RNN

Name	Nathan Varghese
Identity Key	nava3000

	Level	Completed
	Beginner	8
	Intermediate	2
	Advanced	0
	Expert	0

Goal	
5722	10

Total Completed
10

Recurrent Neural Network

CSCI 5722: Computer Vision

Fall 2024

Dr. Tom Yeh

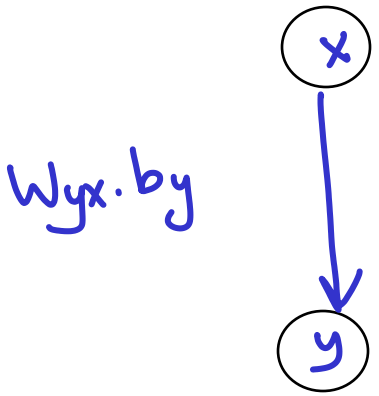
Evolution to RNNs

CSCI 5722 Computer Vision

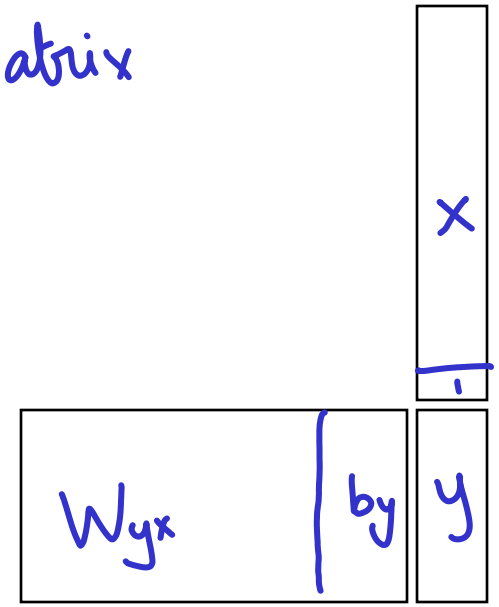
University of Colorado Boulder

Linear transformation

graphical



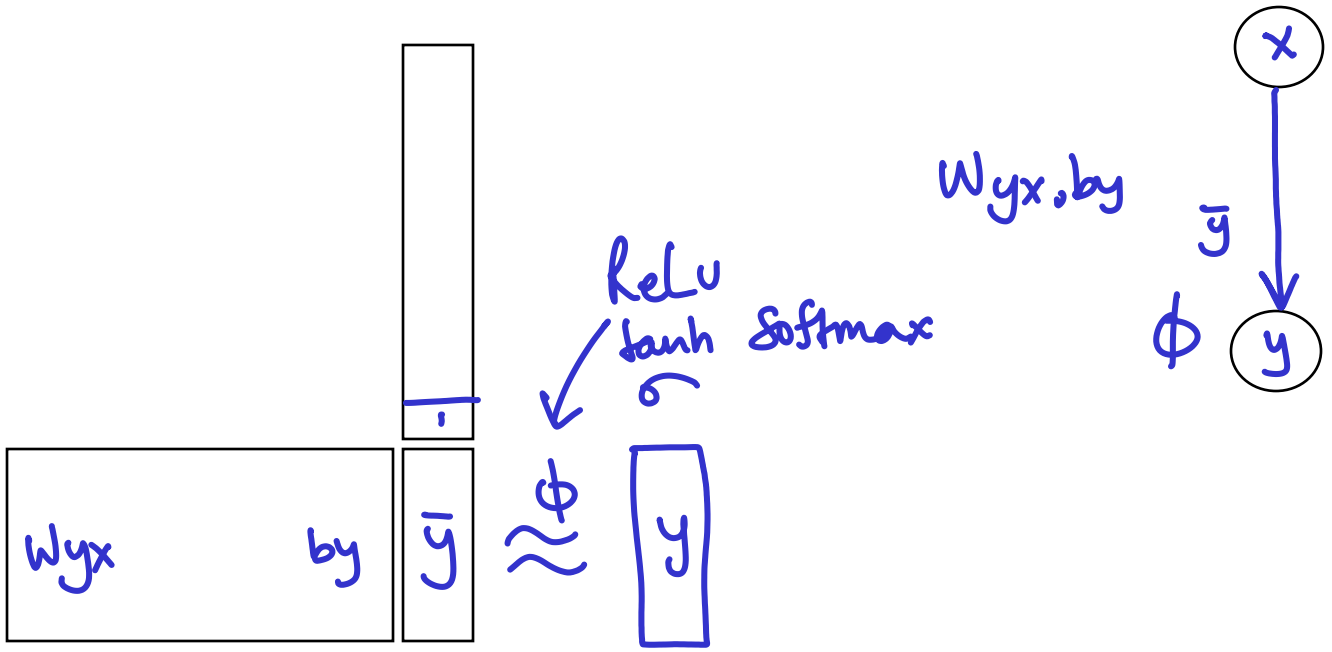
matrix



algebraic

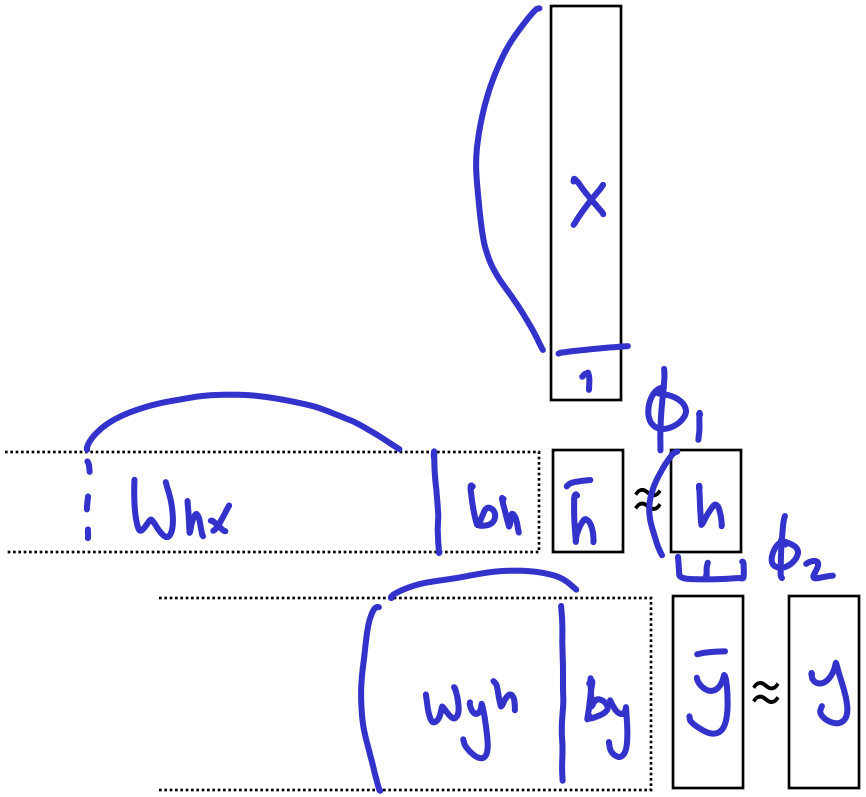
$$y = W_{yx} * X + b_y$$

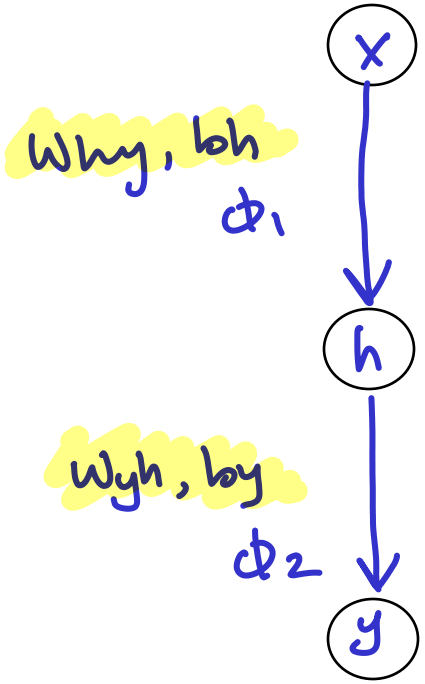
+ non-linearity



$$y = \phi(W_{yx} \cdot X + b_y)$$

+ hidden layer





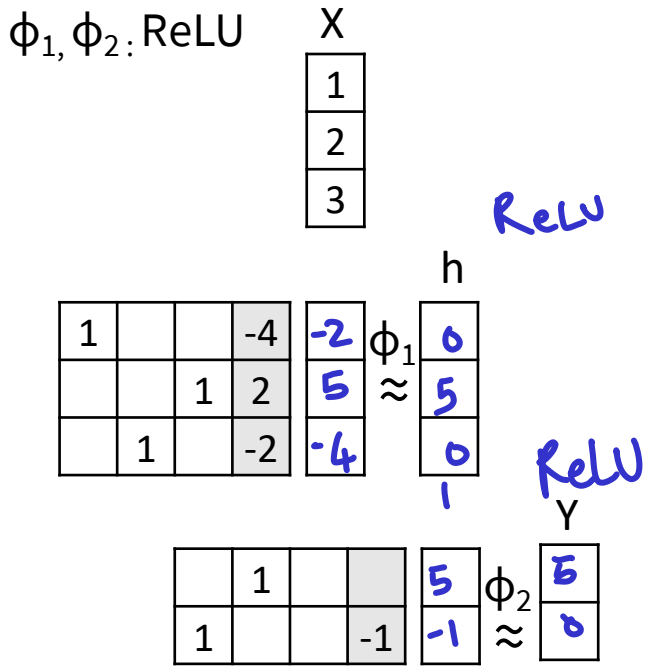
- trainable parameters

$h = \phi_1(W_{hx} \cdot X + b_h)$

$y = \phi_2(W_{yh} \cdot h + b_y)$

$= \phi_2(W_{yh} \cdot (\phi_1(W_{hx} \cdot X + b_h)) + b_y$

Calculate an NN with a hidden layer



$$Y = \phi_2(\mathrm{W}_2 \cdot \phi_1(W_1 \cdot X + b_1) + b_2)$$

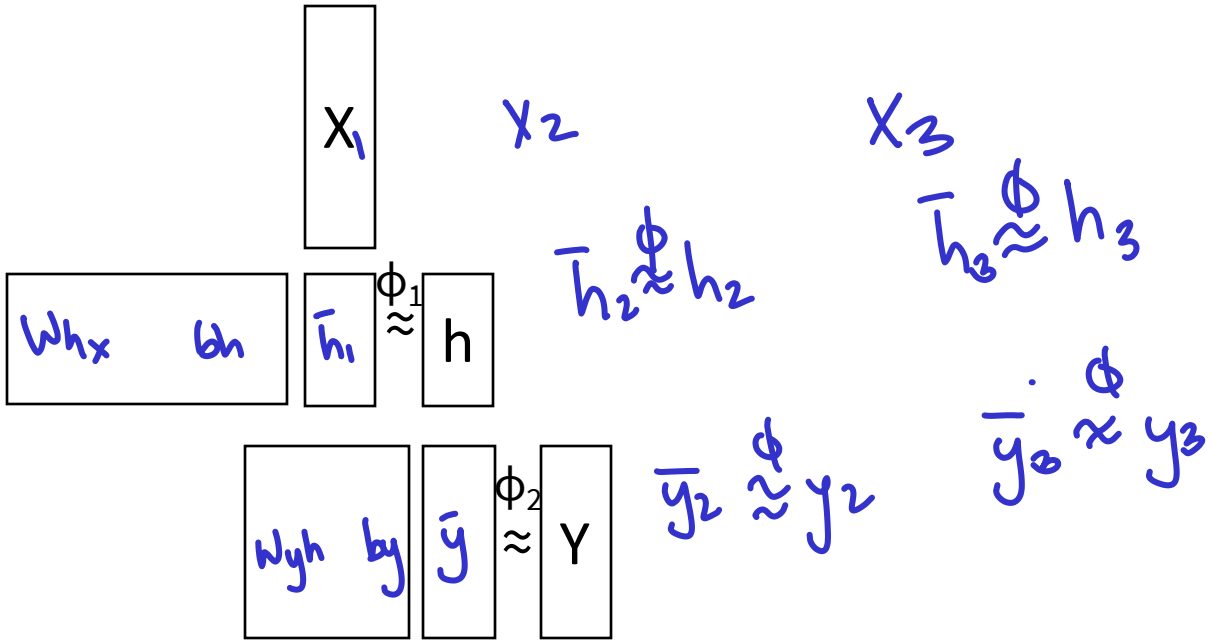
size(W_1) = 3 x 3

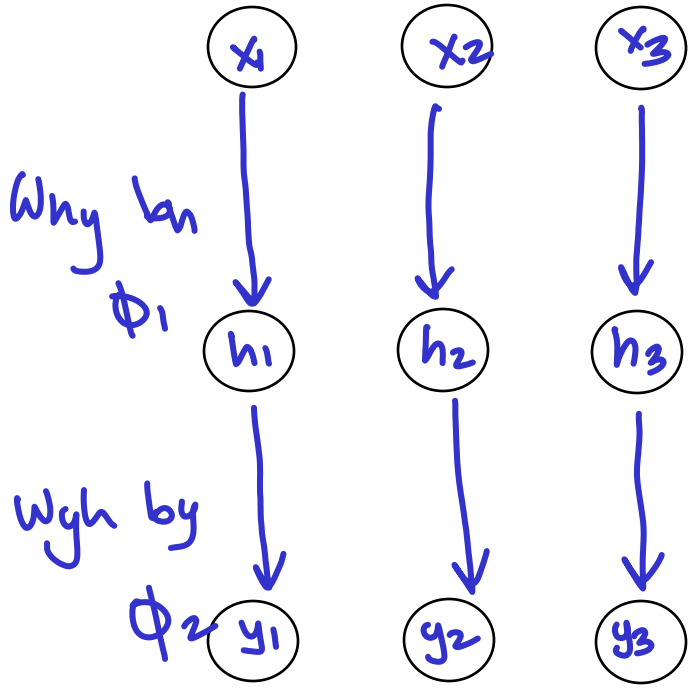
size(b_1) = 3 x 1

size(W_2) = 2 x 3

size(b_2) = 2 x 1

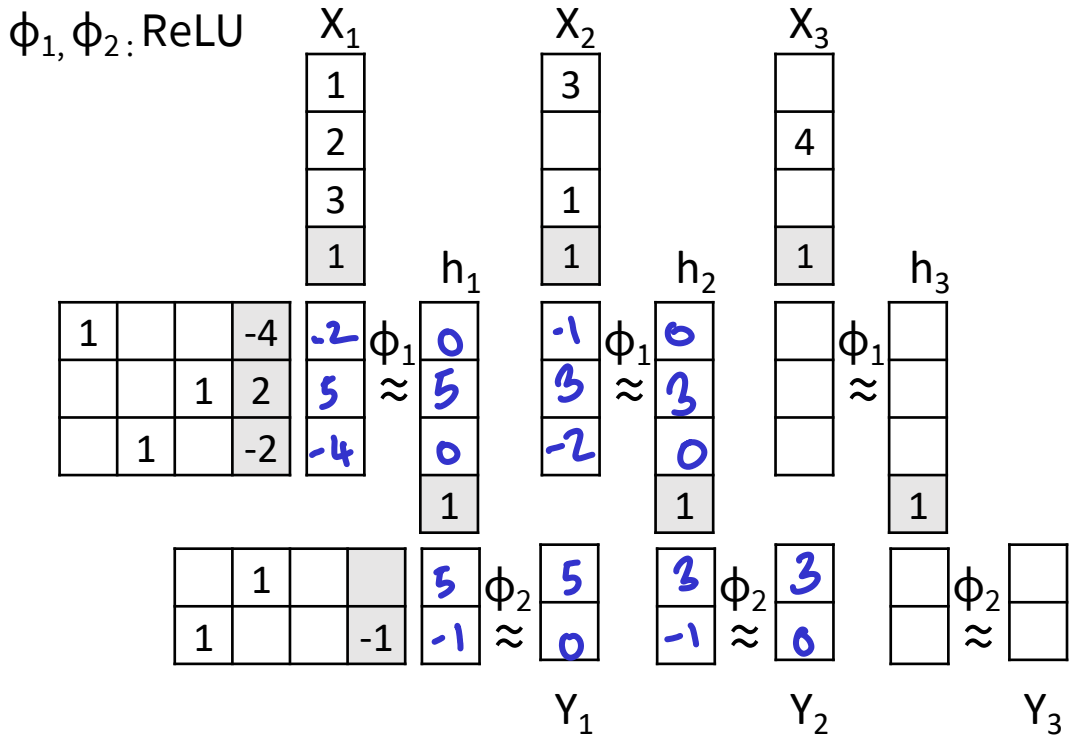
+ sequence





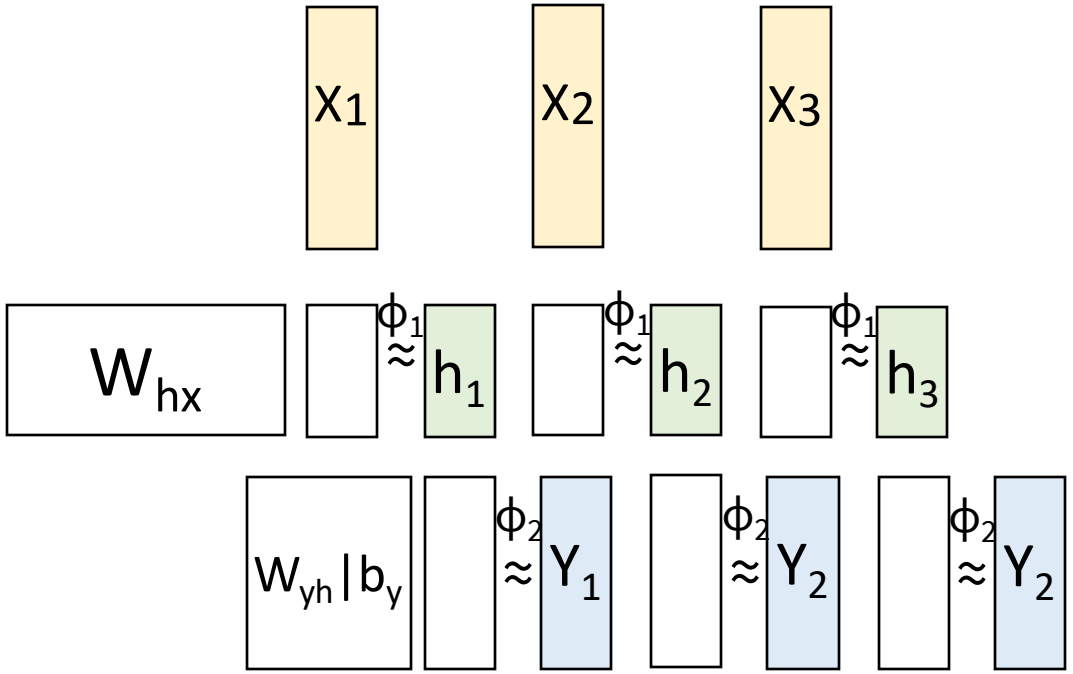
$$y_t = \Phi_2\left(W_{yh}\,\Phi_1(W_{hx}X_t + b_h) + b_y\right) \qquad t = 1,2,3\ldots$$

Calculate an NN for a batch of inputs

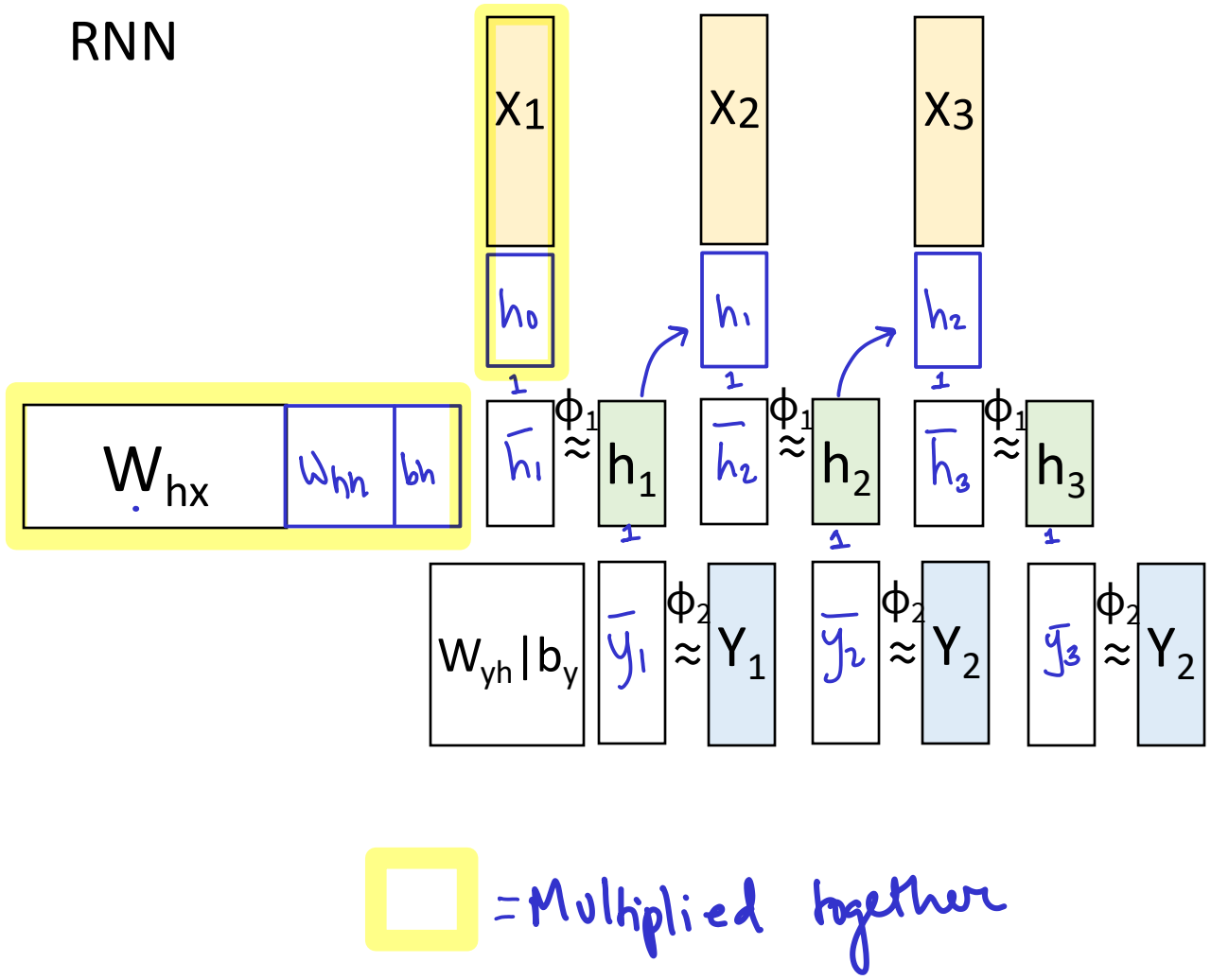


+ dependence on previous inputs

Independent NN



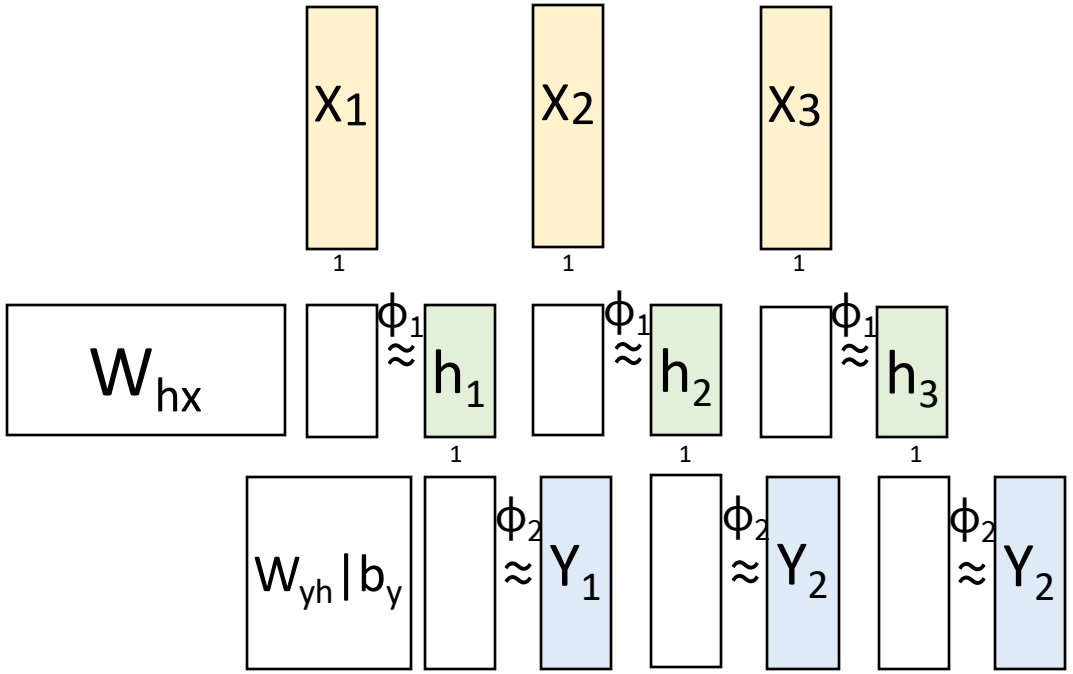
RNN



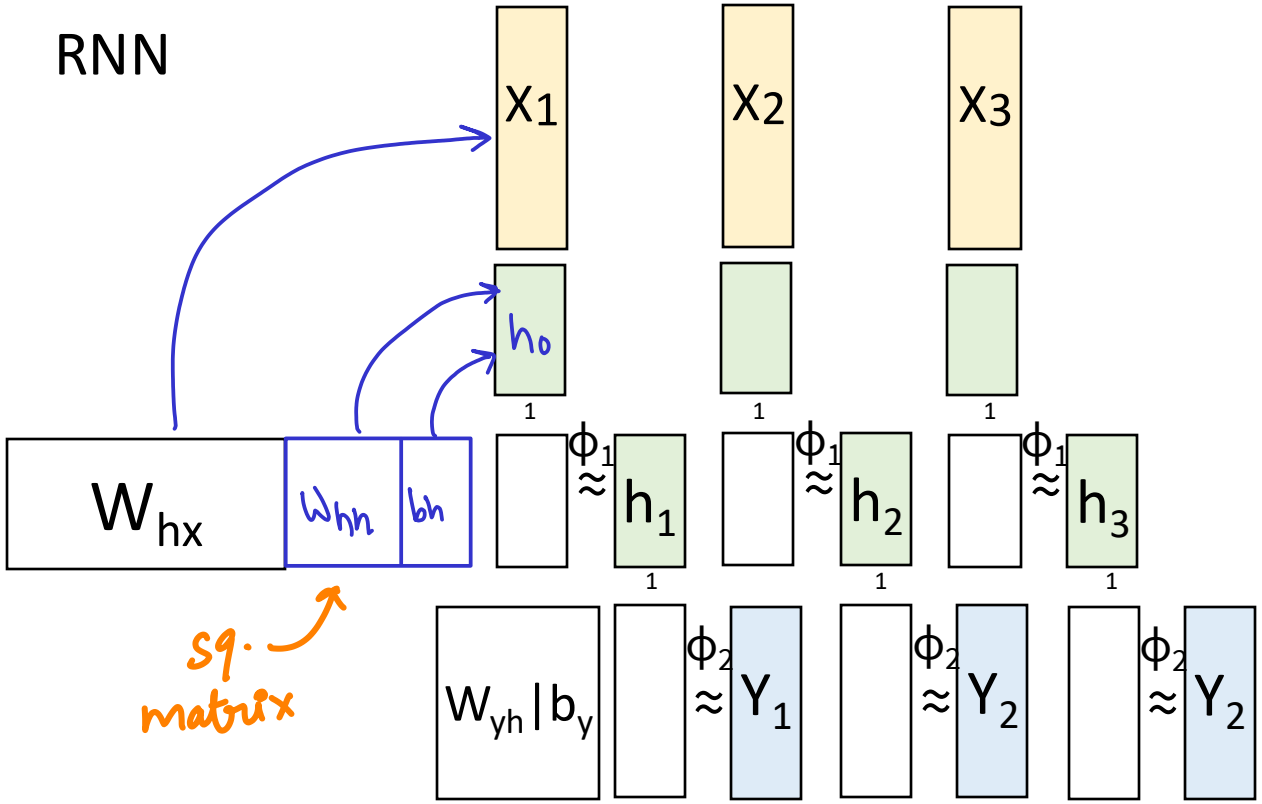
NN vs RNN: Math

- New term

Independent NN



RNN



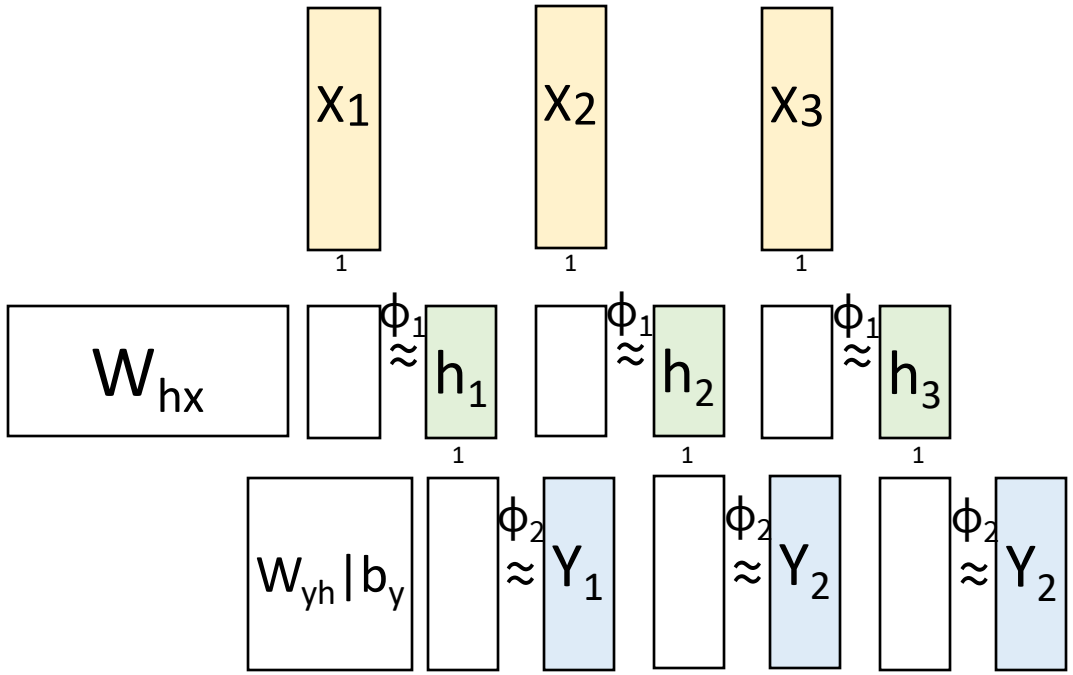
$$h_t = \phi_1(W_{hx} \cdot X_t + bh)$$

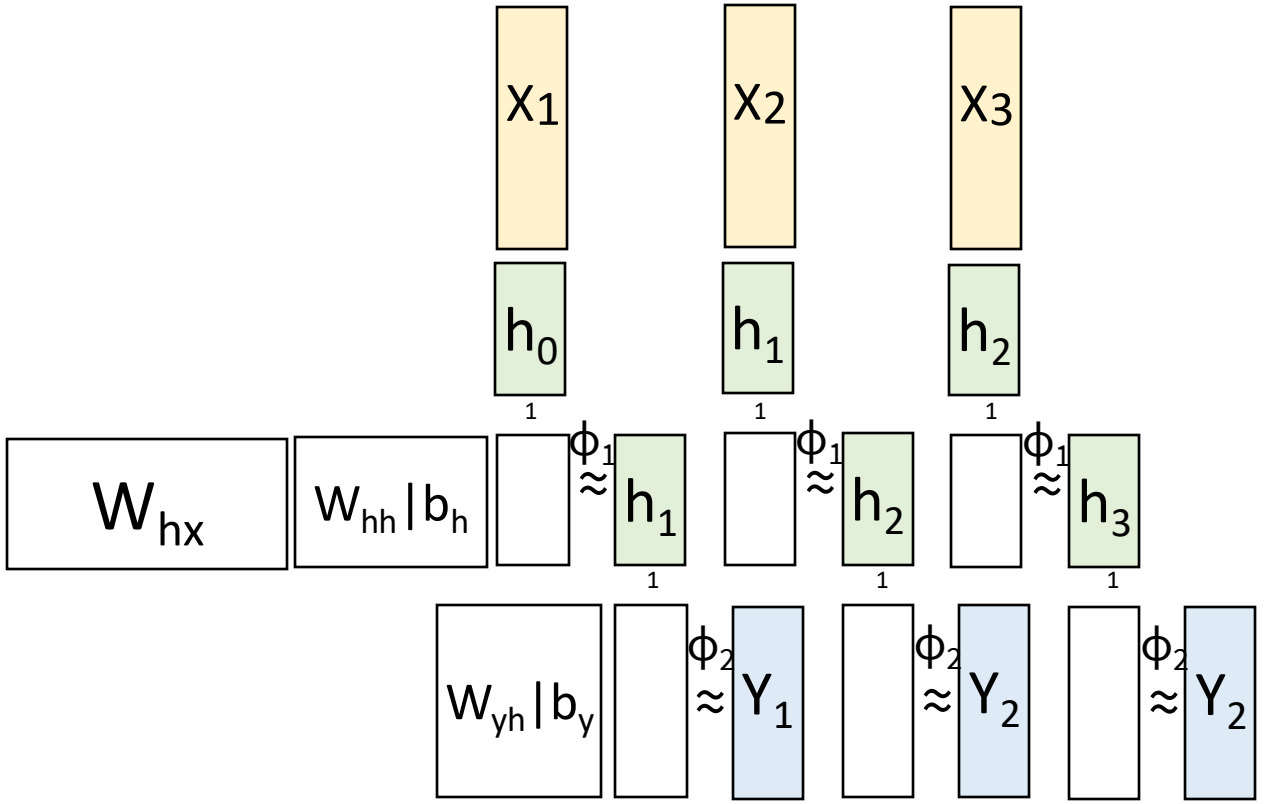
$$Y_t = \phi_2(W_{yh} \cdot h_t + by)$$

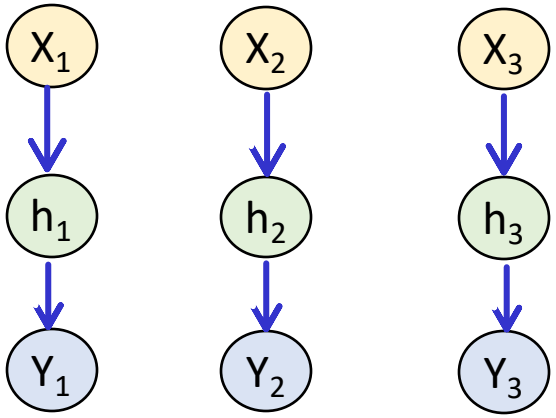
$h_t =$ $\phi_1 (W_{hx} \cdot X_t + W_{hh} \cdot h_{t-1} + bh)$

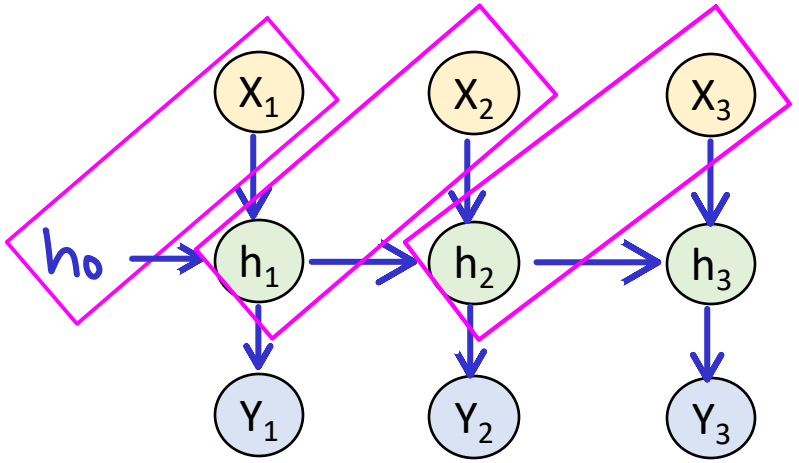
$Y_t =$ $\phi_2 (W_{yh} \cdot h_t + by)$

NN vs. RNN: Graph

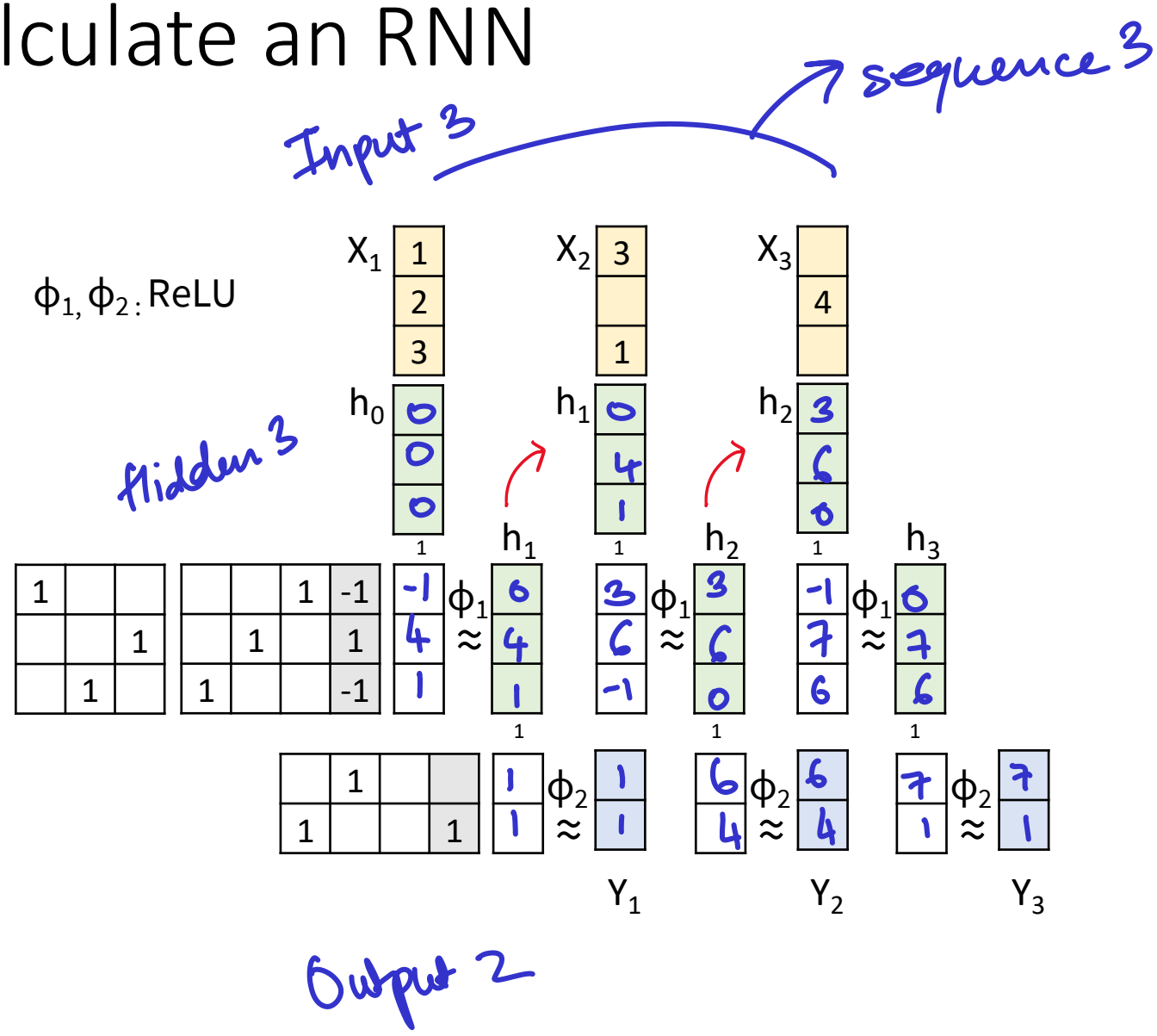




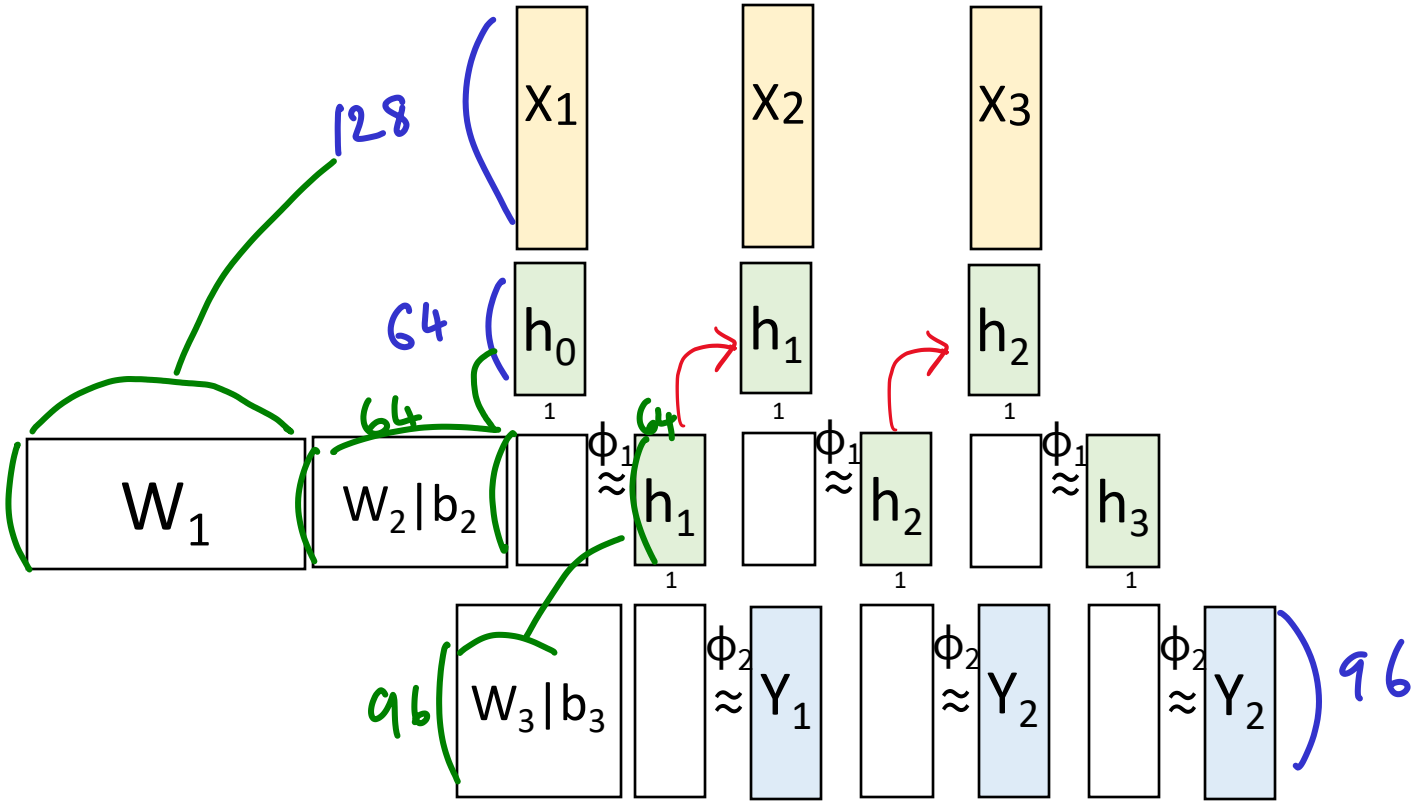




Calculate an RNN



Counting Parameters



Given :
size(X) = 128
size(h) = 64
size(Y) = 96

Determine:

size(W_1) = 64 x 128

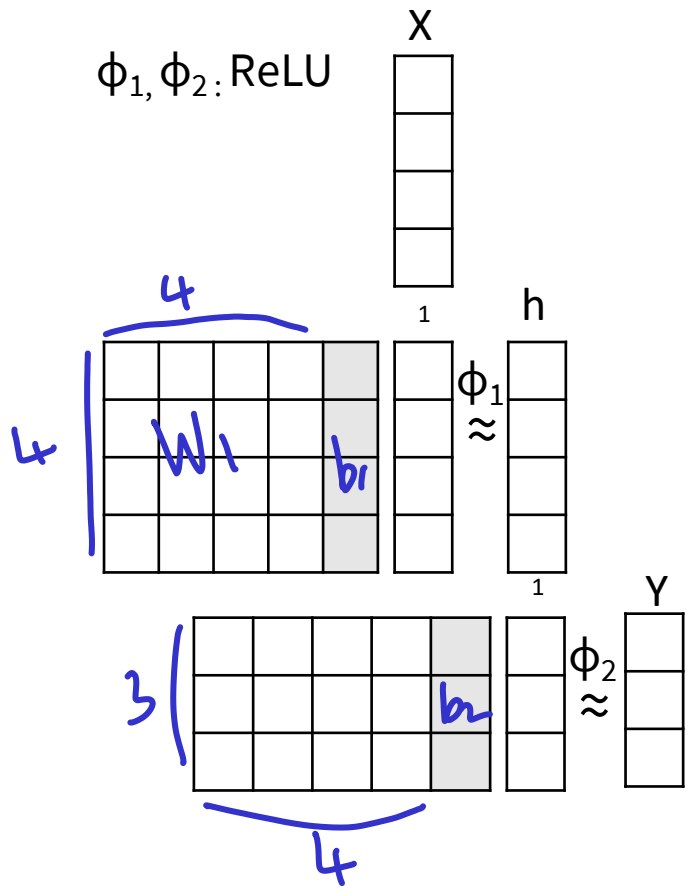
size(W_2) = 64 x 64

size(b_2) = 64 x 1

size(W_3) = 96 x 64

size(b_3) = 96 x 1

MLP Parameter Sizes



size(W_1) = 4 x 4

size(b_1) = 4 x 1

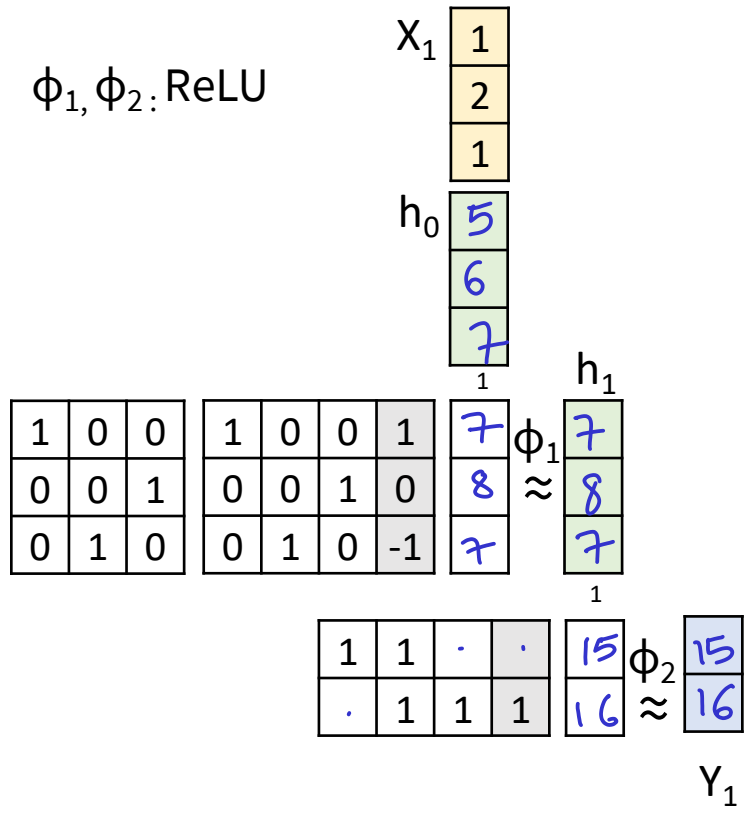
size(W_2) = 3 x 4

size(b_2) = 3 x 1

$$Y = \phi_2(\mathrm{W}_2 \cdot \phi_1(W_1 \cdot X + b_1) + b_2)$$

Calculate an RNN (t = 1)

This activity is standalone, not dependent on other activities.



$1+5+1=7$

$1+7=8$

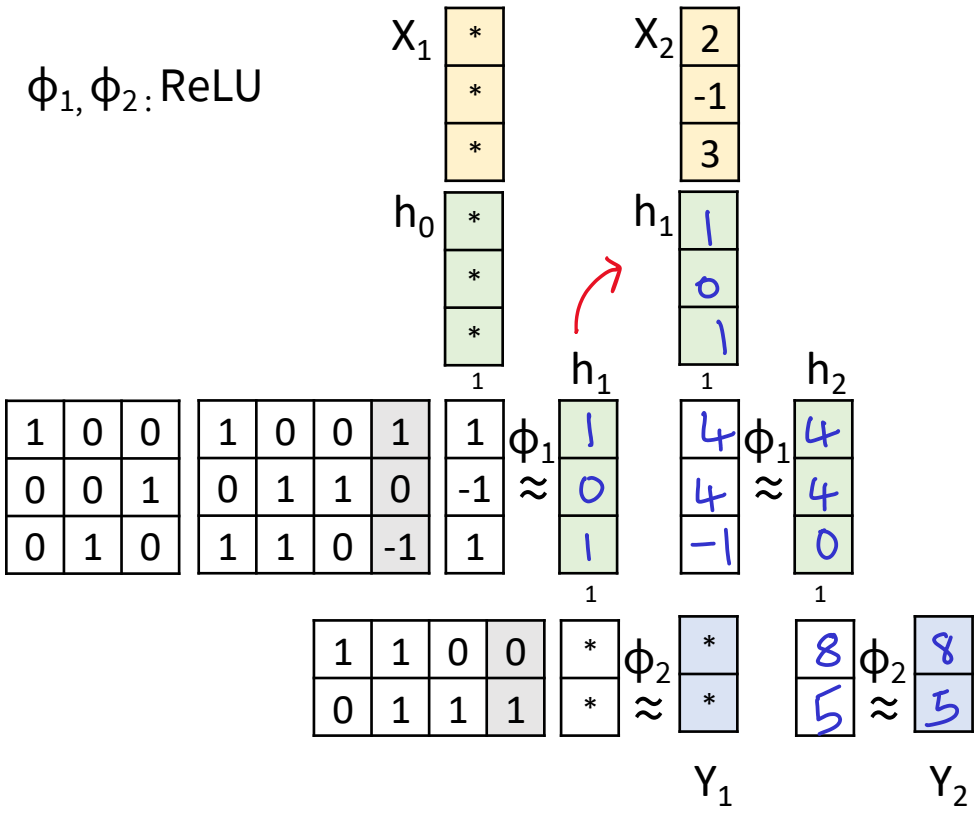
$2+6-1=7$

$7+8=15$

$8+7+1=16$

Calculate an RNN (t = 2)

This activity is standalone, not dependent on other activities.



$2+1+1=4$

$3+1=4$

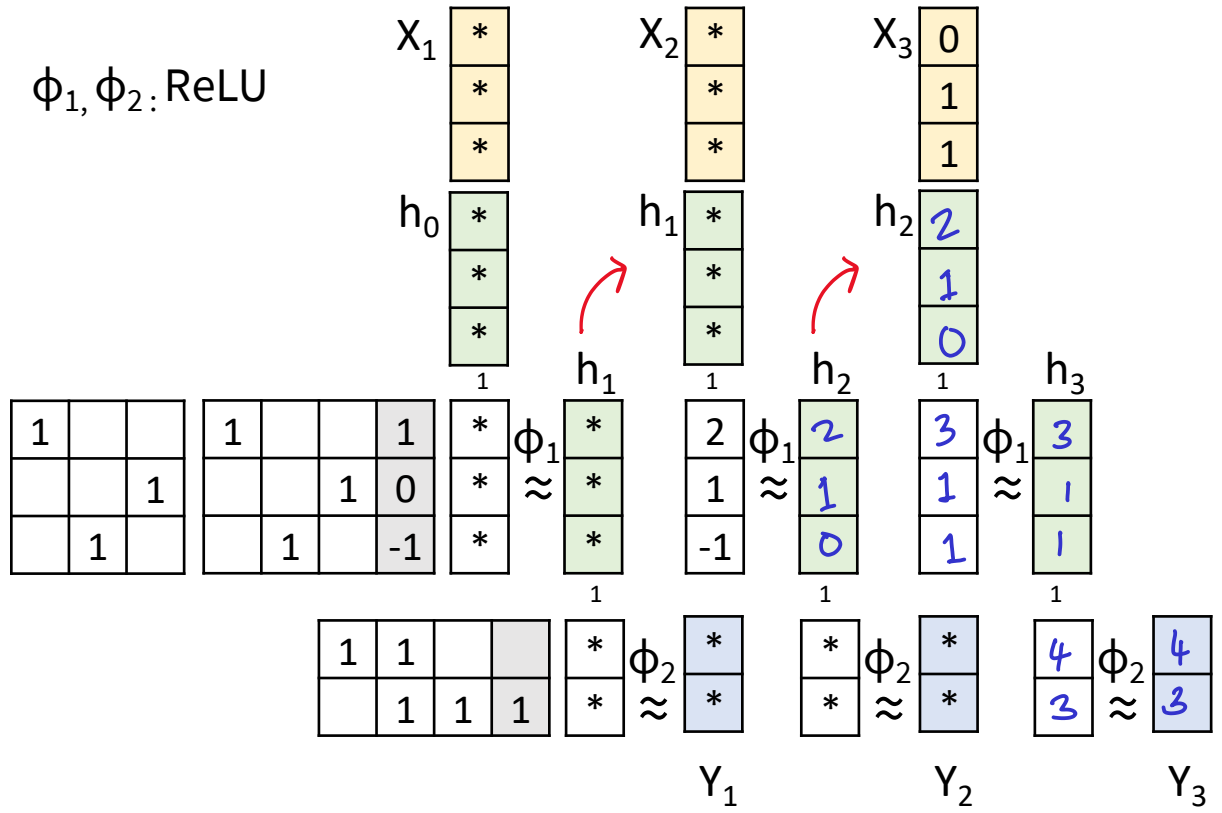
$-1+1-1=-1$

$4+4=8$

$4+1=5$

Calculate an RNN (t = 3)

This activity is standalone, not dependent on the previous one.



2 + 1 = 3

1 + 0 = 1

1 + 1 − 1 = 1

3 + 1 = 4

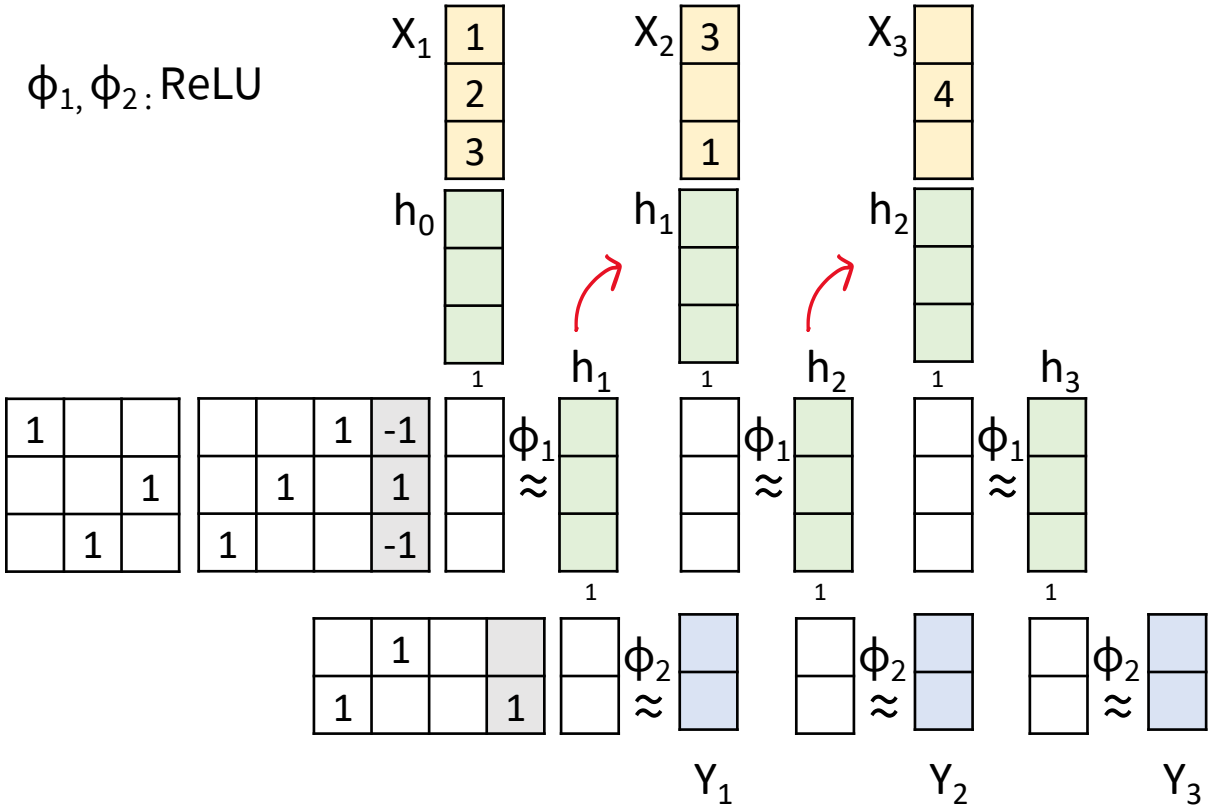
1 + 1 + 1 = 3

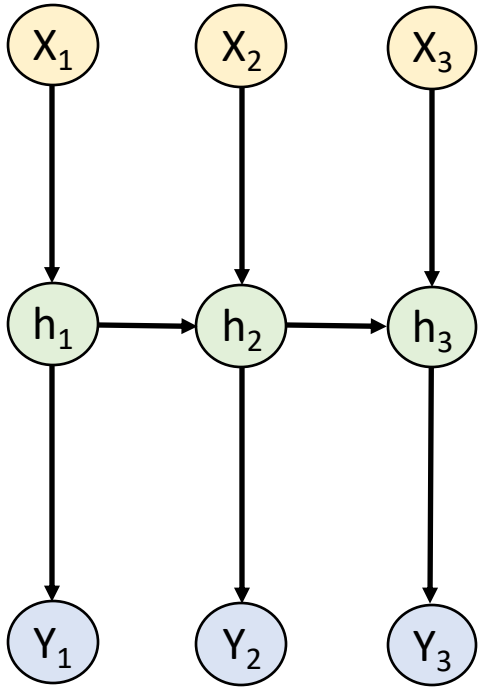
Variations of RNNs

CSCI 5722 Computer Vision

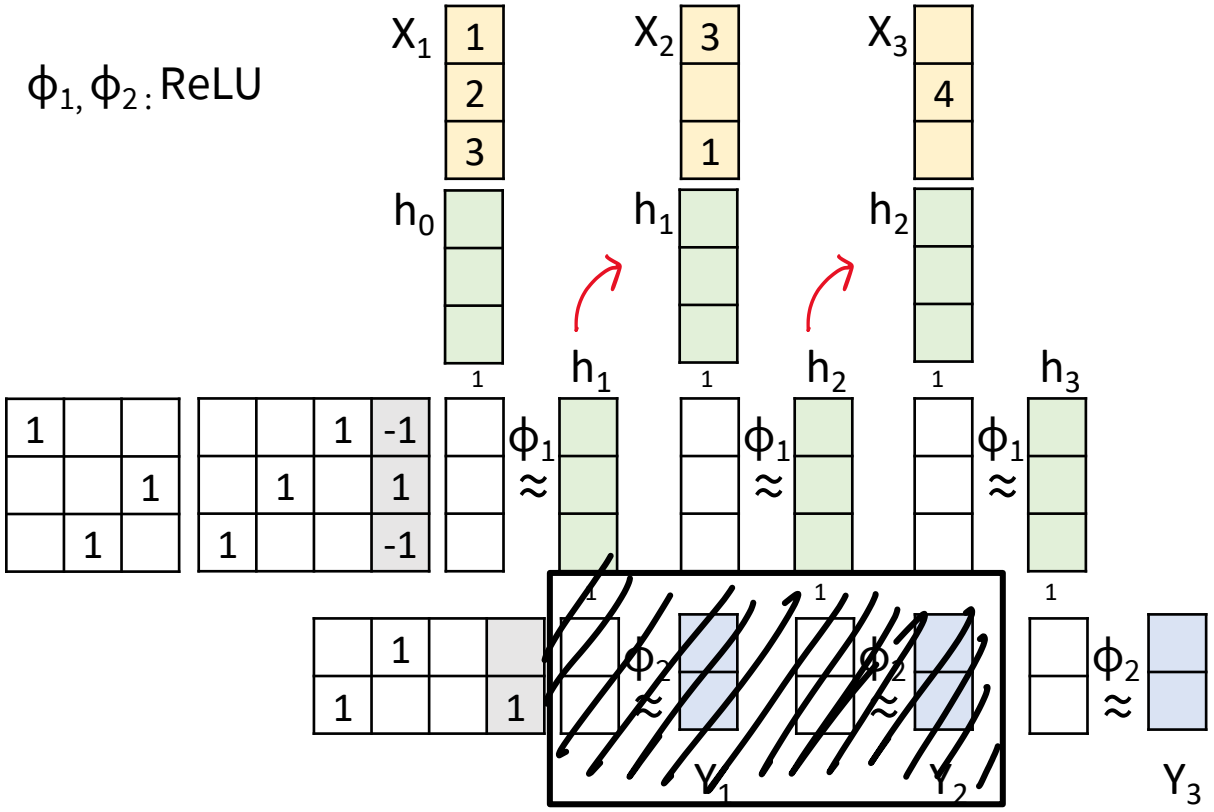
University of Colorado Boulder

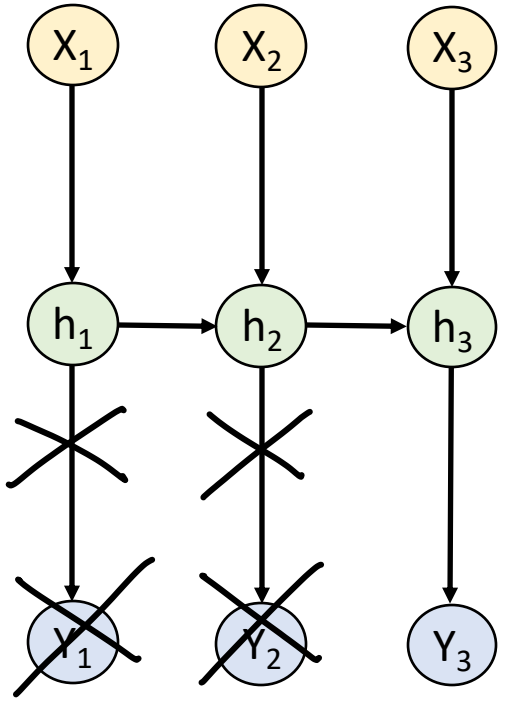
Many to Many


ϕ_1, ϕ_2 : ReLU
X_1
X_2
X_3
h_0
h_1
h_2
h_3
ϕ_1
ϕ_2
Y_1
Y_2
Y_3

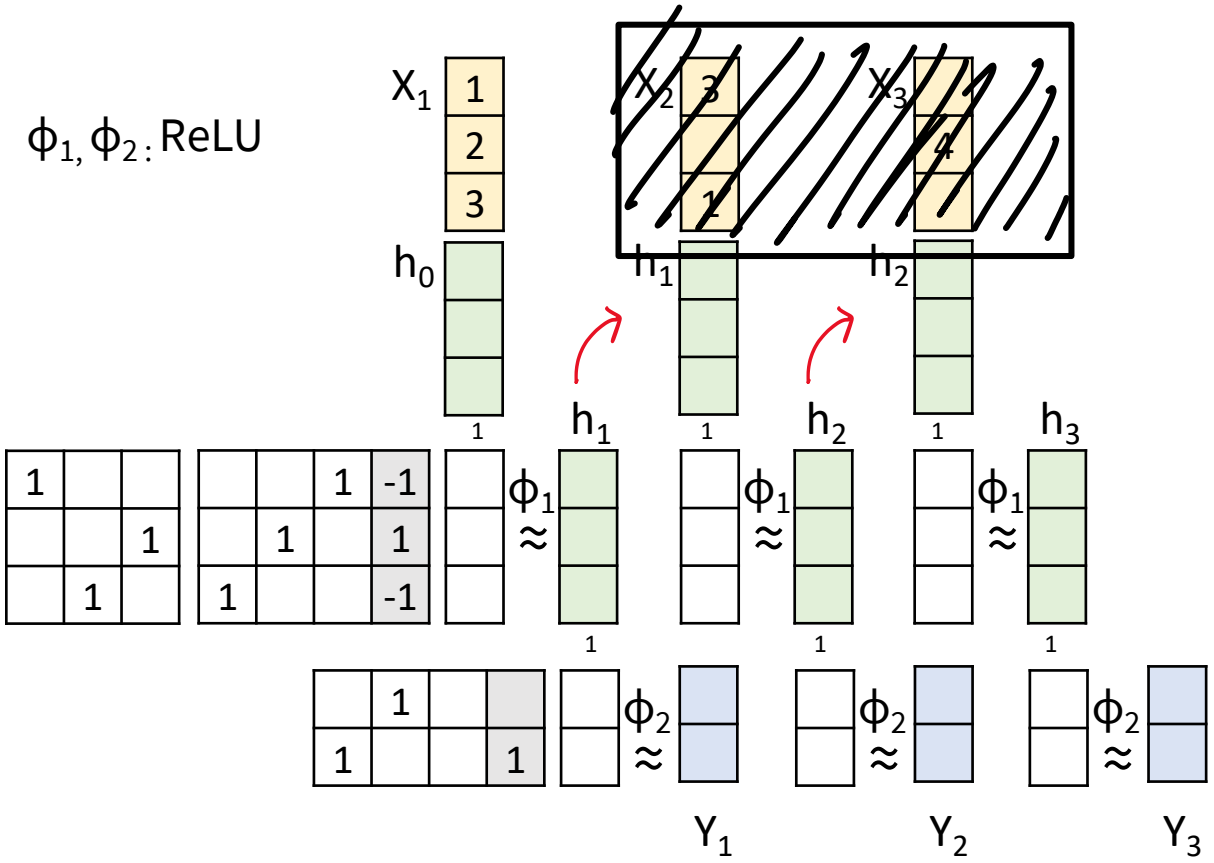

X_1
X_2
X_3
h_1
h_2
h_3
Y_1
Y_2
Y_3

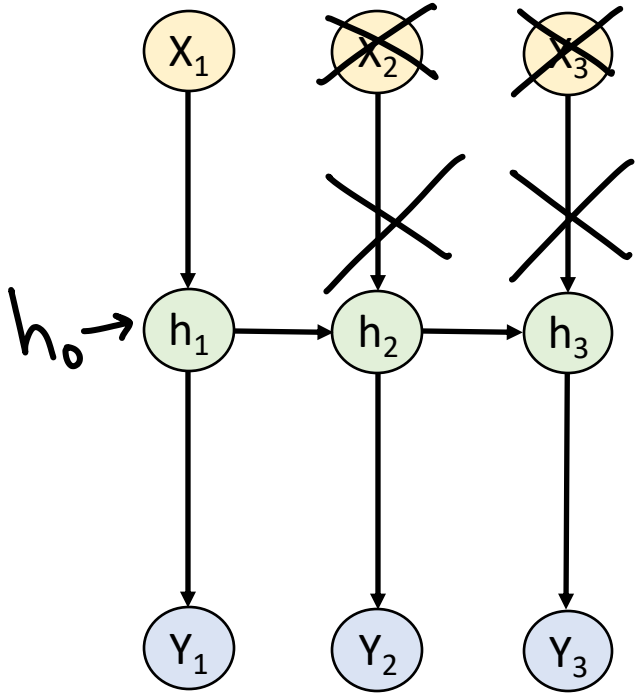
Many to One


ϕ_1, ϕ_2 : ReLU
X_1
X_2
X_3
h_0
h_1
h_2
h_3
ϕ_1
ϕ_2
Y_1
Y_2
Y_3

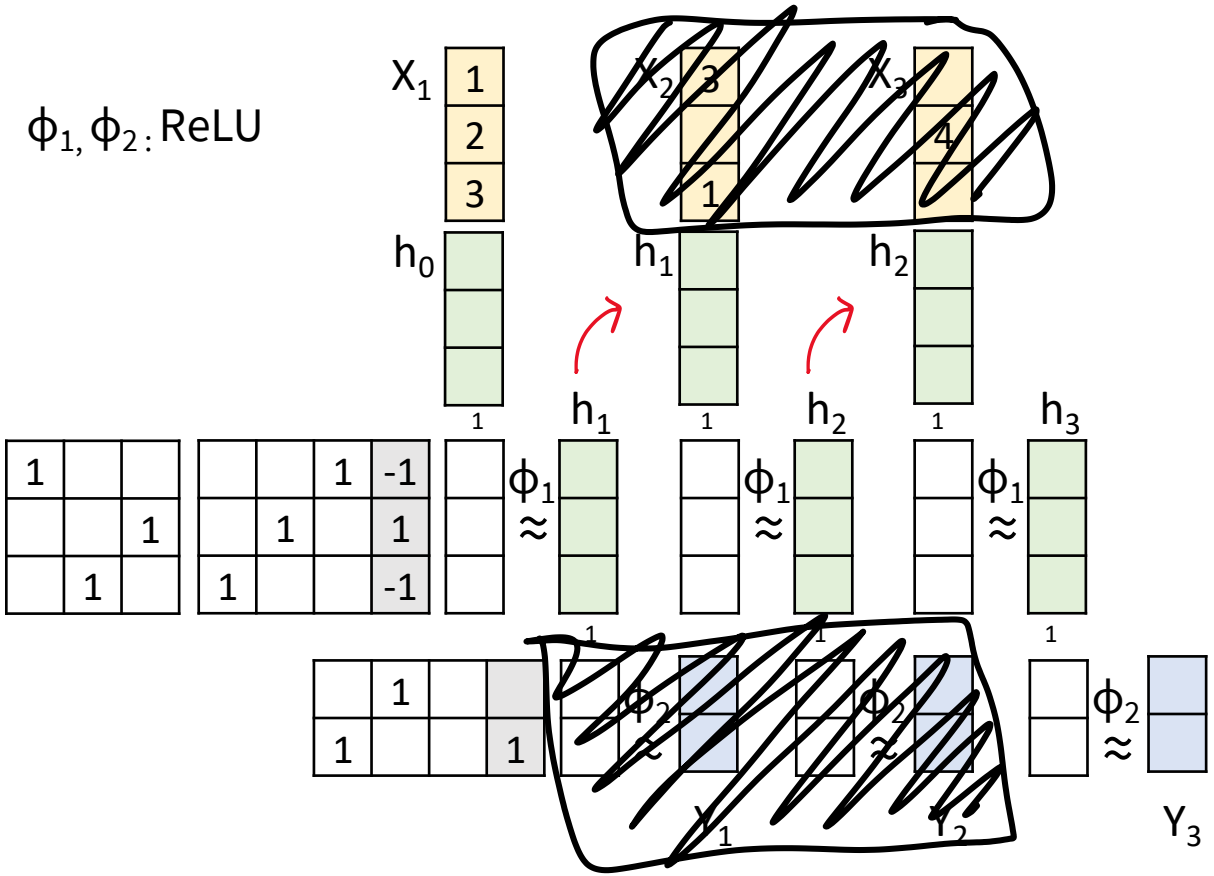

X_1
X_2
X_3
h_1
h_2
h_3
Y_1
Y_2
Y_3

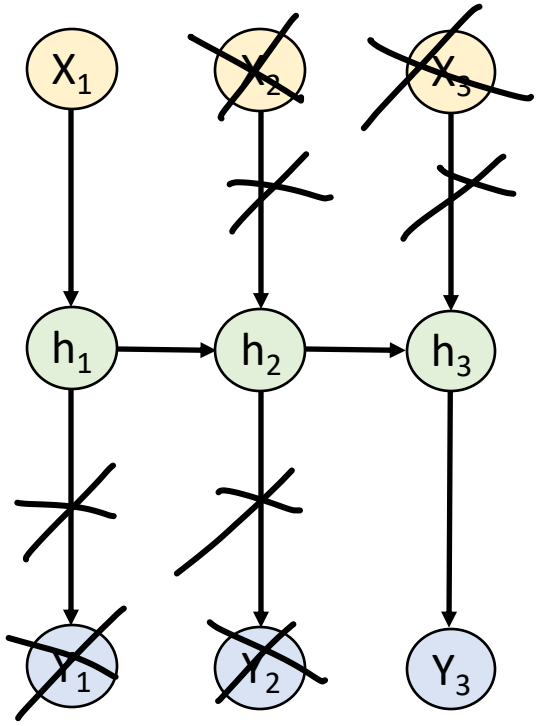
One to Many


ϕ_1, ϕ_2 : ReLU
X_1
X_2
X_3
h_0
h_1
h_2
h_3
ϕ_1
ϕ_2
Y_1
Y_2
Y_3


X_1
X_2
X_3
h_0
h_1
h_2
h_3
Y_1
Y_2
Y_3

One to One


ϕ_1, ϕ_2 : ReLU
X_1
X_2
X_3
h_0
h_1
h_2
h_3
ϕ_1
ϕ_2
Y_1
Y_2
Y_3

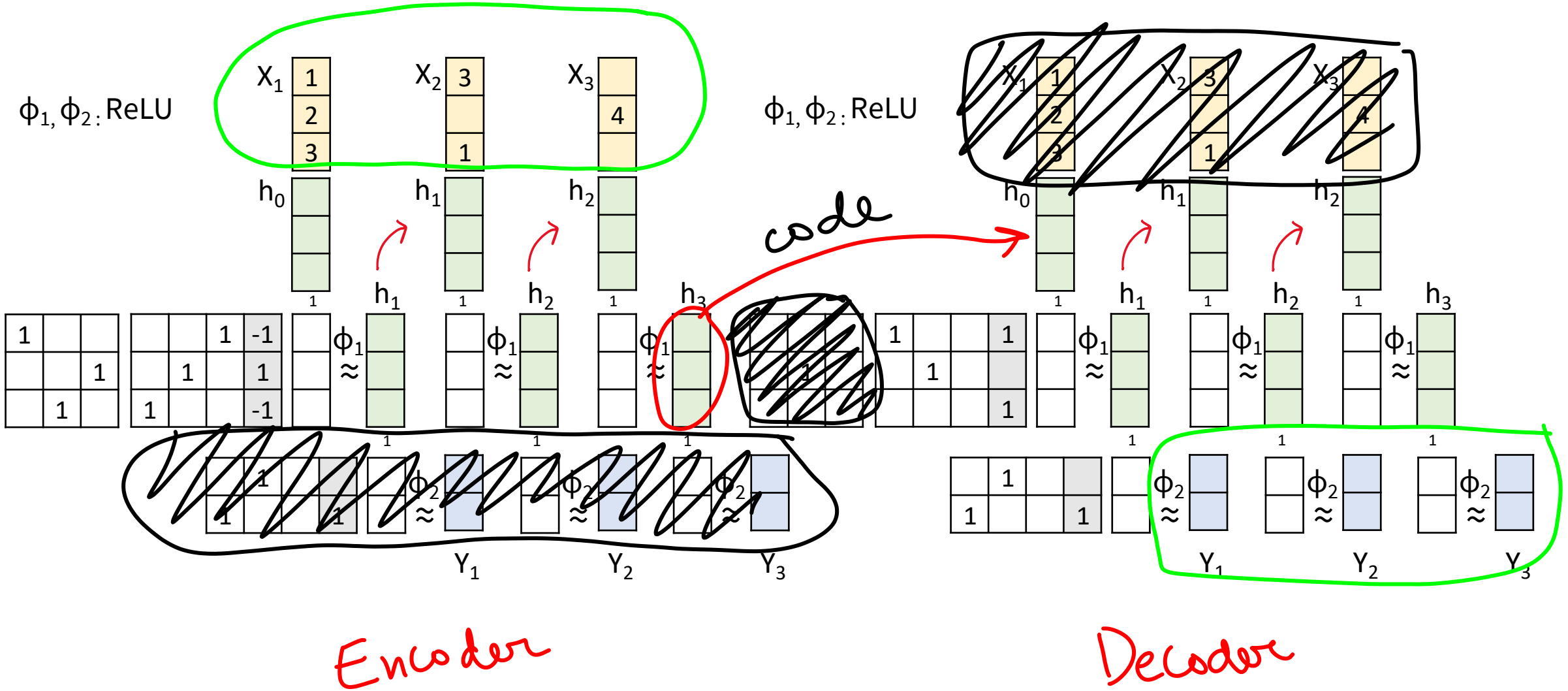

X_1
X_2
X_3
h_1
h_2
h_3
Y_1
Y_2
Y_3

Advanced RNN Architectures

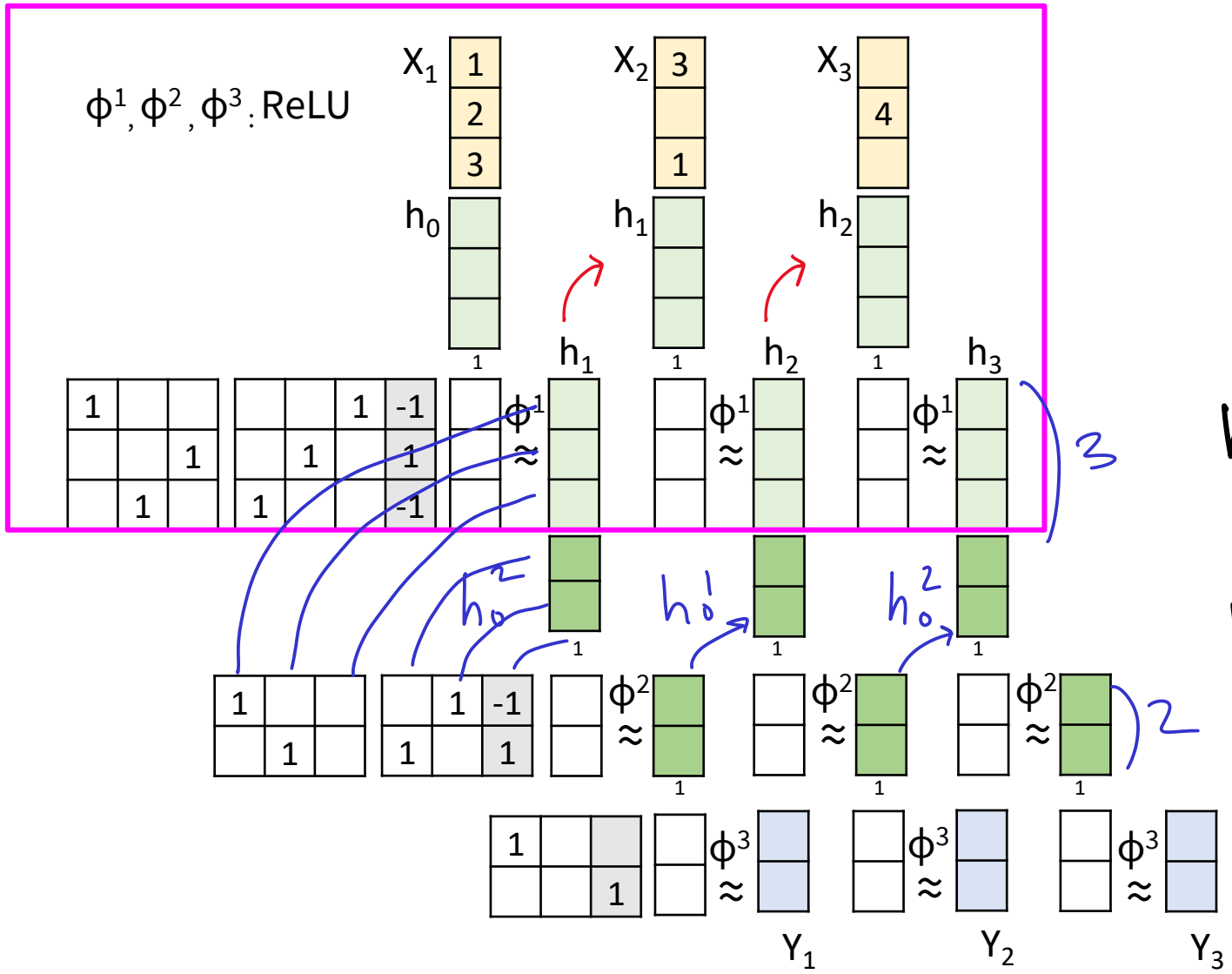
CSCI 5722 Computer Vision

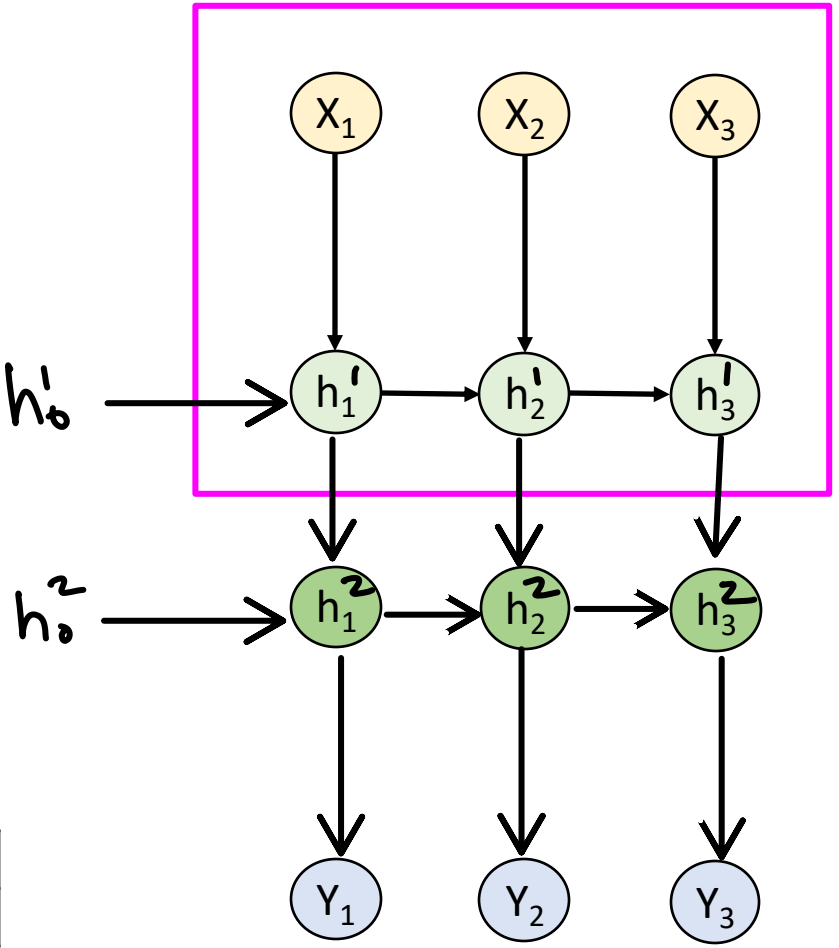
University of Colorado Boulder

Sequence to Sequence

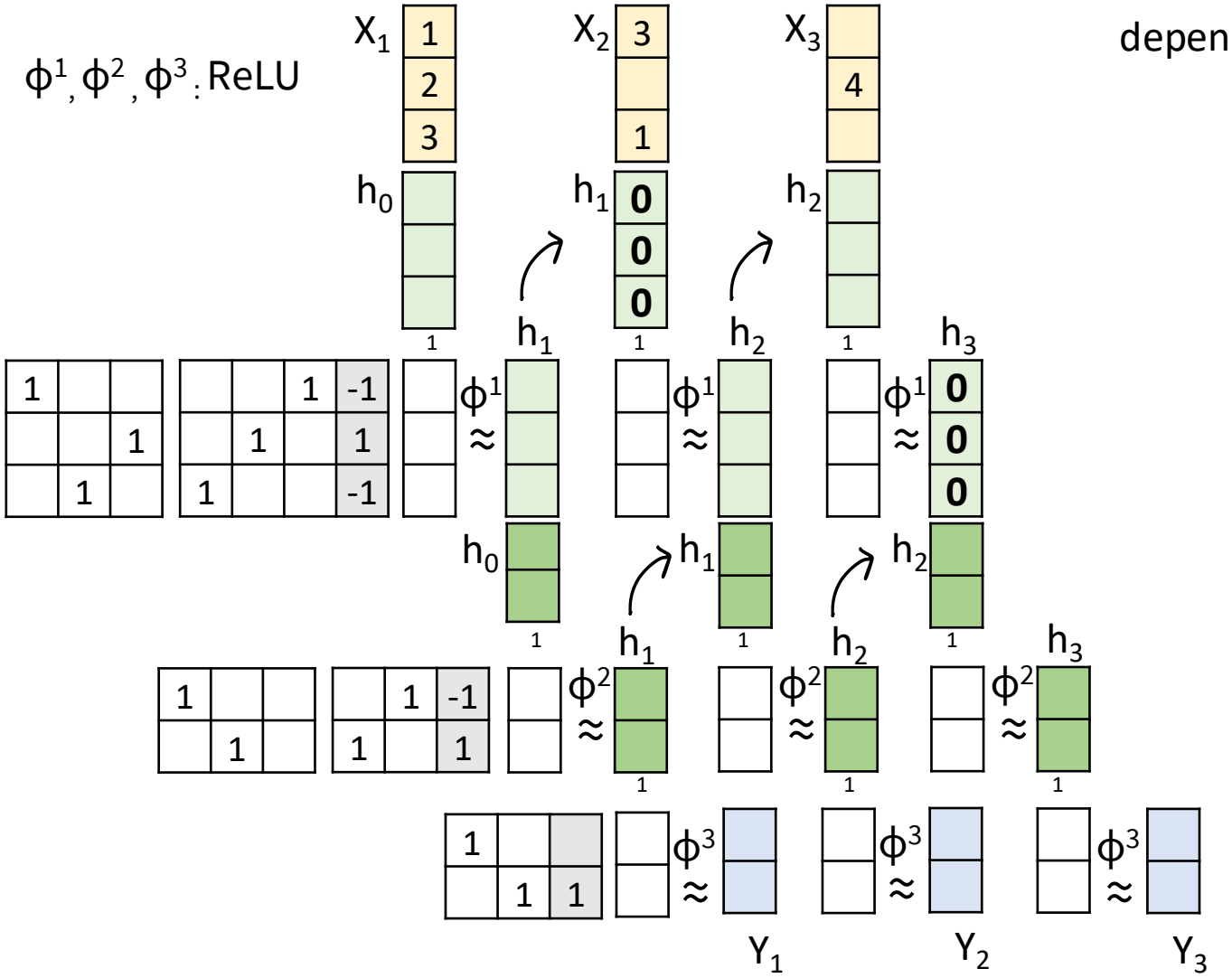


Multilayer RNN

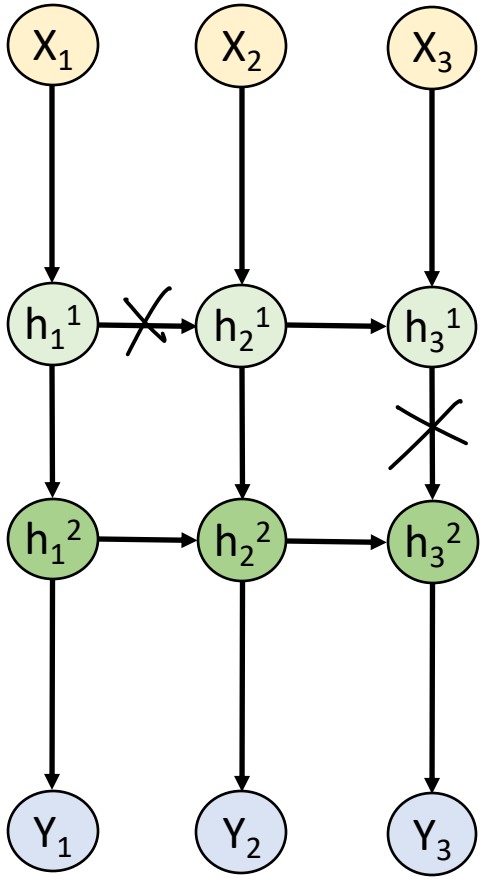




Identify links

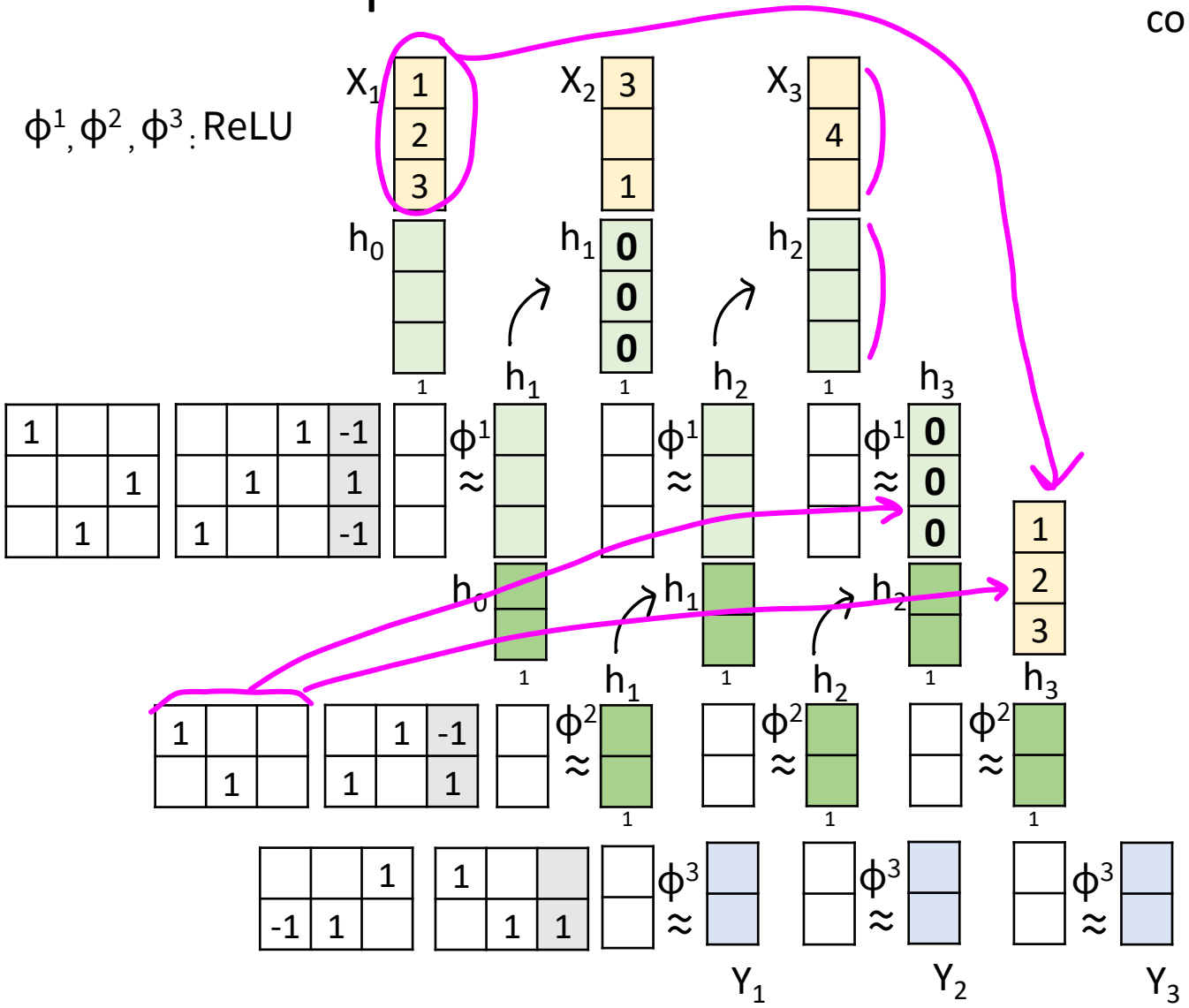


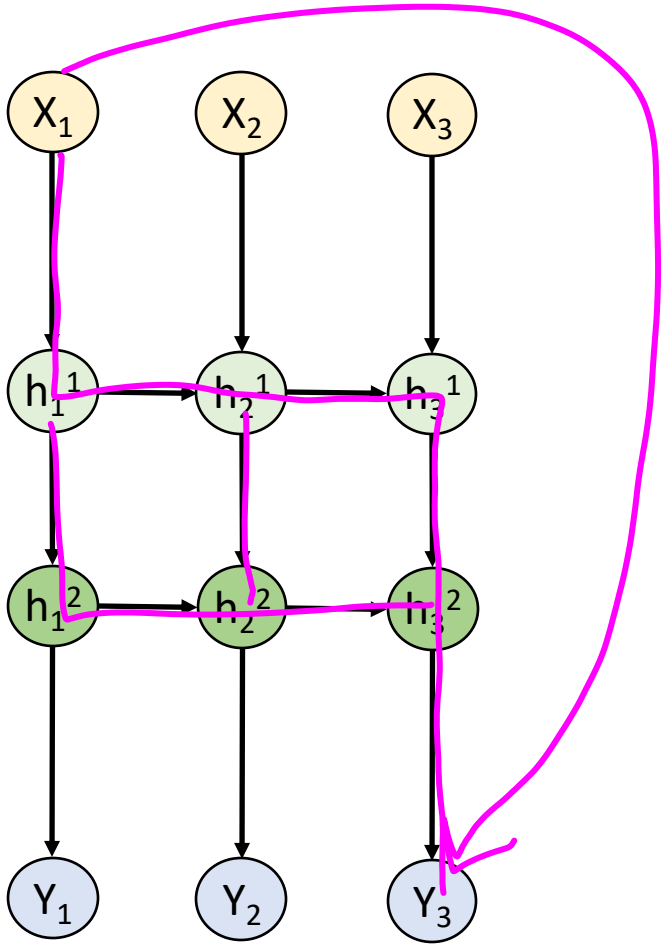
Suppose we set some hidden states to zeros on purpose. Cross out the affected dependency links.



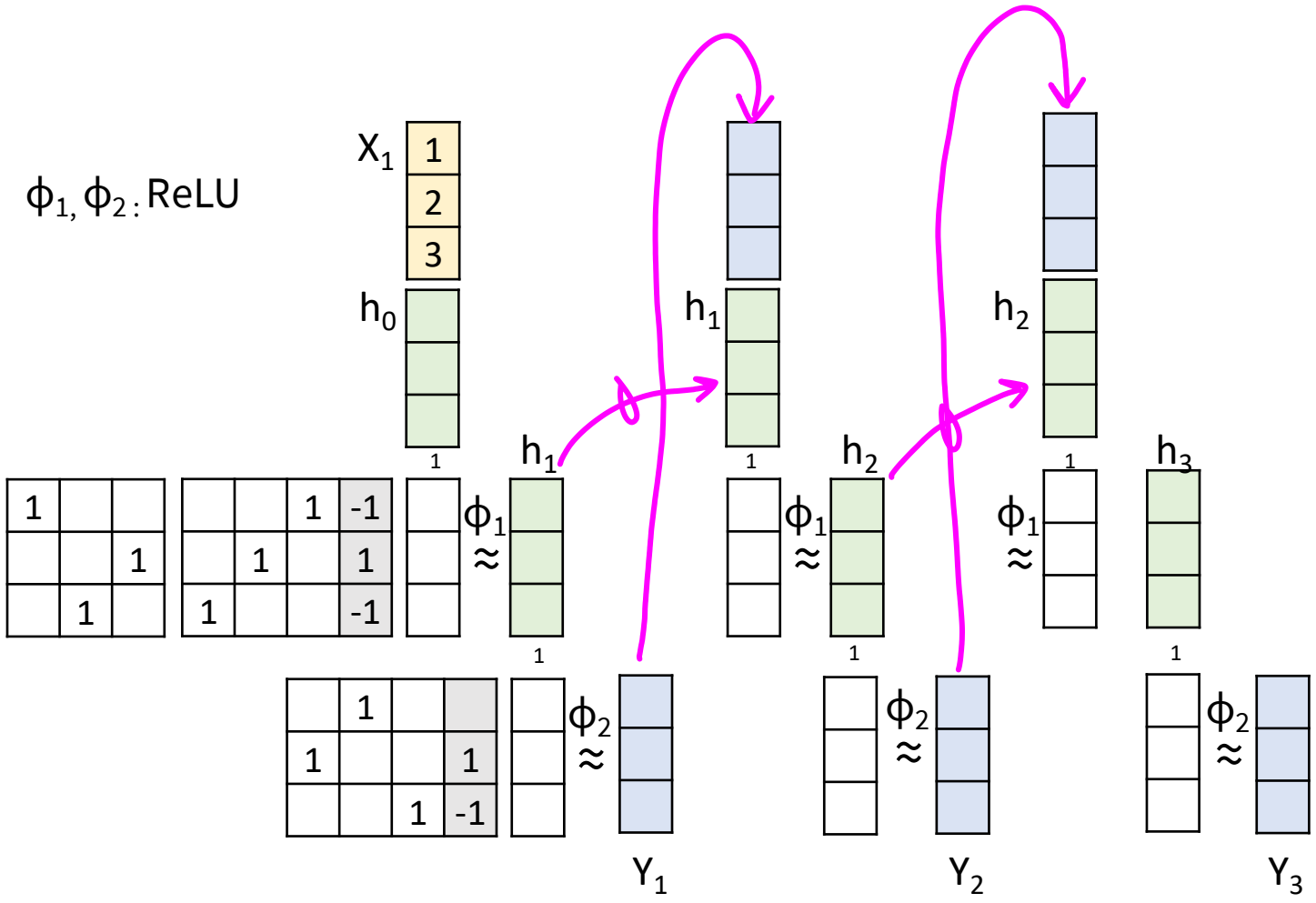
Add a skip connection

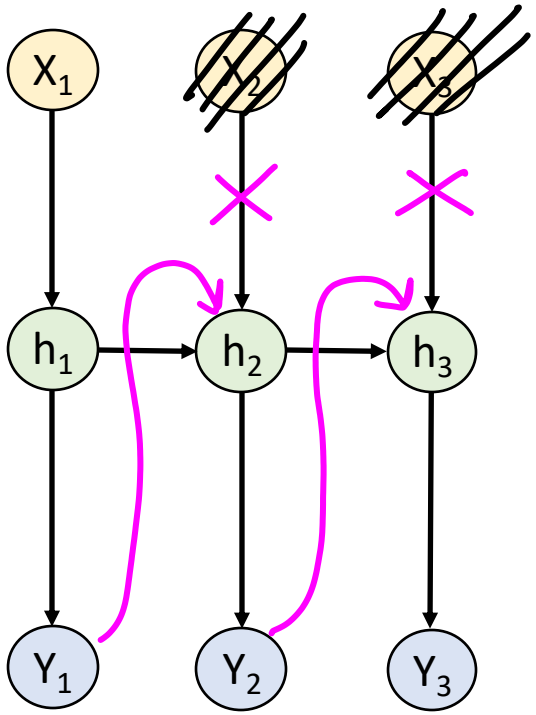
Draw a new arrow to illustrate the skip connection from X_1.



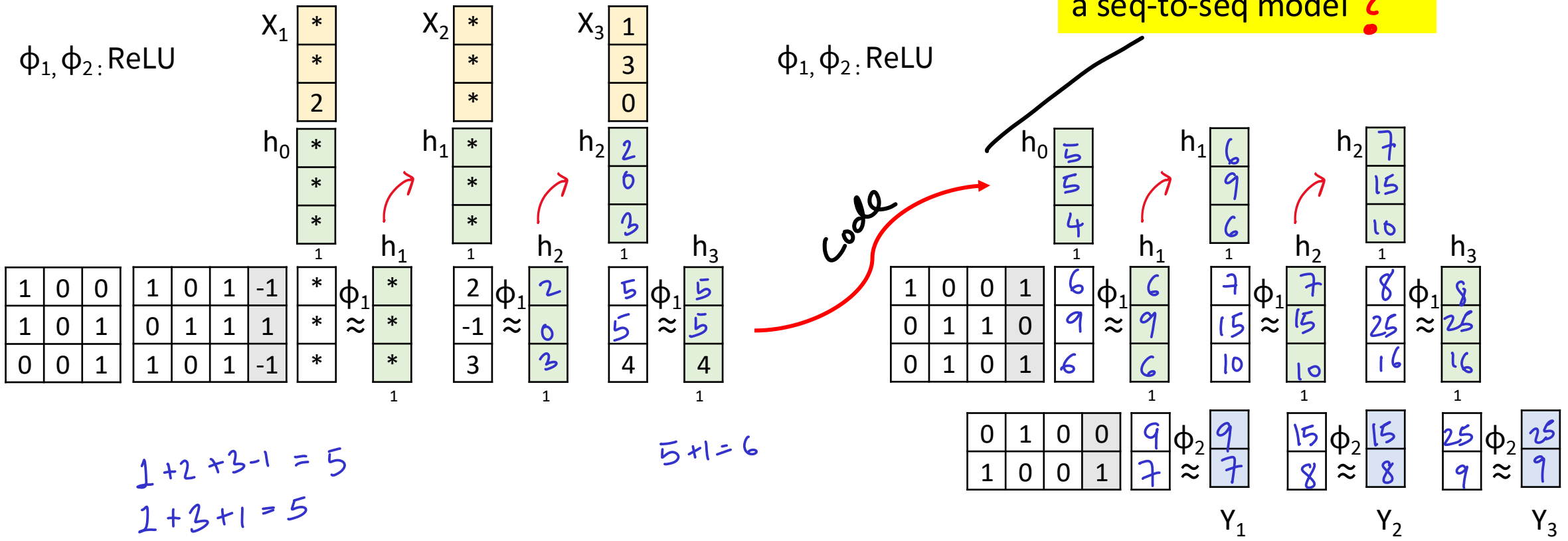


Auto-Regressive RNN

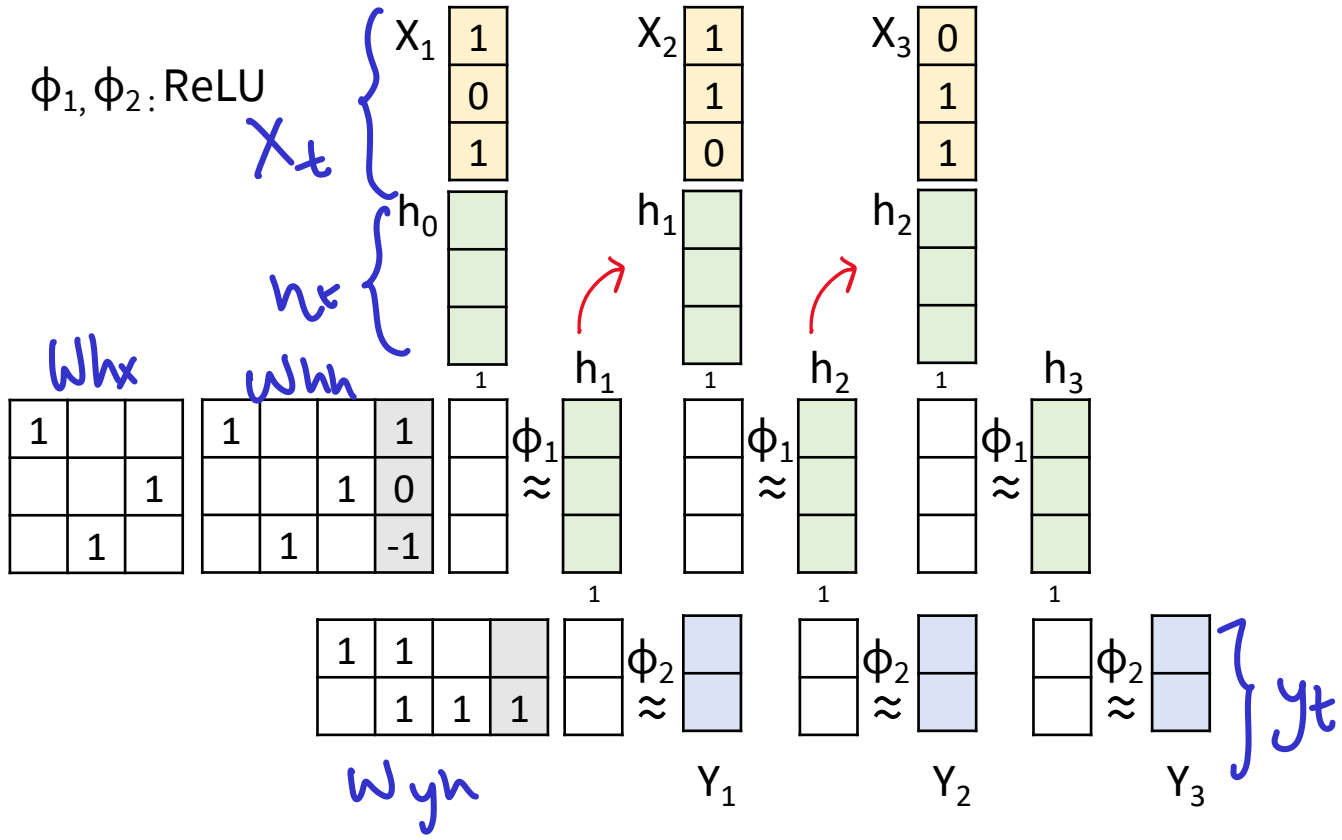




Sequence to Sequence



Counting Parameters (small)



size(X_t) = 3x1

size(h_t) = 3x1

size(y_t) = 2x1

size(W_{hx}) = 3x3

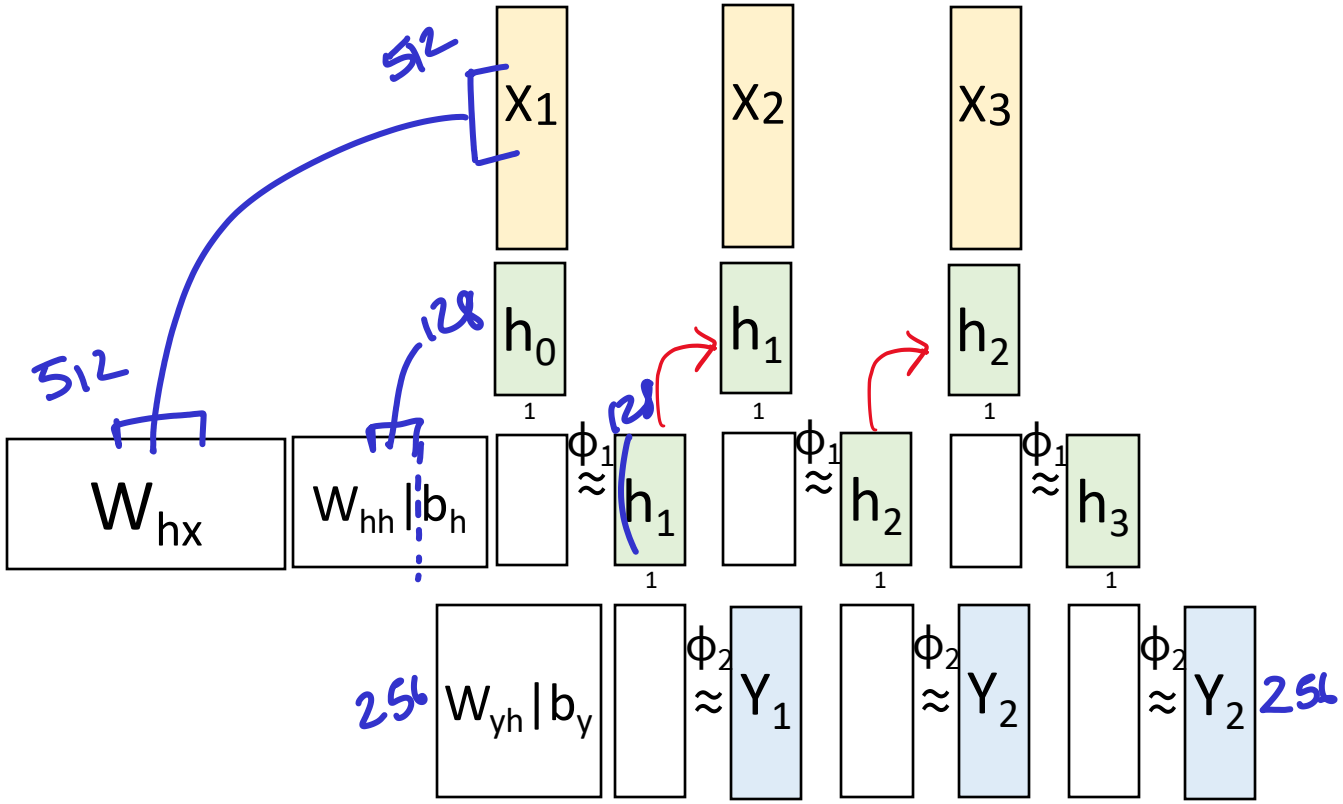
size(b_h) = 3x1

size(W_{hh}) = 3x3

size(W_{yh}) = 2x3

size(b_y) = 2x1

Counting Parameters (large)



size(X_t) = 512 x 1

size(h_t) = 128 x 1

size(Y_t) = 256 x 1

size(W_{hx}) = 512 x 128

size(b_h) = 128 x 1

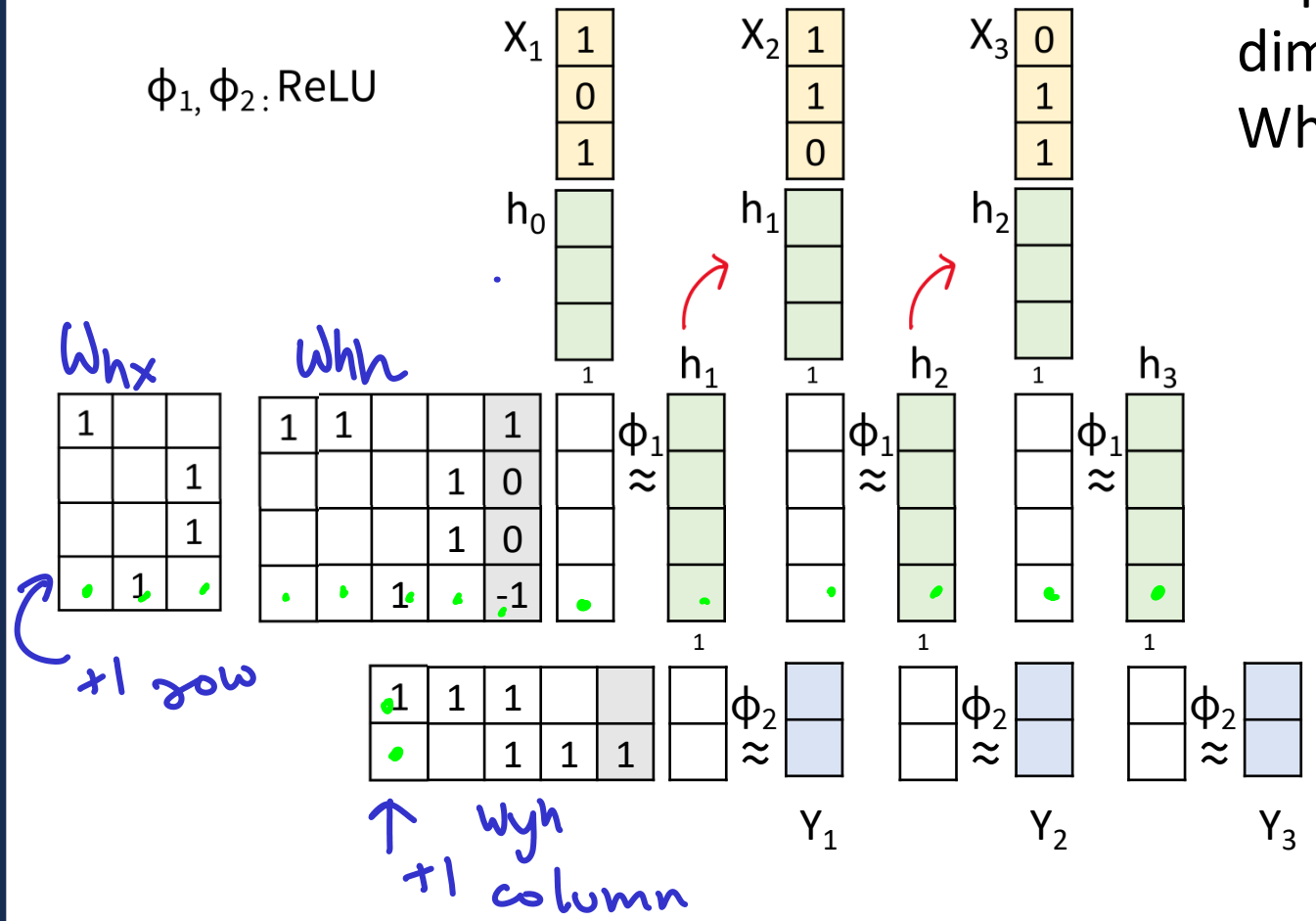
size(W_{hh}) = 128 x 128

size(W_{yh}) = 256 x 128

size(b_y) = 256 x 1

Adding Parameters

I copy pasted a screenshot



Suppose we increase the hidden state's dimension by 1.
What would be the new parameter sizes?

size(W_{hx}) = 4x3

size(b_h) = 4x1

size(W_{hh}) = 4x4

size(W_{yh}) = 2x4

size(b_y) = 2x1

# of New Parameters = 16

(Hint: You can try to draw the extra cells as visual aid)

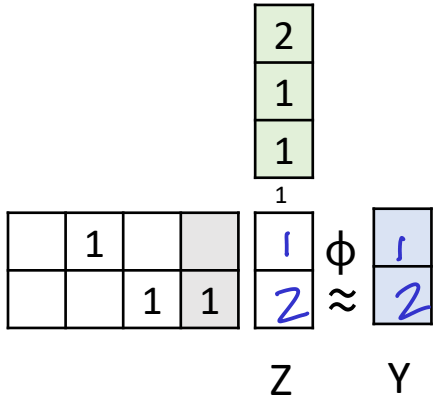
Modeling Probabilities

CSCI 5722 Computer Vision

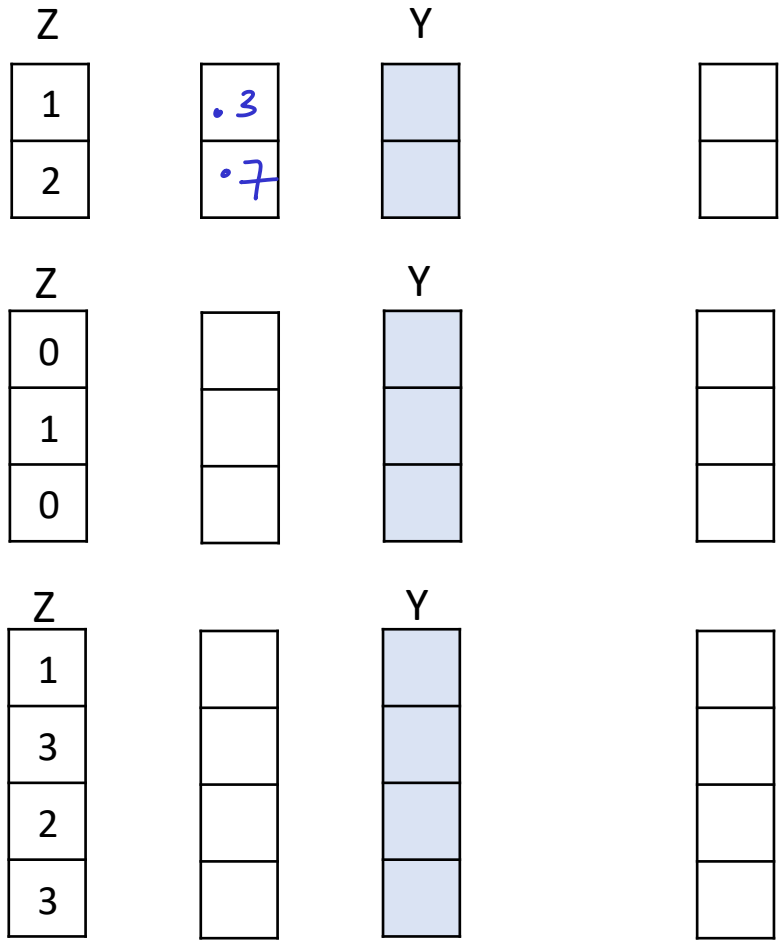
University of Colorado Boulder

Output Values → Probability Distribution

ϕ = ReLU



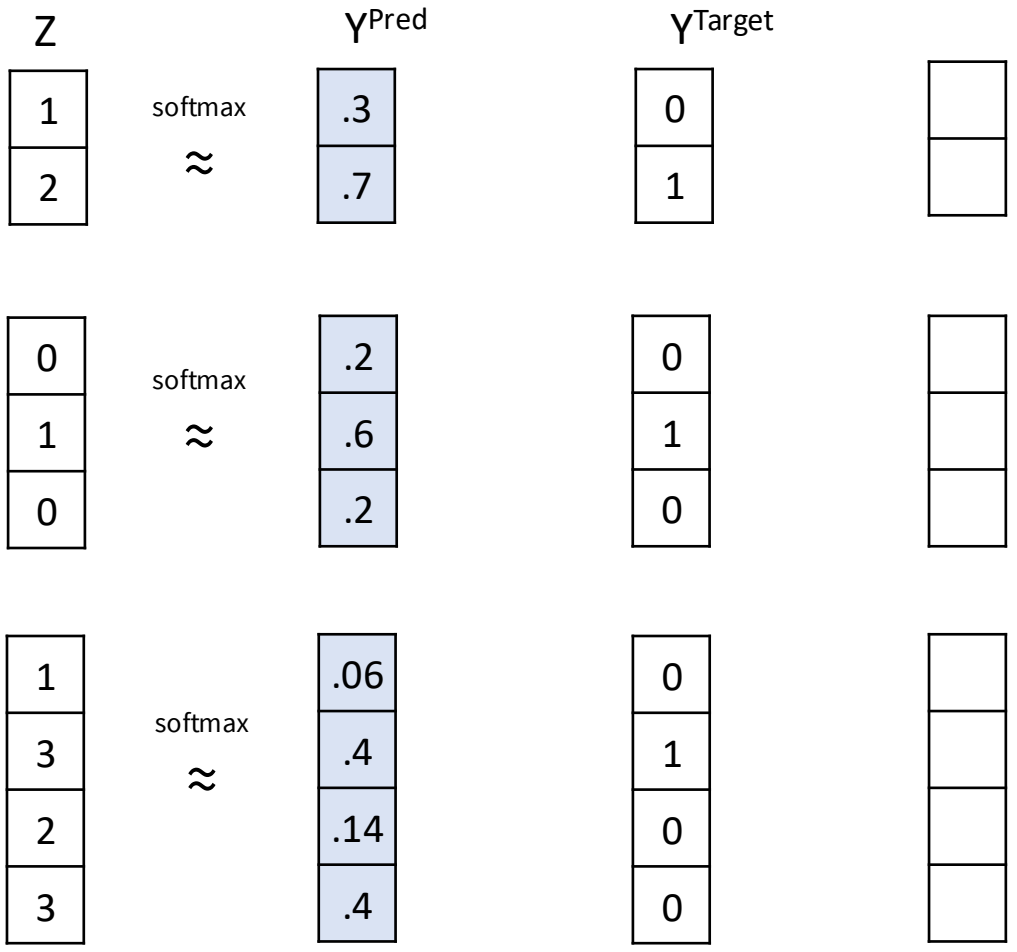
$Y = \phi(Z) =$



x	e^x	round
0	1	1
1	2.71828	3
2	7.38906	7
3	20.08554	20
4	54.58715	55
5	148.41316	148

Gradient of Softmax + CE Loss

$$\frac{\partial L}{\partial Z} = \_\_\_\_\_$$

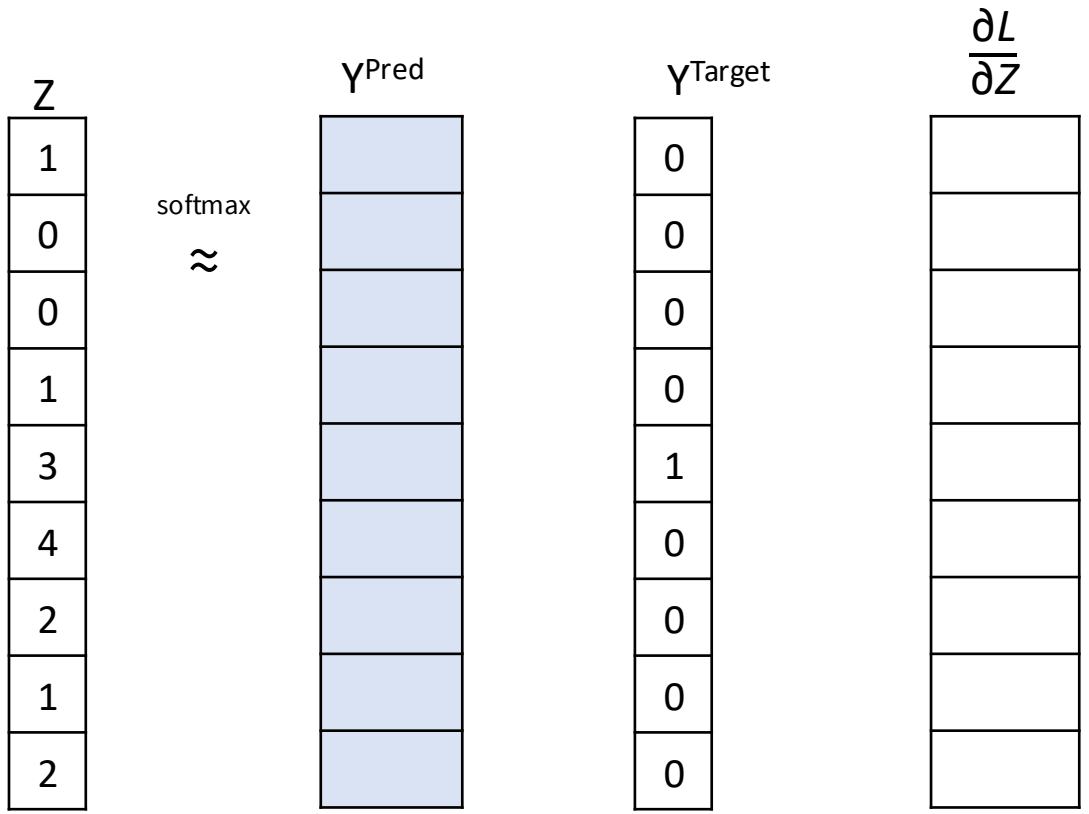


$$Softmax = S(y_i) = \frac{e^{y_i}}{\sum e^{y_j}}$$

$$CE\ loss = \frac{-1}{N} \sum y_i \log(\hat{y}_i)$$

Calculate the Gradient of Softmax + CE Loss

x	e^x	round
0	1	1
1	2.71828	3
2	7.38906	7
3	20.08554	20
4	54.58715	55
5	148.41316	148



Calculate the Gradient of Softmax + CE Loss

$$\text{CE loss} = -\frac{1}{N}\sum y_i \log(\hat{y}_i)$$

$$S(y_i) = \frac{e^{y_i}}{\sum e^{y_i}}$$

x	e^x	round
0	1	1
1	2.71828	3
2	7.38906	7
3	20.08554	20
4	54.58715	55
5	148.41316	148

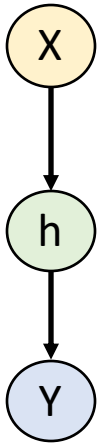
Z	softmax ≈	Y^{Pred}	Y^{Target}	$\frac{\partial L}{\partial Z}$
0		0.01	0	0.01
1		0.03	0	0.03
4		0.55	0	-0.55
1		0.03	0	0.03
0		0.01	1	0.01
3		0.20	0	0.20
2		0.07	0	0.07
2		0.07	0	0.07
1		0.01	0	0.02

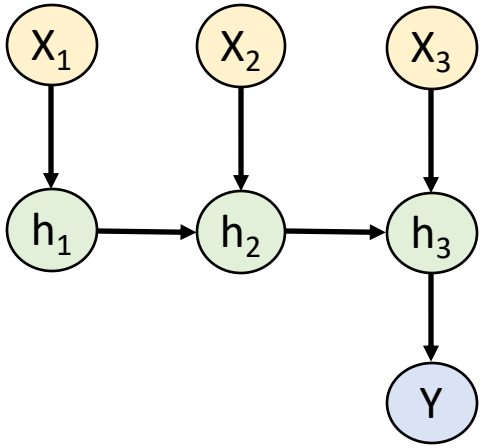
$2e^0 + 3e^1 + 2e^2 + 1e^3 + 1e^4$

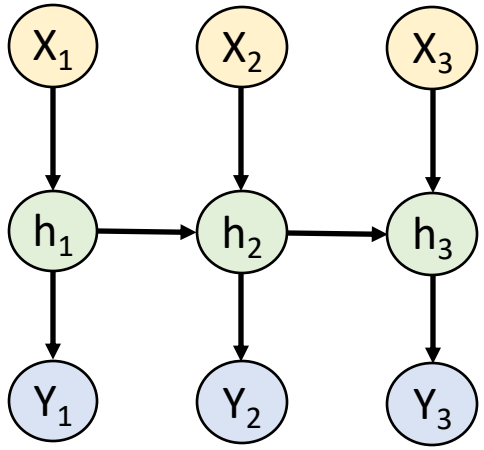
$= 2(1) + 3(3) + 2(7) + 20 + 55$

$= 100$

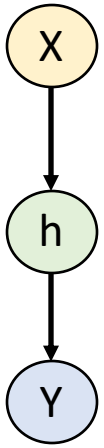
Model → Function

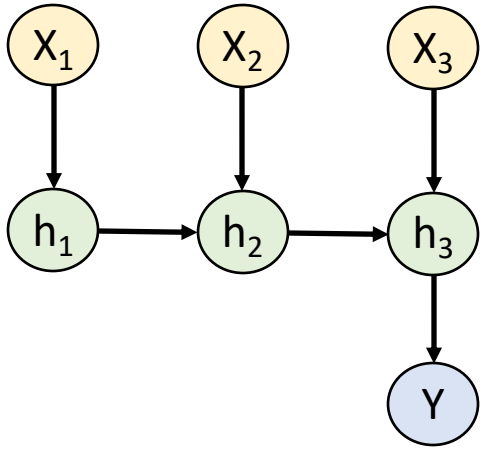


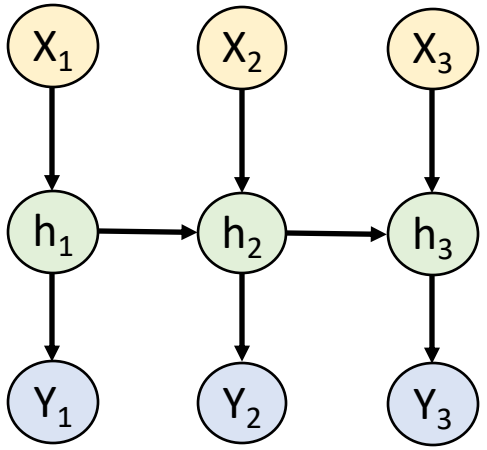




Function → Conditional Probability Function







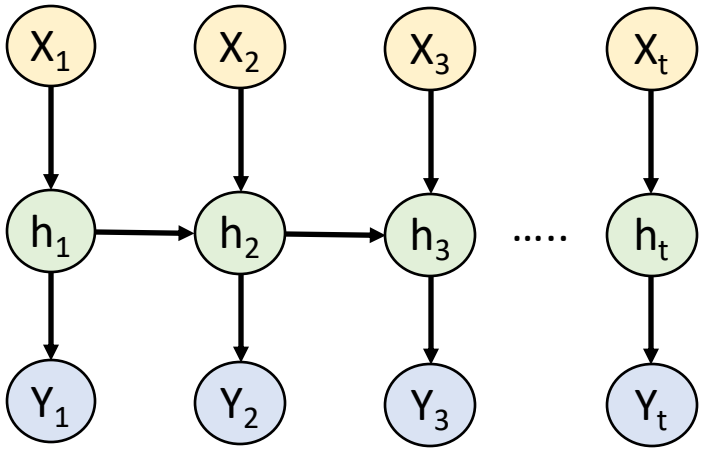
$Y = f(X)$

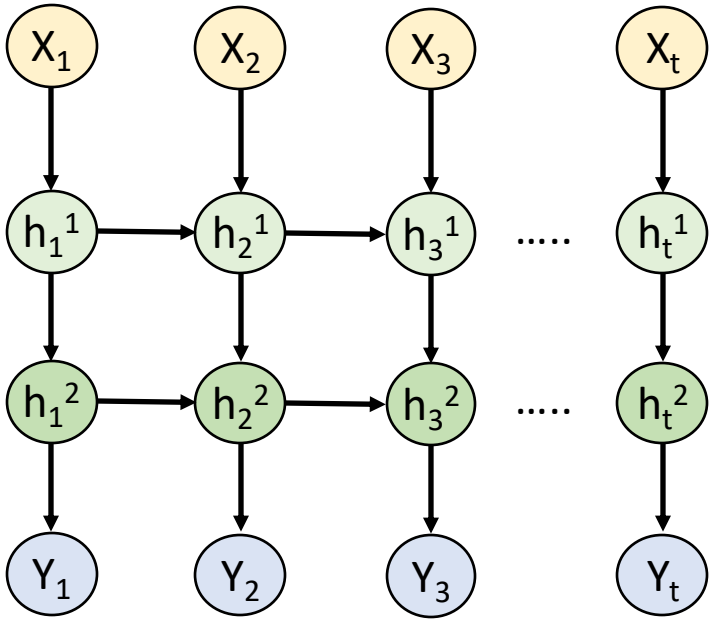
$Y = f(X_1, X_2, X_3)$

$Y = f(Y_t \mid X_t, X_{t-1}, \ldots X_1)$

Model → Conditional Probability Distributions

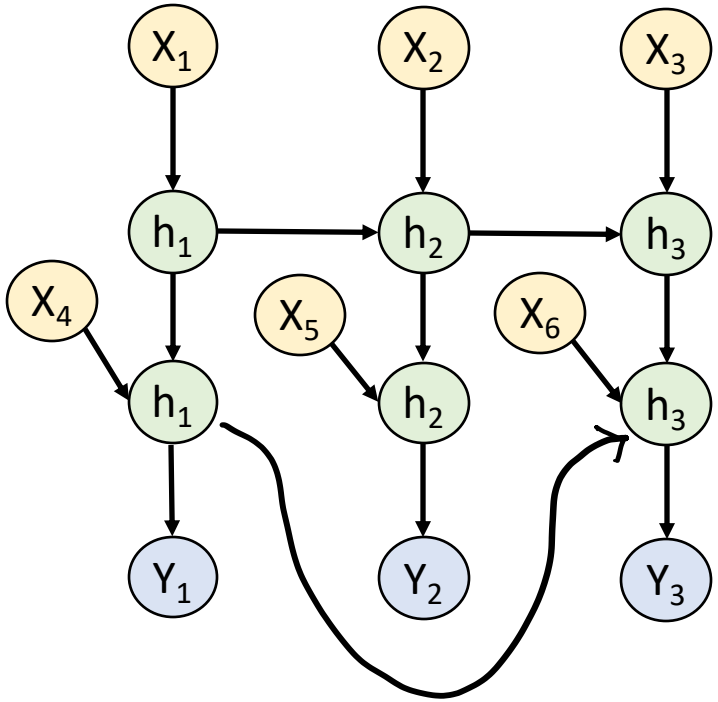
What conditional probability distributions could these two models estimate?





$P(Y_t \mid \quad)$ $P(Y_t \mid \quad)$

Model → Conditional Probability Distributions



What conditional probability distributions could this model learn to estimate?

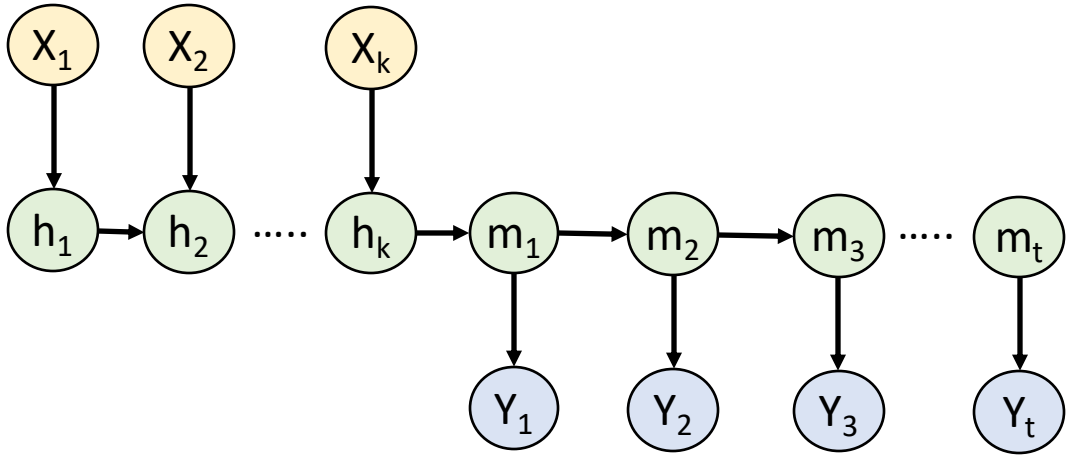
$P(Y_1 \mid \quad)$

$P(Y_2 \mid \quad)$

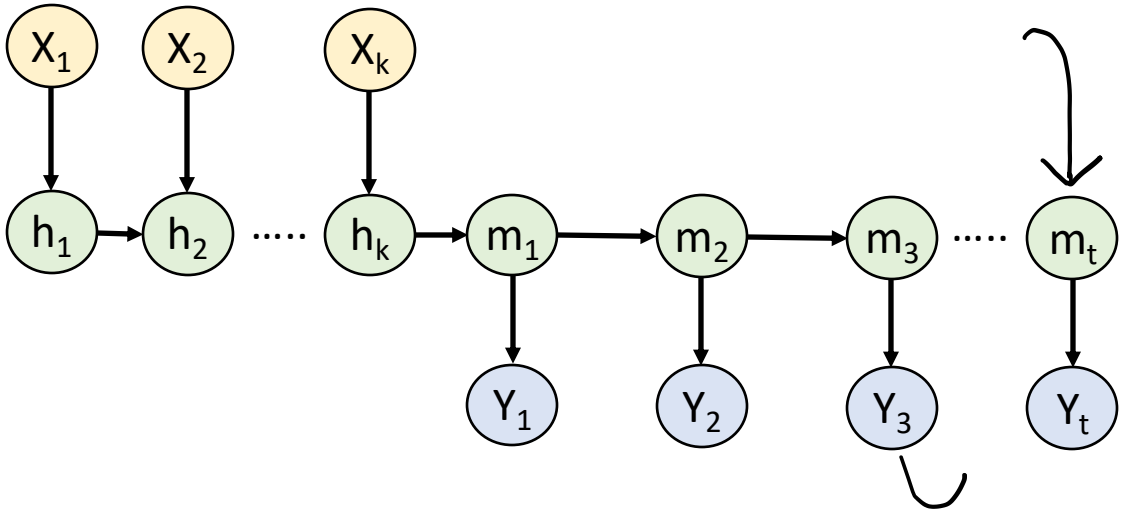
$P(Y_3 \mid \quad)$

Model → Conditional Probability Distributions

What conditional probability distributions could these two seq-to-seq models estimate?



$P(Y_t \mid \quad)$



$P(Y_t \mid \quad)$

Identify "No Dependency" Links

Cross out the dependency links to match the matrix form.