Topics

① word2vec

② RNN (simple RNN)
- ✓ — Sequential data
  - Time series data
  - Sentences
- Theory
- Time series data
- Char level prediction
- Word level prediction
- Implementation (Keras / Tensorflow)

③ LSTM
- Implementation
- Theory

Concept
- Theory using example
- math intuition + derivation
- Implementation

① word2vec

② RNN
- Simple RNN
- LSTM

③ Attention models / Transformers

④ Real time case studies
- ↳ Speech recognition
- ↳ chat bot (RASA)
- ↳ Sentiment analysis

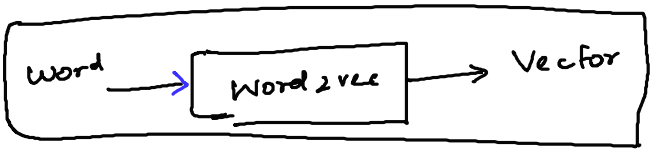
Word 2 vec

→ Document term matrix
- Sparse (0's)
- High dimension
- Content

Words → ??? → Vectors

→ LSA

→ Word embeddings
- word2vec ✓
- Glove ✓
- fast text



$[d_1, d_2, \ldots d_M]$

$M \simeq 300$

Objective
- Should be dense
- Should be in lower dimension (300)
- Should represent meaning of the word

  Sim(benz, Audi) ~ high

  Sim(King, man) ~ high

  Semantic field ~ hospital (Doctors, nurses.., surgeon, ..)
- Should be comparable with each other

Distributional semantics

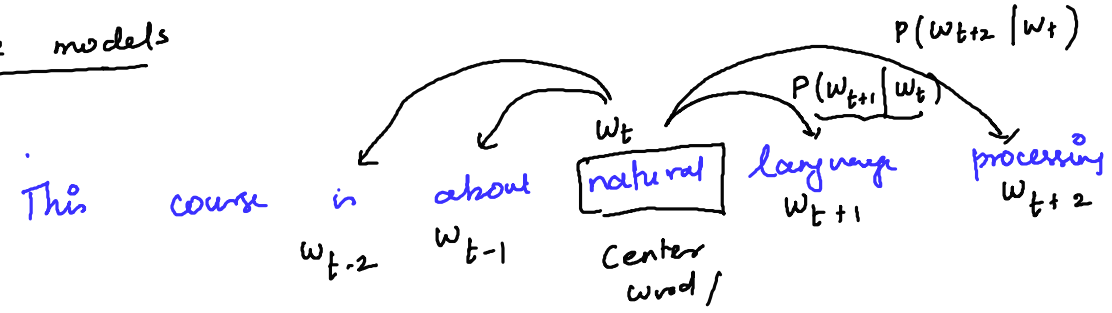
"You shall know a word by the company it keeps"
(J. R. Firth 1957)

debt on a partial banking sector, crisis

shore, rivers, banking, sand

shore, rivers → Content words
banking → target word
sand → Content words

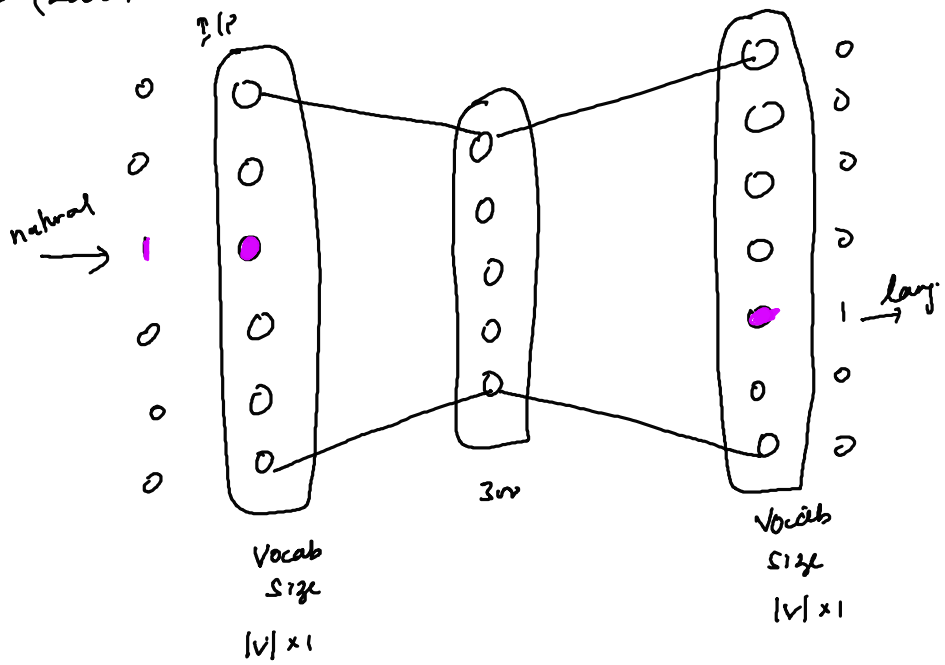
Language models



→ Probability model

→ Markov model

→ (2003) Neural network language model (NNLM) - Bengio et al 2003



Context	Target (t)
natural	language
language	processing

→ NNLM → Output layer → |V| x 1 → high

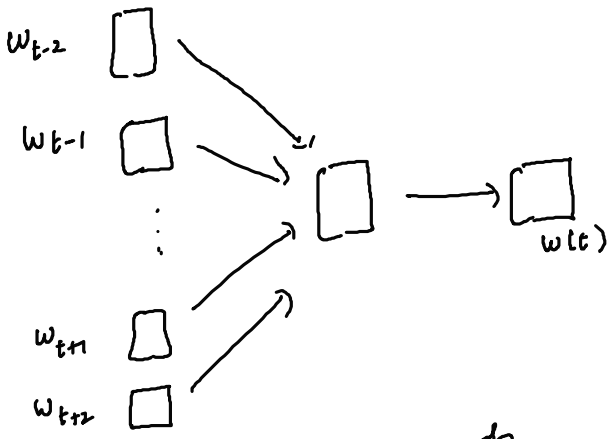
→ 2013 → word2vec

Word2 vec

→ Skip gram

→ Continuous bag of words (CBOW)

CBOW

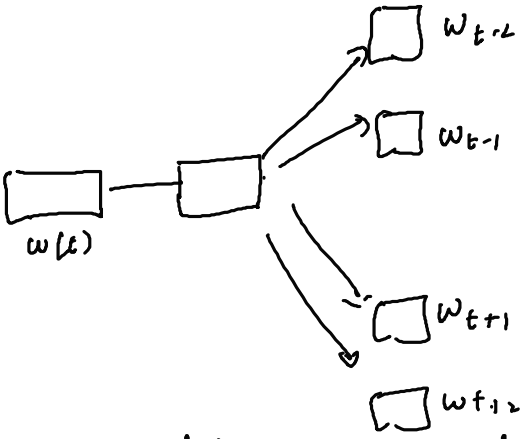


→ Predicts the center words given the context words

→ ✓ Extremely faster compared to skip gram

→ Good representation for frequent words

Skip Gram



→ Predict the context words given the target word

→ Slower compared to CBOW

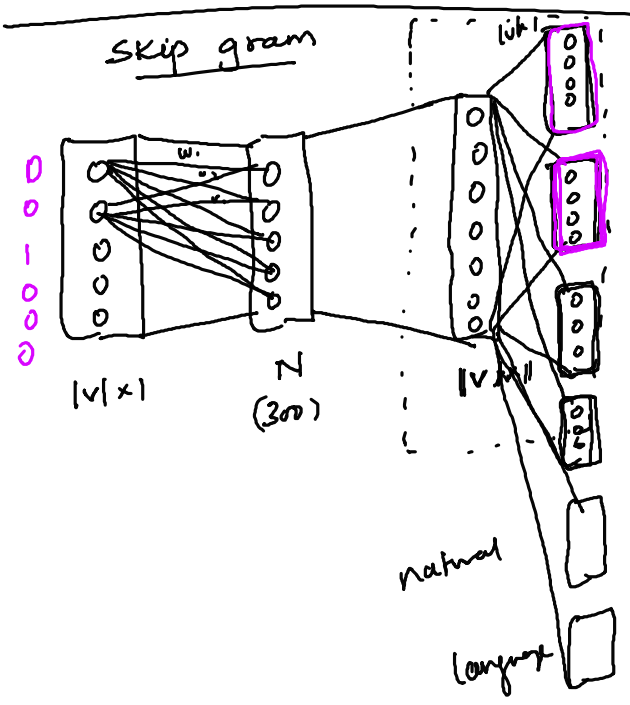
→ Good representation for rare words

CBOW & Skip gram → predict words ✓

Vector representation of words

Window size => 1

Skip gram



this

Sent1 → This course is about natural language

Sent2 → Next course is big data

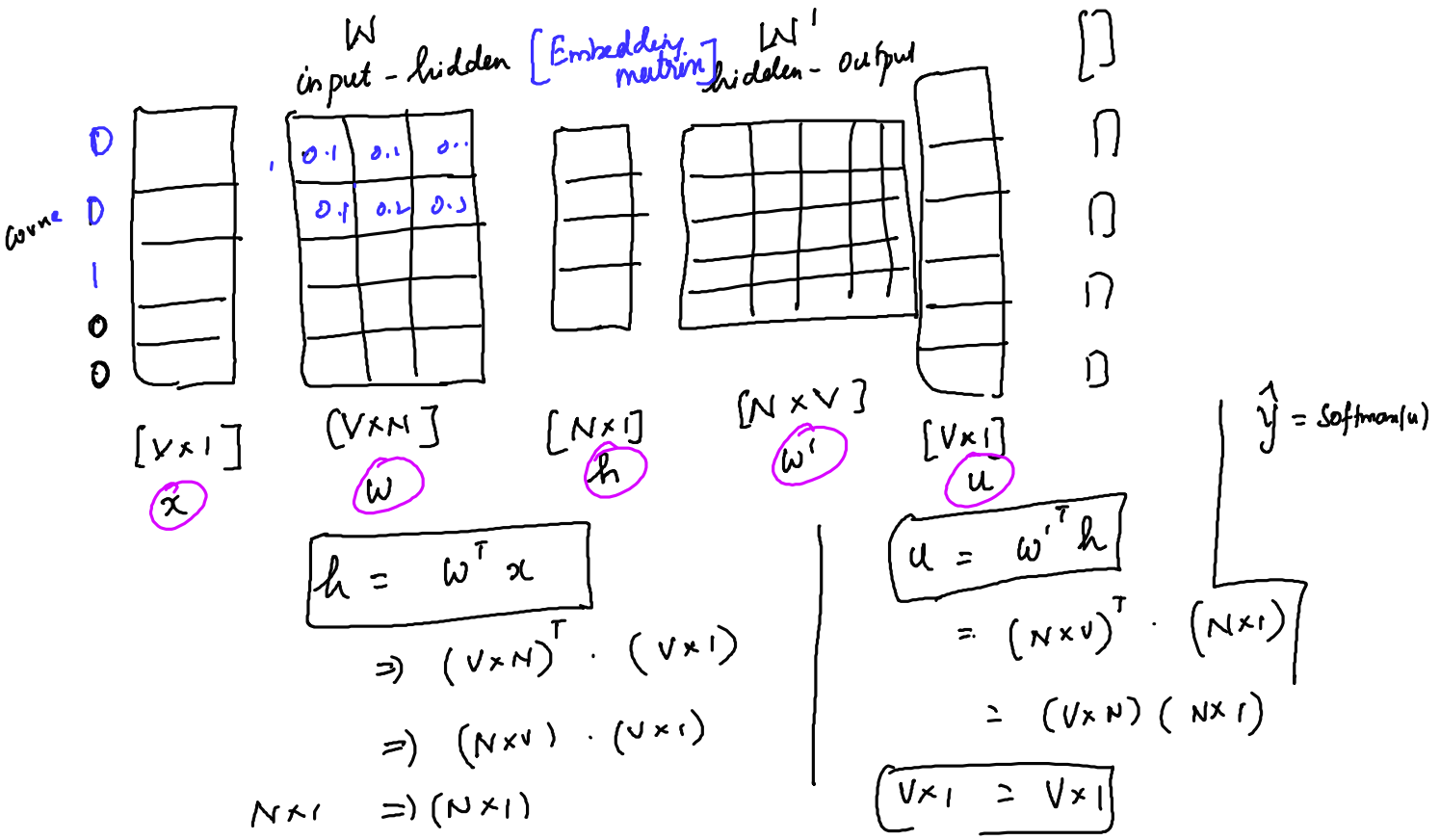
is

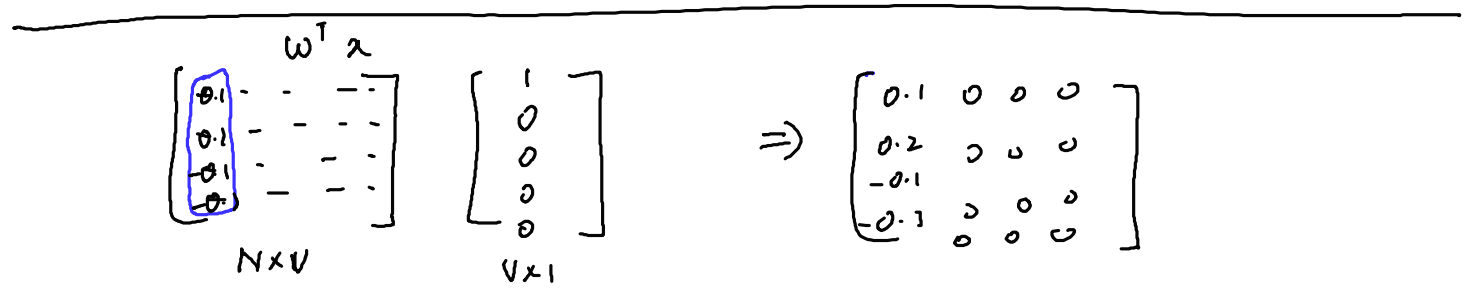
course ✓

about

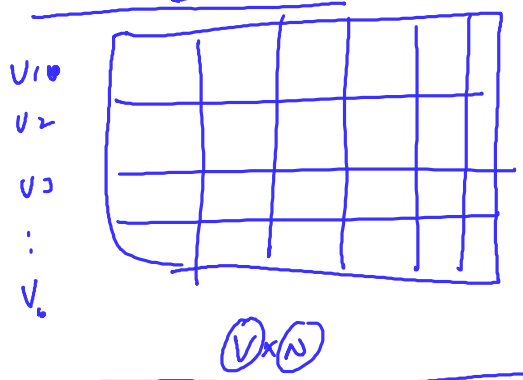
Center word	Content words
This ✓	course ✓
course	(this, is)
is	(this, course, about, natural)
course	(next, is)

This	1	0	0	0	0	0
is	0	1	0	0	0	0
course	0	0	1	0	0	0
about	0	0	0	1	0	0
natural	0	0	0	0	1	0
language	0	0	0	0	0	1
next						
big						
data						





Word Embedding matrix



window-size=1

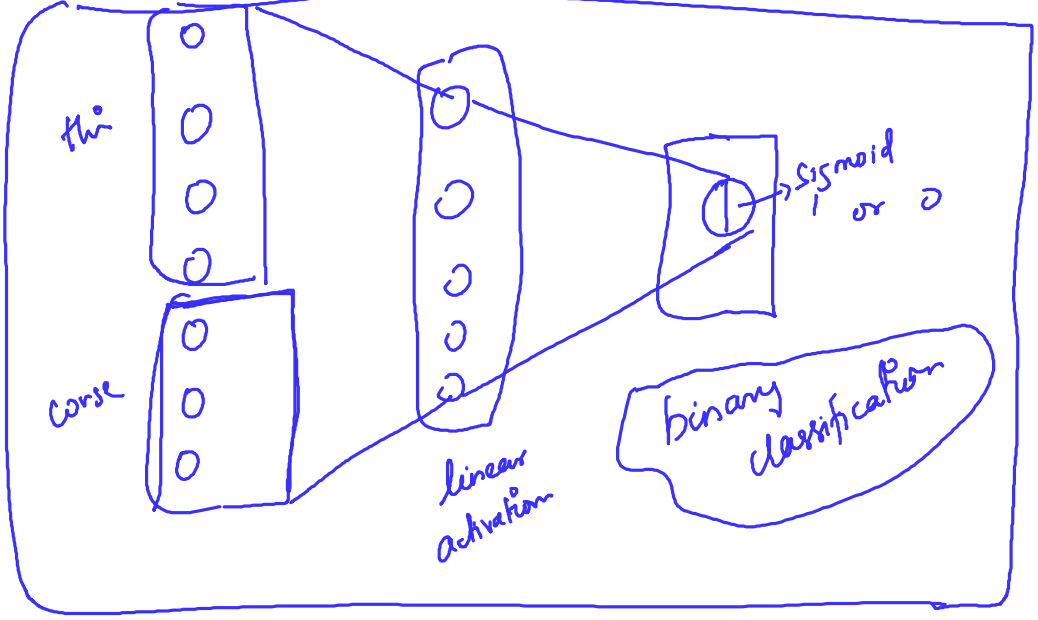
Center word	Context word
this	course
course	(this, is)
is	this, math

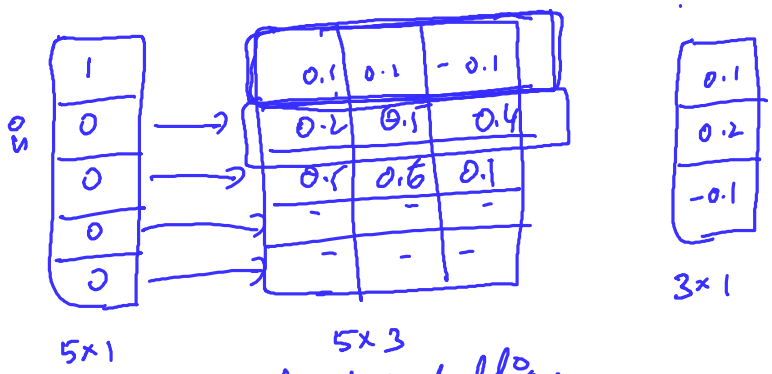
negative sampling

Center word	Context word	
this	course	positive pair
this	math	→ negative pair
this	natural	→ negative pair
coove	this	→ positive pair
courie	is	→ positive pair
coru	math	→ negative pair

Word2vec

Skipgram with negative sampling





Vocab [this, is, natural, language, processing]

This	1	0	0	0	0
is	0	1	0	0	0
:					
Processing	0	0	0	0	1

2003 → NNLM

2013 → word2vec

RNN

- LSTM — 1997 — Complex
- GRU — 2014 — Simpler

→ [This, course, is, on, NLP] Vocab

→ [1, 2, 100, 5, 300] this — 1

course — 2

---

① Tokenization

② Pad sequences to make size of all document equal

Doc 1 : [This is about nlp]

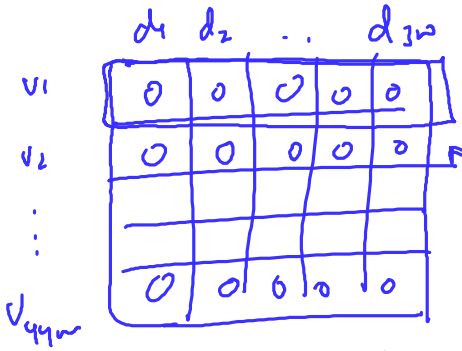
⇒ [1, 100, 300, 5, 0, 0, 0, 0, 0, 0]

Doc 2 : [This course teaches nlp basics and can ...]

⇒ [1, 50, 350, ..., 0, 0, 0]

doc ⇒ 50

Vocab : {a, the, movie, good, bad, ...} → 44,000



embedding ['a'] : [ ] 300

embedding ['the'] : [ ]

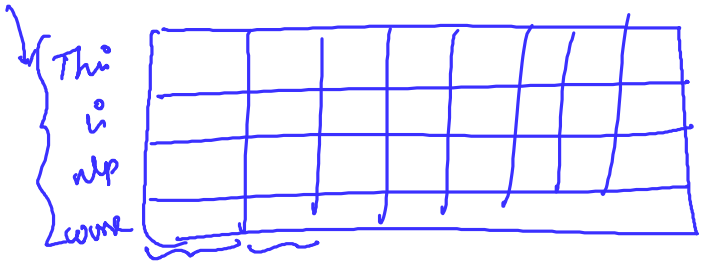
Word embedding matrix

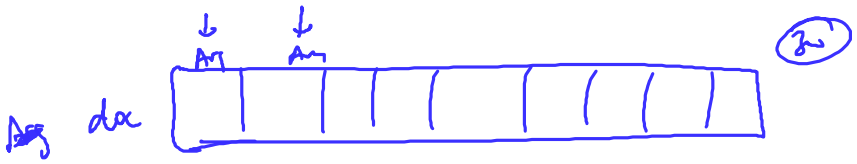
44000 × 300

Zero vector × 300

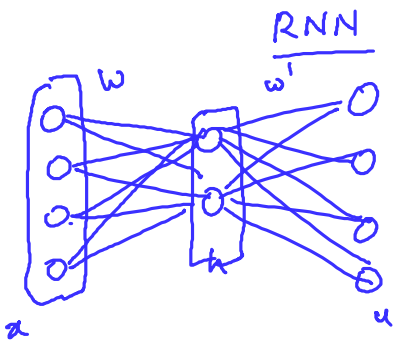
---

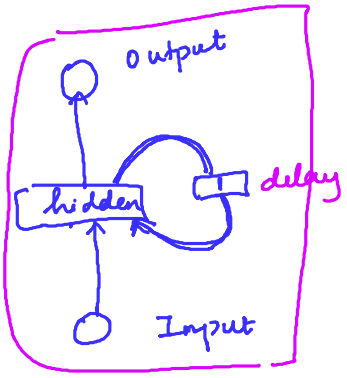
doc ⇒ [This a nlp course]



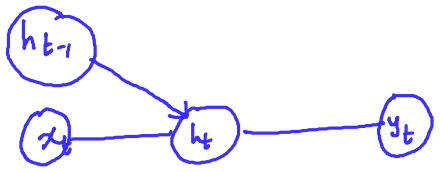


RNN



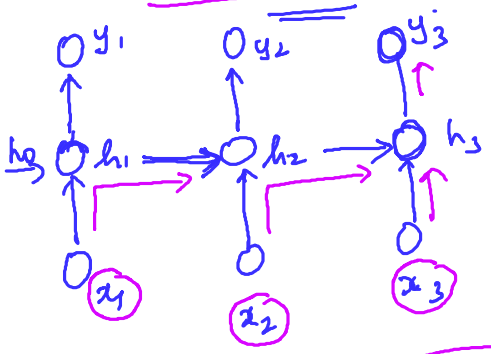


Forward pass $h = w^T x$ $\quad u = w'^T h$

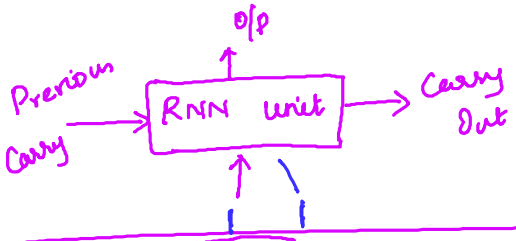


$$h_t = f(x_t, \overset{w}{\theta})$$

Unfolded version



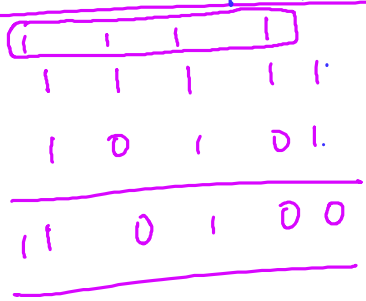
$$y_t = \theta y_{t-1} + \text{error}$$



$$h_1 = f(x_1, h_0)$$

$$h_2 = f(x_2, h_1)$$

$$h_2 = f(x_2, f(x_1, h_0))$$



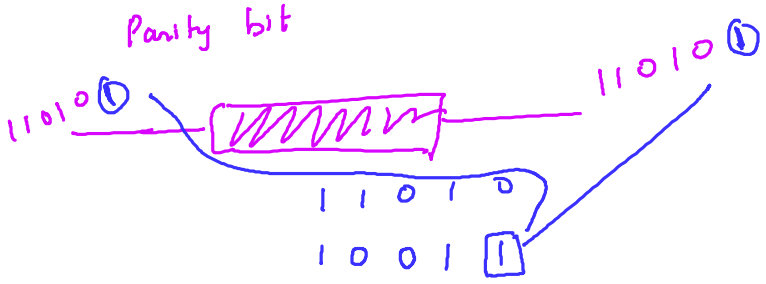
$$h_3 = f(x_3, h_2)$$

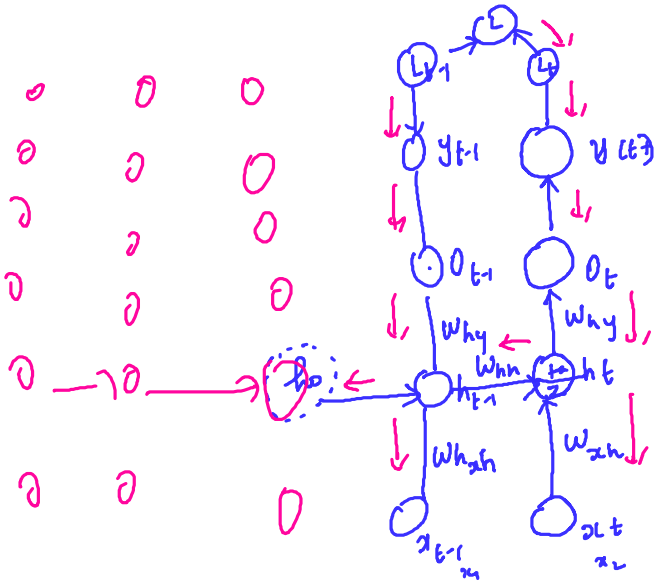
$$= f(x_3, f(x_2, h_1))$$

$$h_3 = f(x_3, f(x_2, f(x_1, h_0)))$$

$$y_n \sim f(x_n, x_{n-1}, x_{n-2}, \ldots x_1)$$

Parity bit





$\theta = (W_{xh}, W_{hy}, W_{hh}, y)$

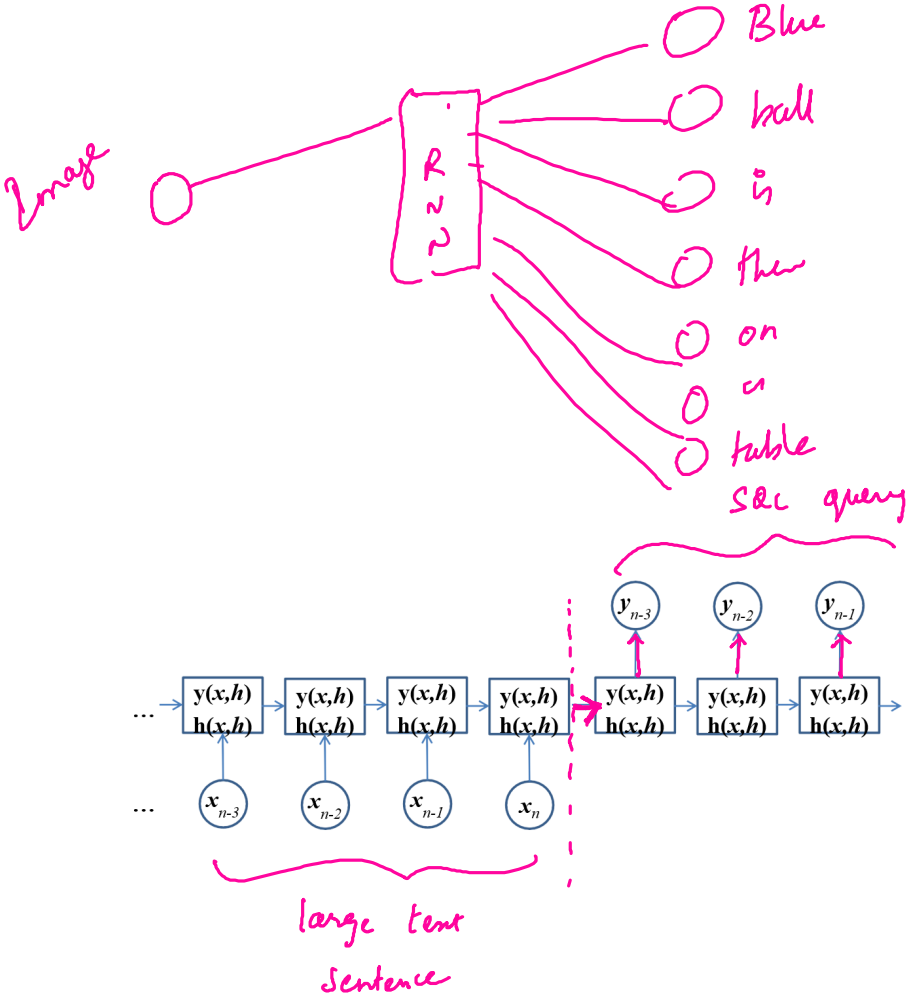
$$z_t = W_{xh} x + W^{hh} h_{t-1}$$

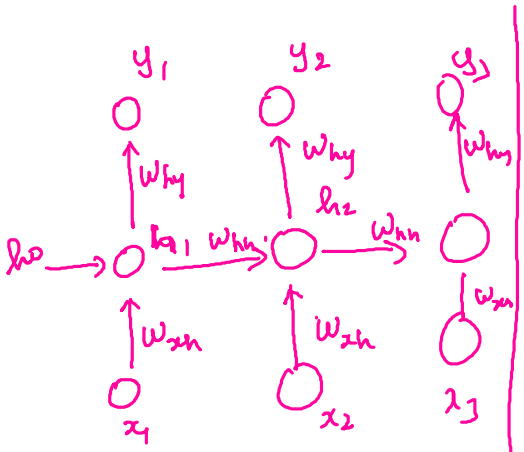
$$h_t = \tan(z_t)$$

$$o_t = W_{hy} h_t$$

$$y_t = \text{softmax}(o_t)$$

Backward propagation through time (BPTT)





$y_3 = f(x_2, h_2)$