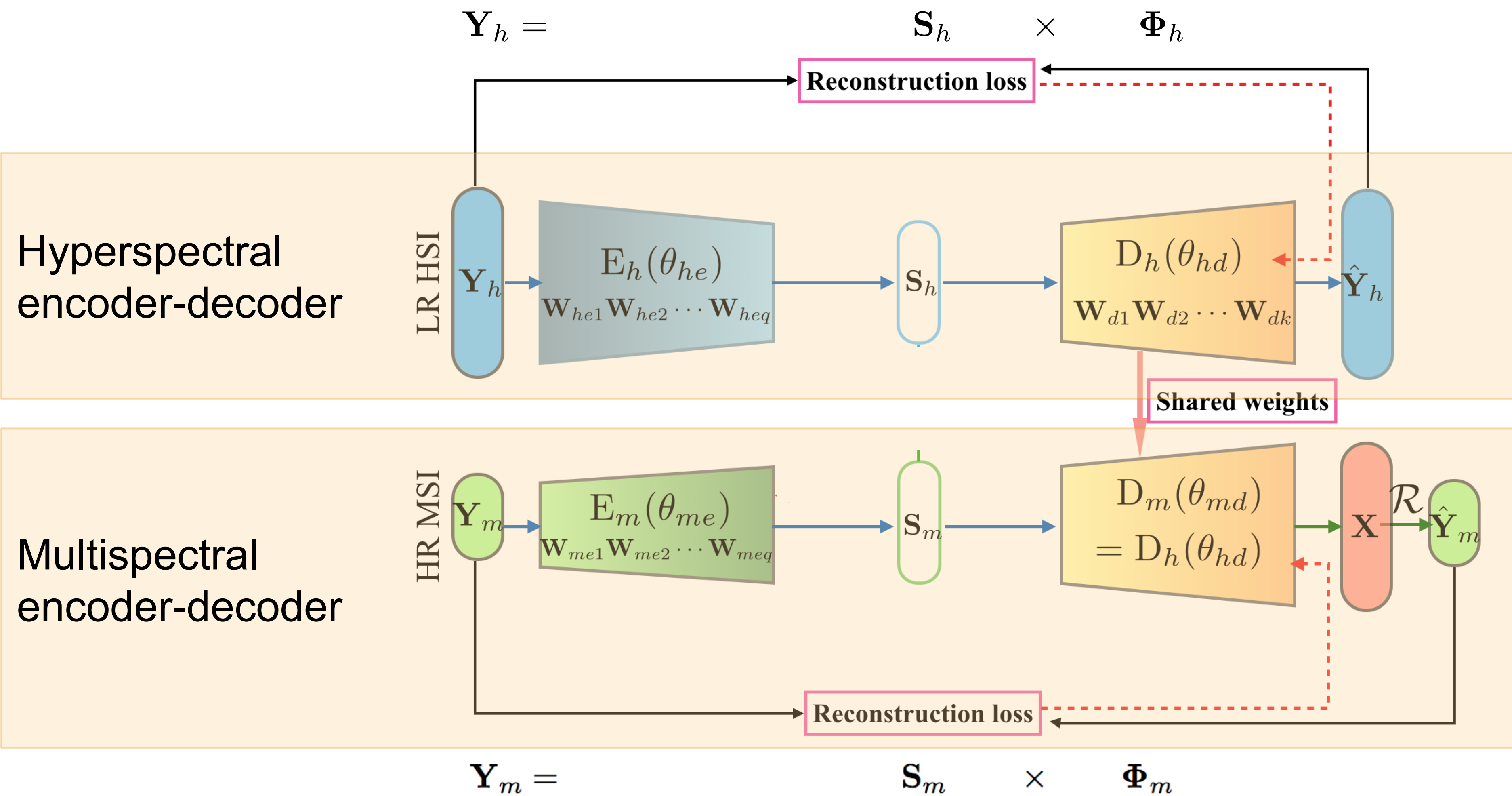
The objective function:

$$P(\mathbf{X}|\mathbf{Y}_h, \mathbf{Y}_m)$$

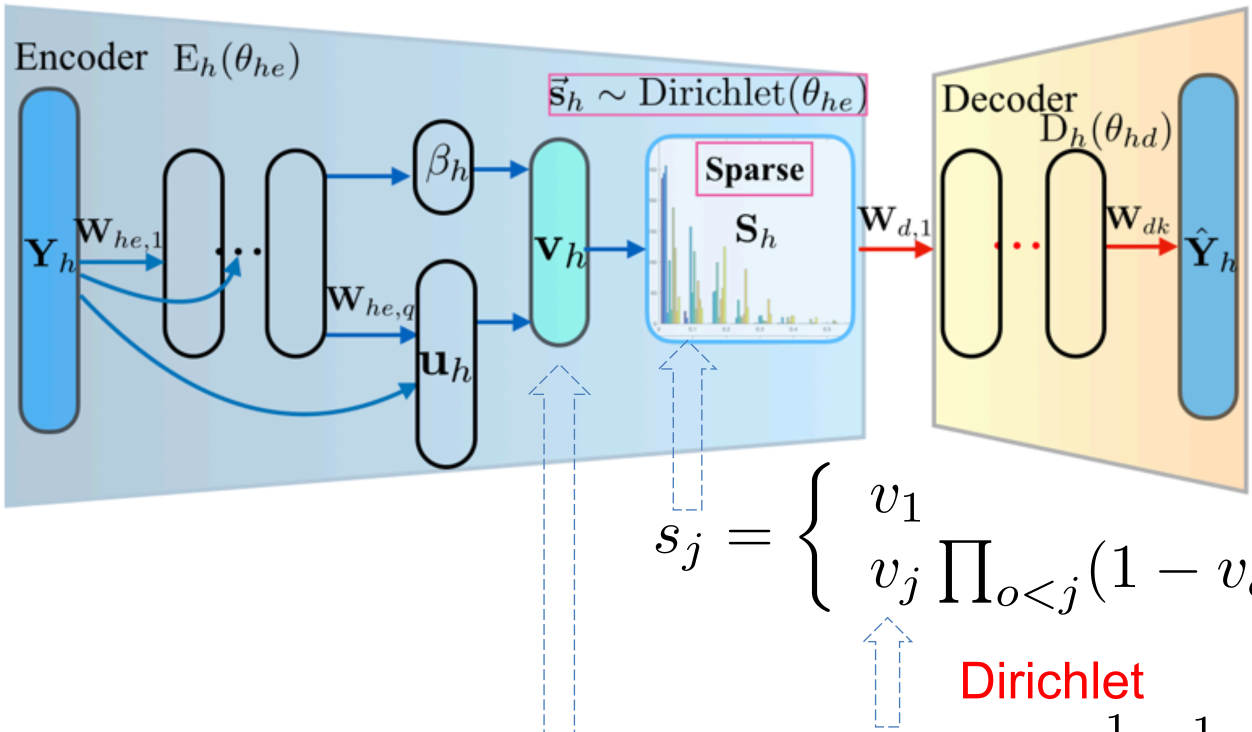
The constraints on S:

- Sum-to-one and non-negative
- Sparse

The deep-learning approach using unsupervised learning



The deep-learning approach with two physical constraints on S (Sum-to-one and Non-negativity)



$$s_j = \begin{cases} v_1 & \text{for} \quad j = 1 \\ v_j \prod_{o<j} (1 - v_o) & \text{for} \quad j > 1 \end{cases}$$

Dirichlet

Kumaraswamy $v_o \sim (1 - (1 - u^{\frac{1}{\beta}})^{\frac{1}{\alpha}}).$

$$v_o \sim \text{Beta}(u, 1, \beta)$$

$$\text{kuma}(u, \alpha, \beta) = \alpha\beta u^{\alpha-1}(1 - u^{\alpha})^{\beta-1}$$

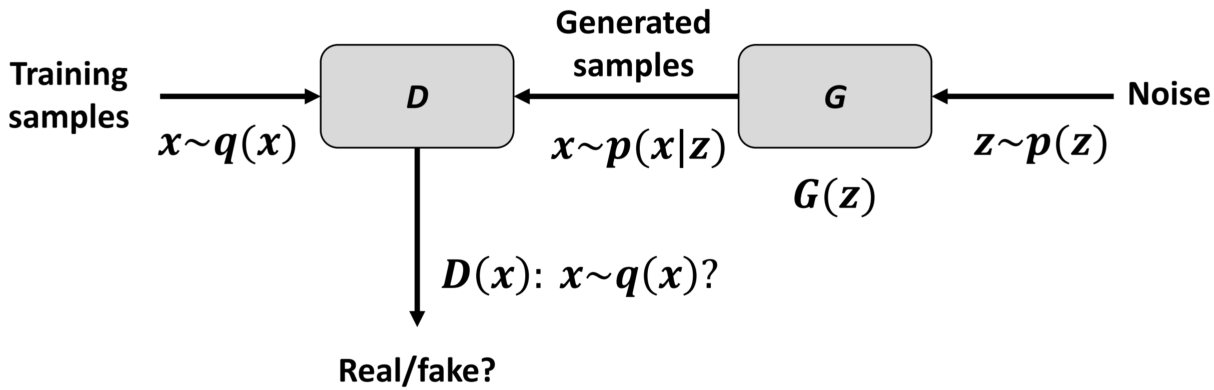
A list of misconceptions

AICIP RESEARCH

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation

GAN

- Two neural networks compete against each other
  - A generator network G: mimic training samples to fool the discriminator
  - A discriminator network D: discriminate training samples and generated samples



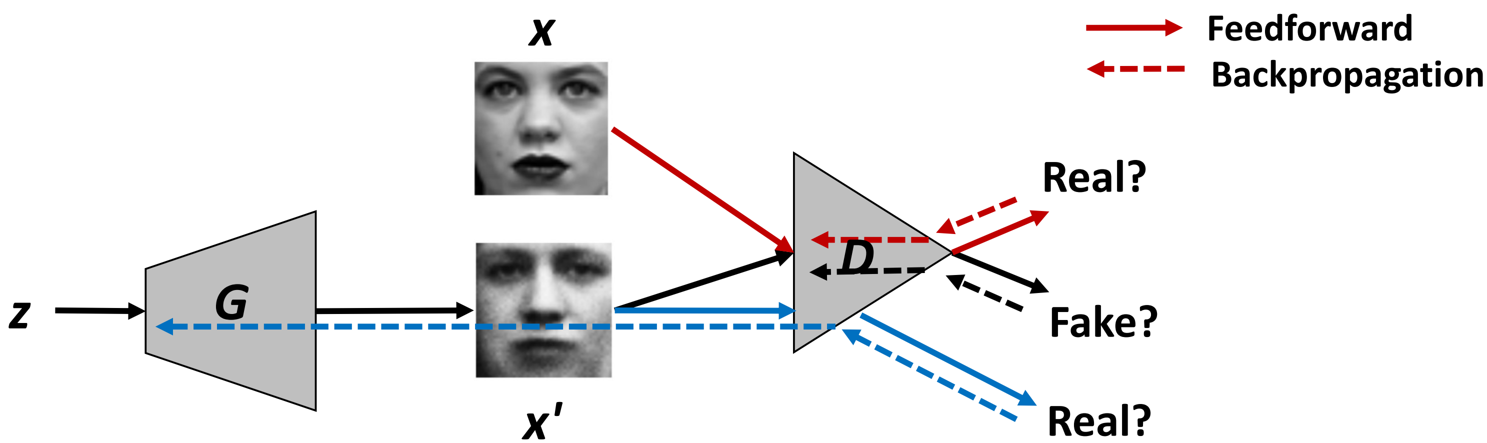
For *D*: $\max_D \; \mathbb{E}_{x\sim q(x)}[\log(D(x))] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]$

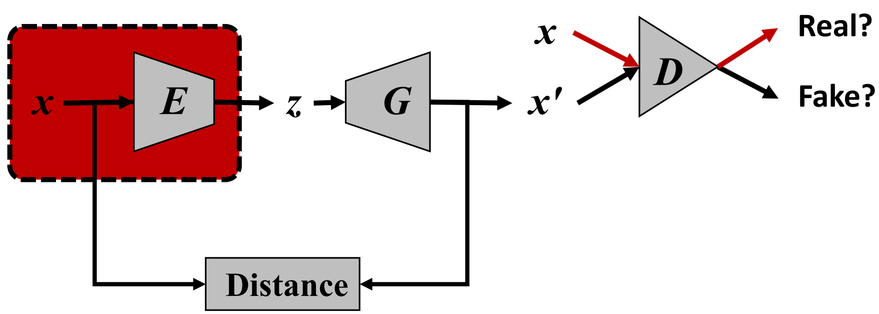
For *G*: $\min_G \; \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]$

GAN

The objective function of GANs:

$$\min_G \max_D \; \mathbb{E}_{x\sim q(x)}[\log(D(x))] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]$$

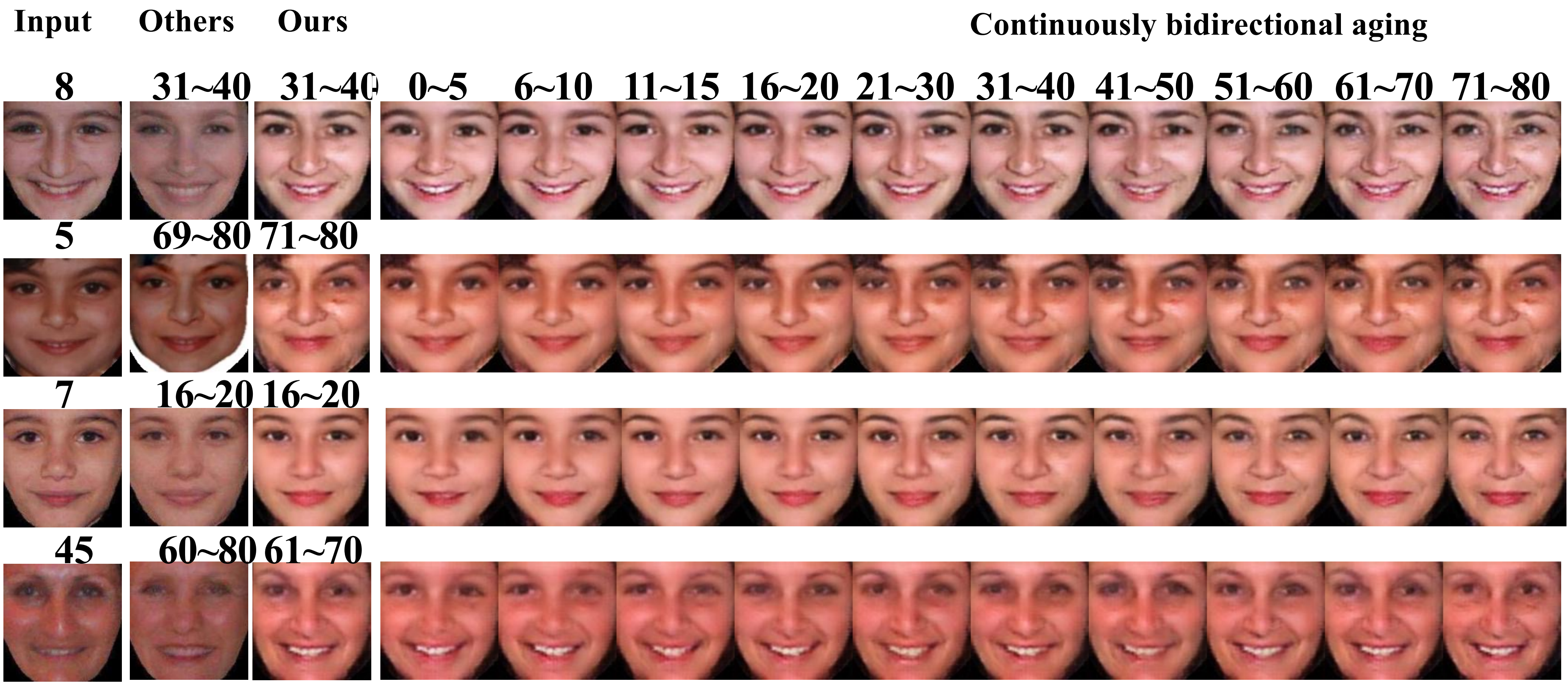




Adding an AE

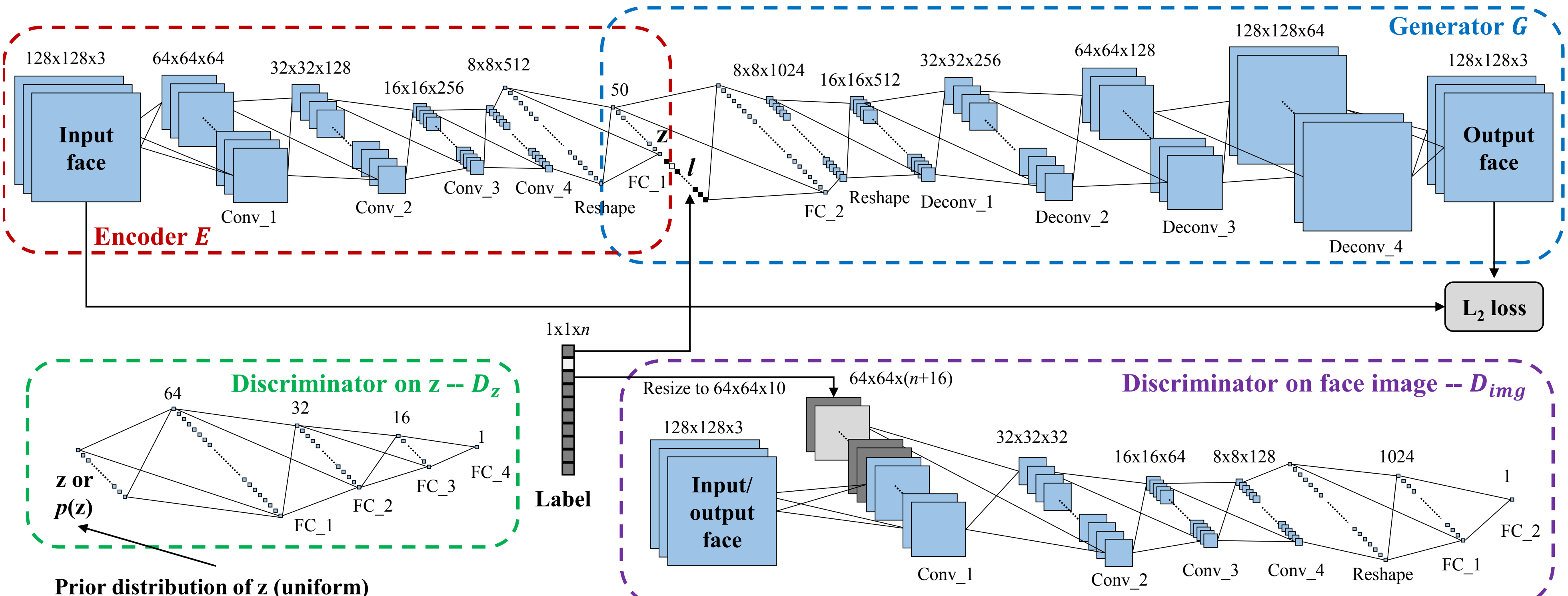
AICIP RESEARCH

Case study: Age progression and regression

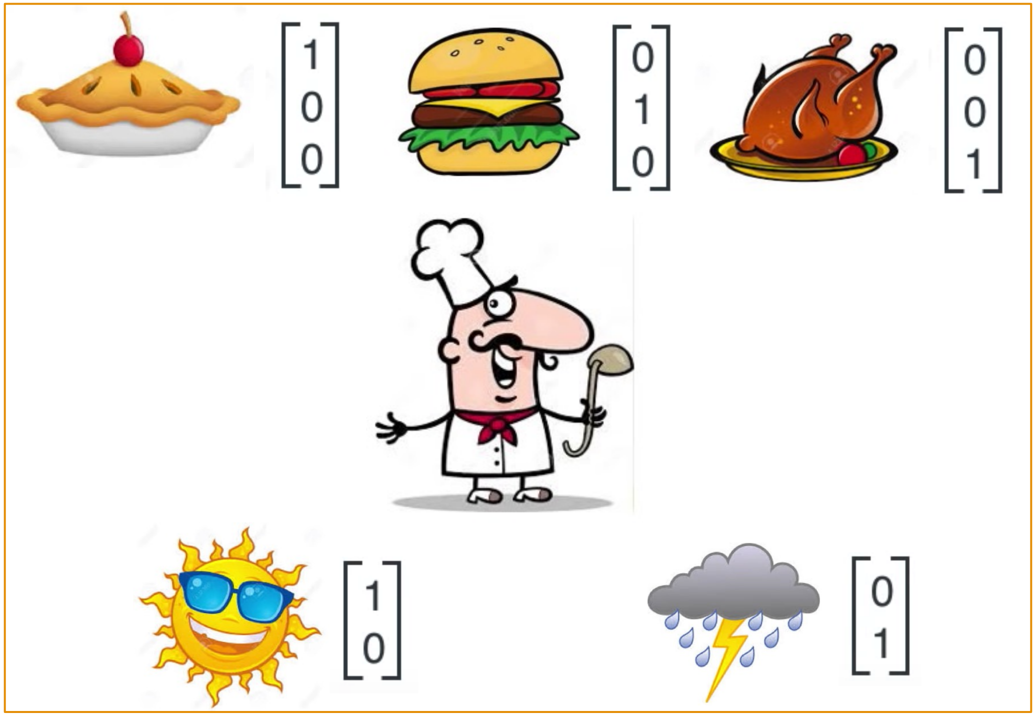


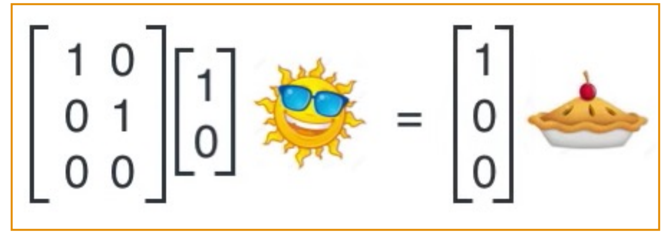
Project page: https://zzutk.github.io/Face-Aging-CAAE

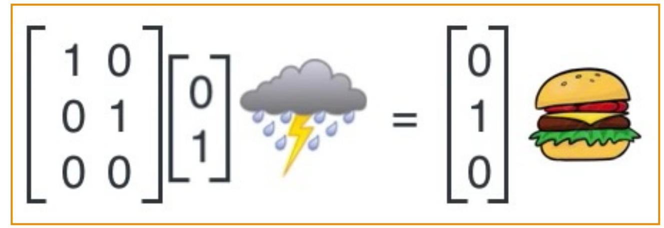
Case study: Conditional Adversarial Autoencoder - CAAE

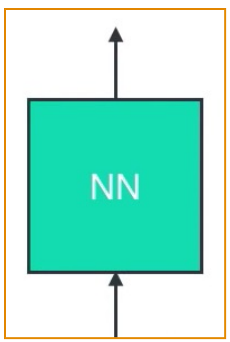


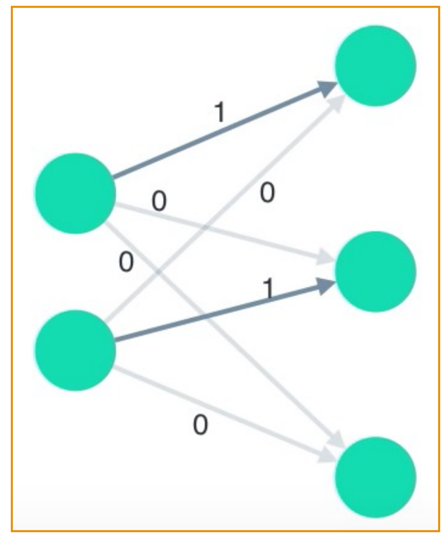
RNN: A friendly introduction to NN



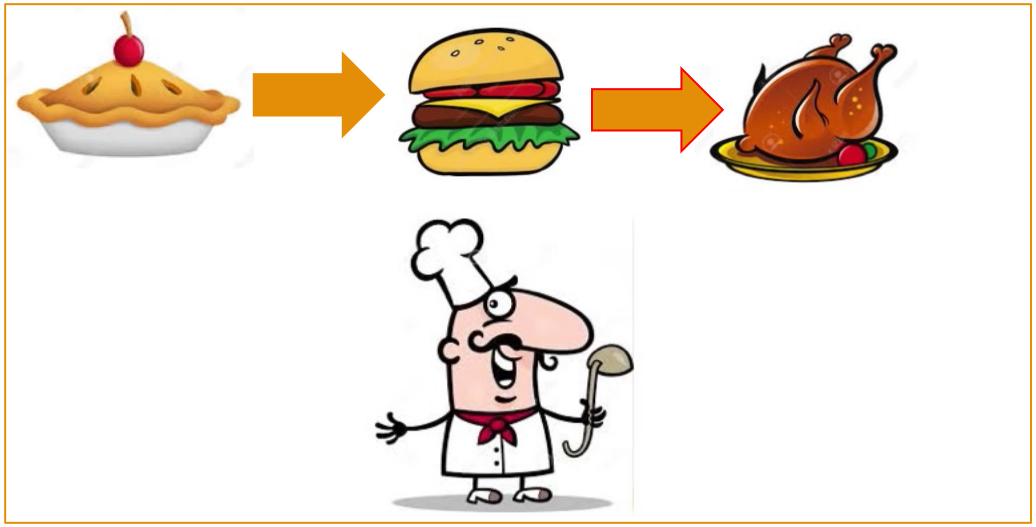




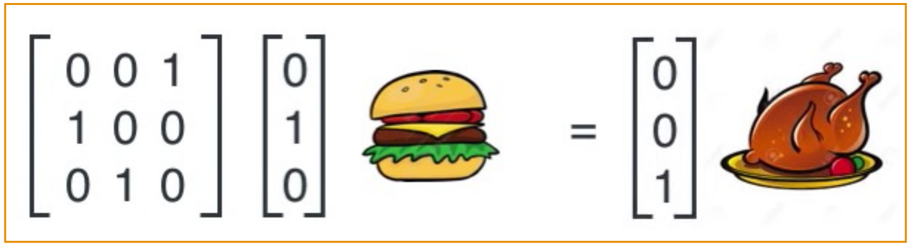




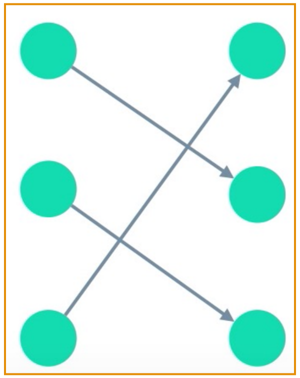
A friendly introduction to RNN

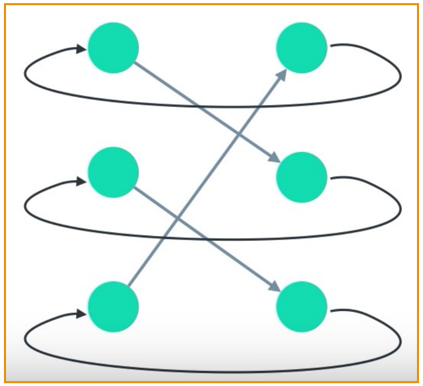


$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

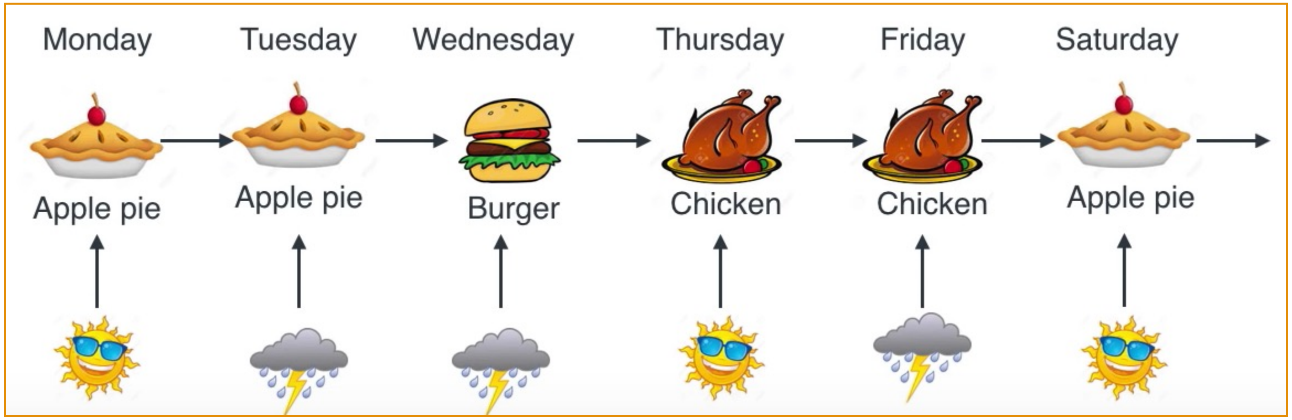


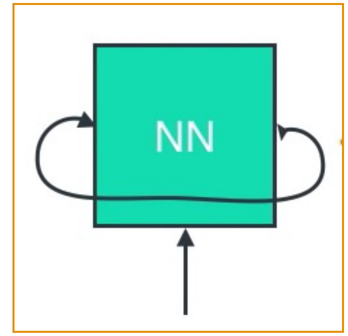
$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

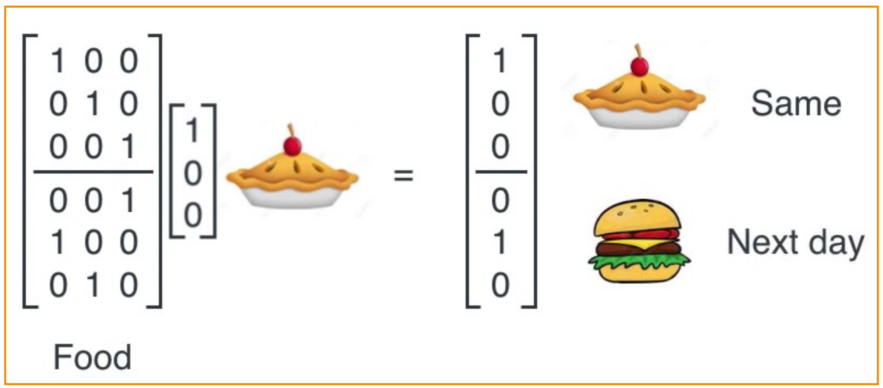


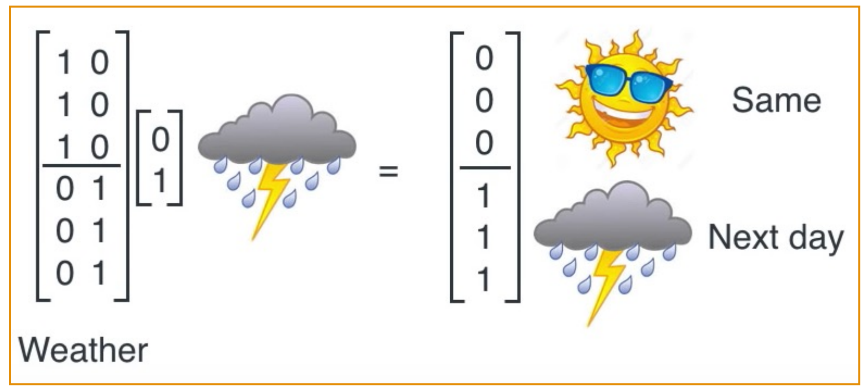


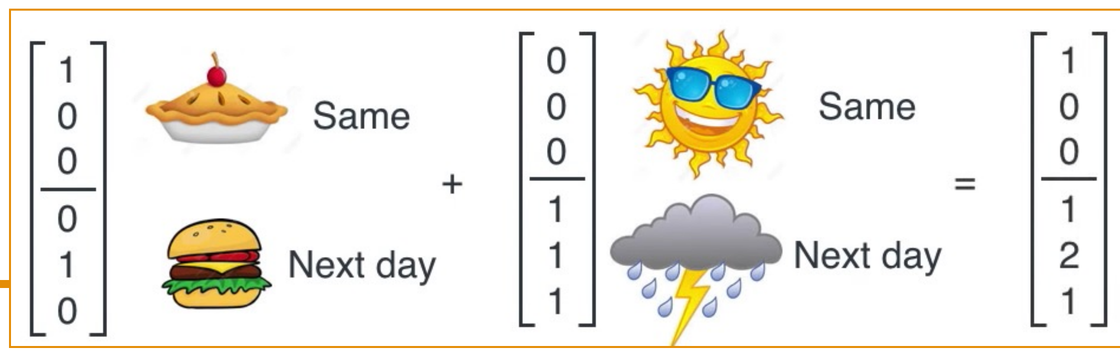
A more complicated case



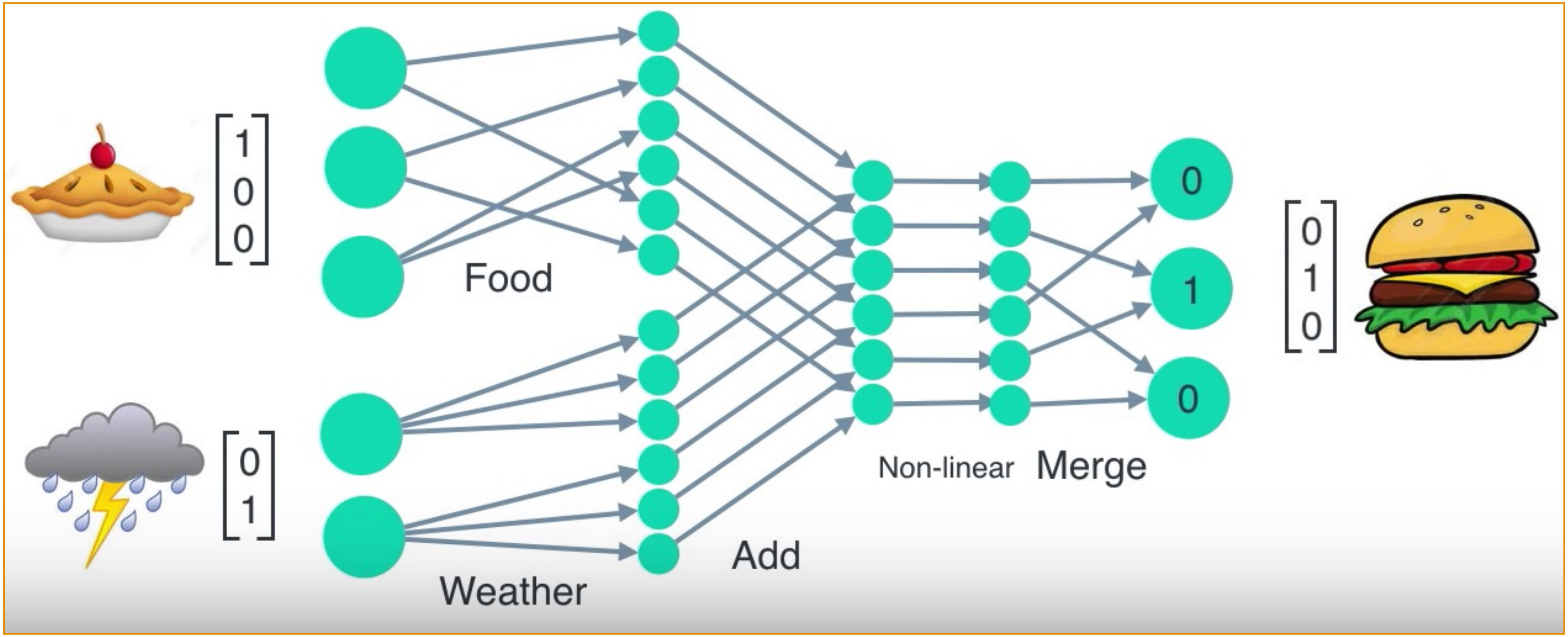




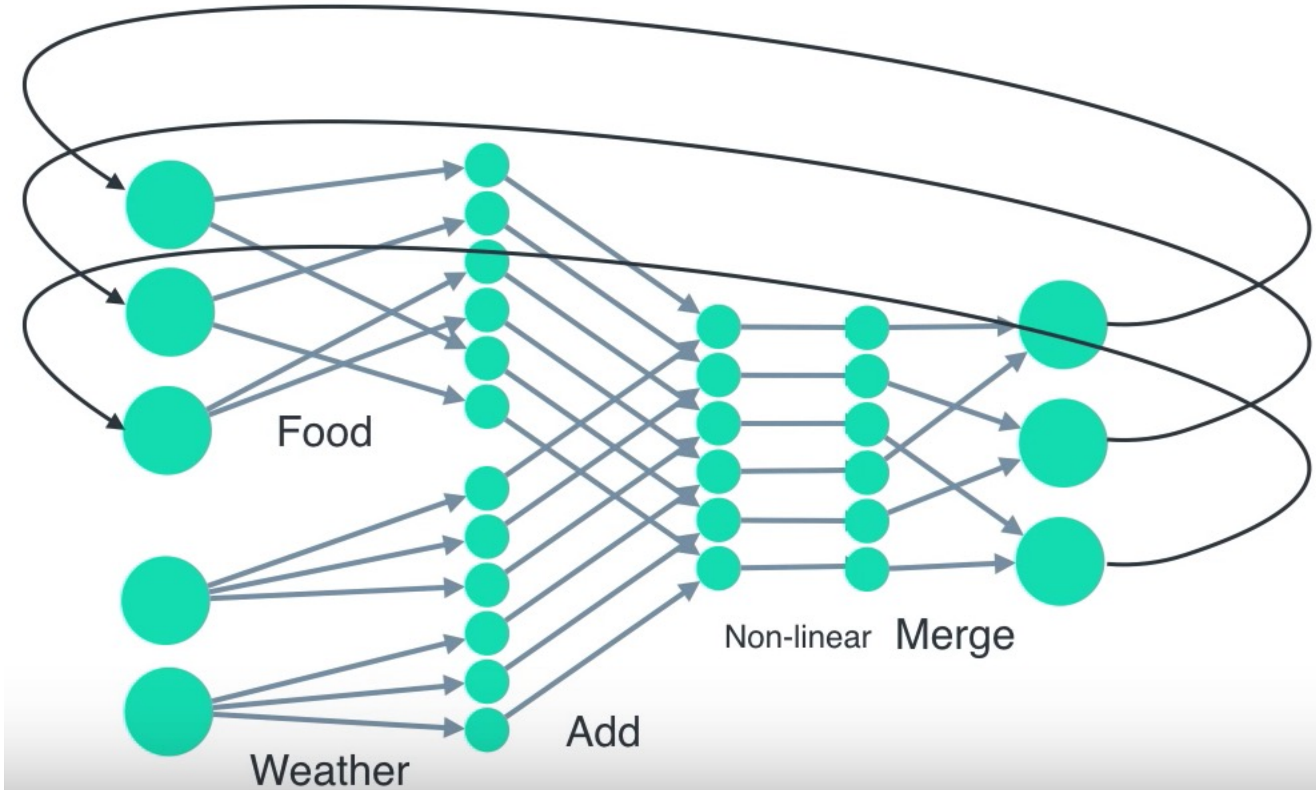




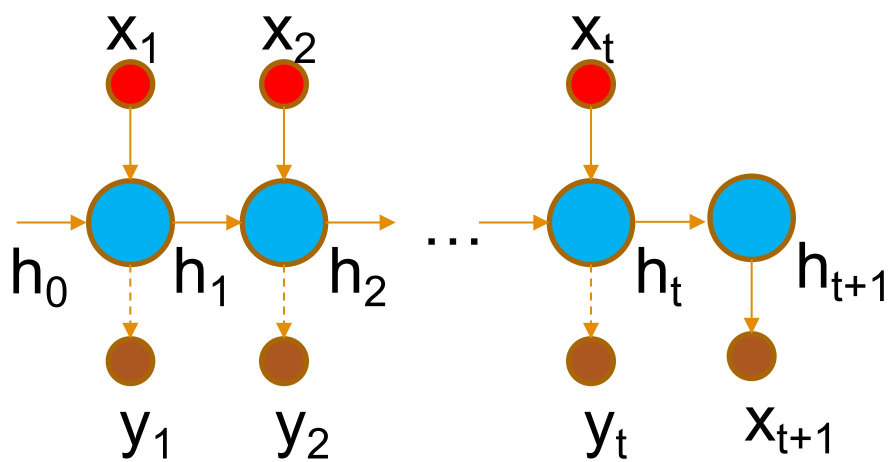
A more complicated case (cont'd)



A more complicated case (cont'd)

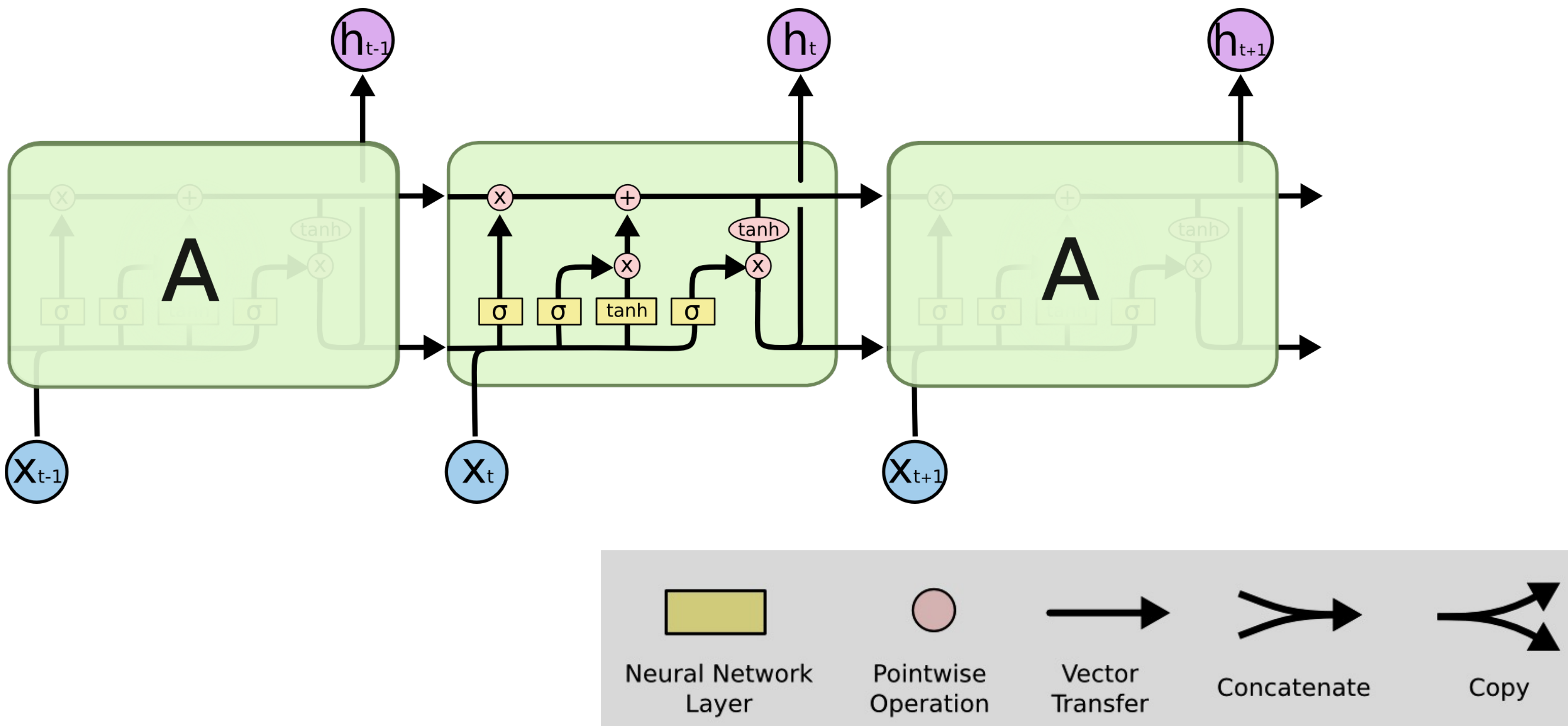


Recurrent neural network (RNN)



The long-short term memory (LSTM) module

LSTMs are explicitly designed to avoid the long-term dependency problem.



Case study: The talking face

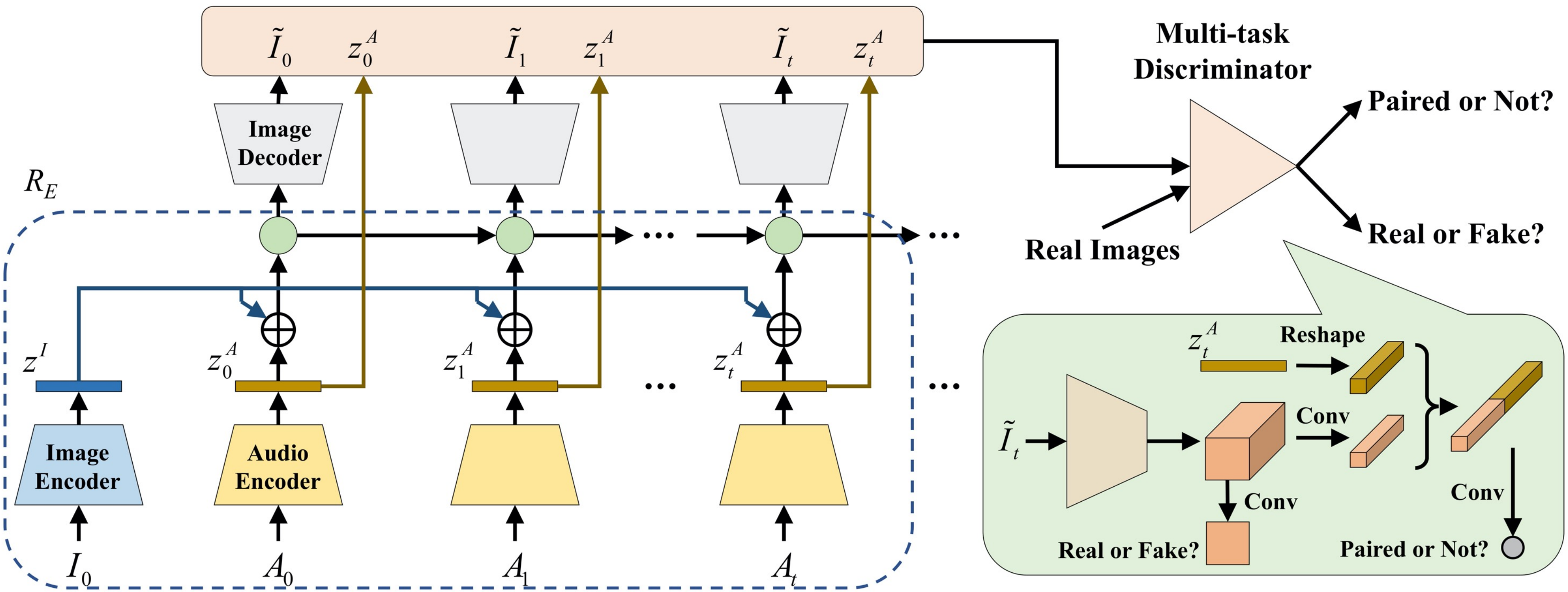
Goal: Given an arbitrary audio clip and a face image, automatically generate realistic and smooth face video with accurate lip sync.

[Suwajanakorn et al., 2017]

Application: Face animation, entertainment, video bandwidth reduction, etc.

The talking face

The proposed method: conditional video generation



http://web.eecs.utk.edu/~ysong18/projects/talkingface/talkingface.html

A list of misconceptions

- Is deep learning merely deeper?
  - The two unique features of convolutional neural network (CNN)
- Is deep learning a classifier?
  - Engineered features vs. automatic features
- Supervised vs. Unsupervised
- Model-based approach vs. Data-driven approach – the two extremes?
- The world beyond CNN
  - GAN, AE, RNN, RL
- Implementation
  - Matlab
  - TensorFlow
  - PyTorch
  - Keras

Education is what remains after one has forgotten everything one learned in school. -- Albert Einstein