Plan

Recap

Reminders

Markov and n-gram

Recurrent Networks

↳ Architecture

↳ Loss

↳ Optimization

Improvements

Recap

Object Detection

1) grad CAM

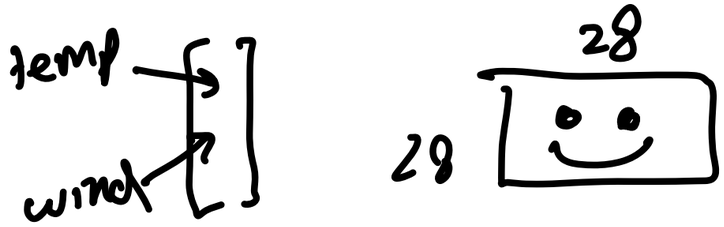
2) bounding boxes ← most popular

3) segmentation mask

Reminders

- Form 20/26
- Homework
  - ↳ 13.5 and above ☺
  - ↳ 12 to 13 :)
  - ↳ below :(talk to me!
  - ↳ Future submissions in LaTeX
- Project Proposal due Monday

Text Applications



- document retrieval
- convert audio to text
- translate between languages
- next word prediction

1) How do we represent words as numbers?

↳ assign numbers to words/letters

problems: distance

↳ one hot encoding

[ ] ← How [0 ⋮ 1 ⋮ 0] ← do

2) Varying length and long range dependency

"I am from France. ... I am going back home to ___."

Assume we've solved 1)

and we have $w \rightarrow$ meaningful vector encoding

$S = (w^1, w^2, w^3, \ldots, w^T)$ ← length of sequence

what is $P(S)$?

$= P(w^1) \cdot P(w^2 | w^1) \cdot P(w^3 | w^2, w^1) \cdot \ldots$

$P(w^2 | w^1) = \frac{P(w^2, w^1)}{P(w^1)}$

2-gram (word 1, word 2)

V^2

$P(w^T | w^{T-1}, w^{T-2}, \ldots, w^2)$

n-gram (word 1, word 2, ..., word n)

V^n

$P(1)$

$= P(w^1) \cdot P(w^2|w^1) \cdot P(w^3|w^2,w^1) \cdots$

Markov

$\approx P(w^1) \cdot P(w^2|w^1) \cdot P(w^3|w^2) \cdots$

2-gram

Predicting next word

When Gamaliel Painter died, he was Middlebury's pride,
A sturdy pioneer without a stain;
And he left his all by will, to the college on the hill,
And included his codicil cane.
Oh, its rap rap rap, and it's tap tap tap,
If you listen you can hear it sounding plain;
For a helper true and tried, as the generations glide,
There is nothing like Gamaliel Painter's cane.[5][6]

What comes after died?

w'

$P(w|w') = \frac{P(w' \text{ and } w)}{P(w')}$

2 grams with he

(he, was) $P(\text{was}|\text{he}) = \frac{P(\text{he was})}{P(\text{he})}$ 1, 2

(rap, rap) $P(\text{rap}|\text{rap}) = \frac{1}{2}$ 1

- v^2 small
- lose long range dependency
- more info
- v^n exponentially big

small — n-gram — big

$$P(w^t \mid w^{t-1}, w^{t-2}, \dots, w^1)$$

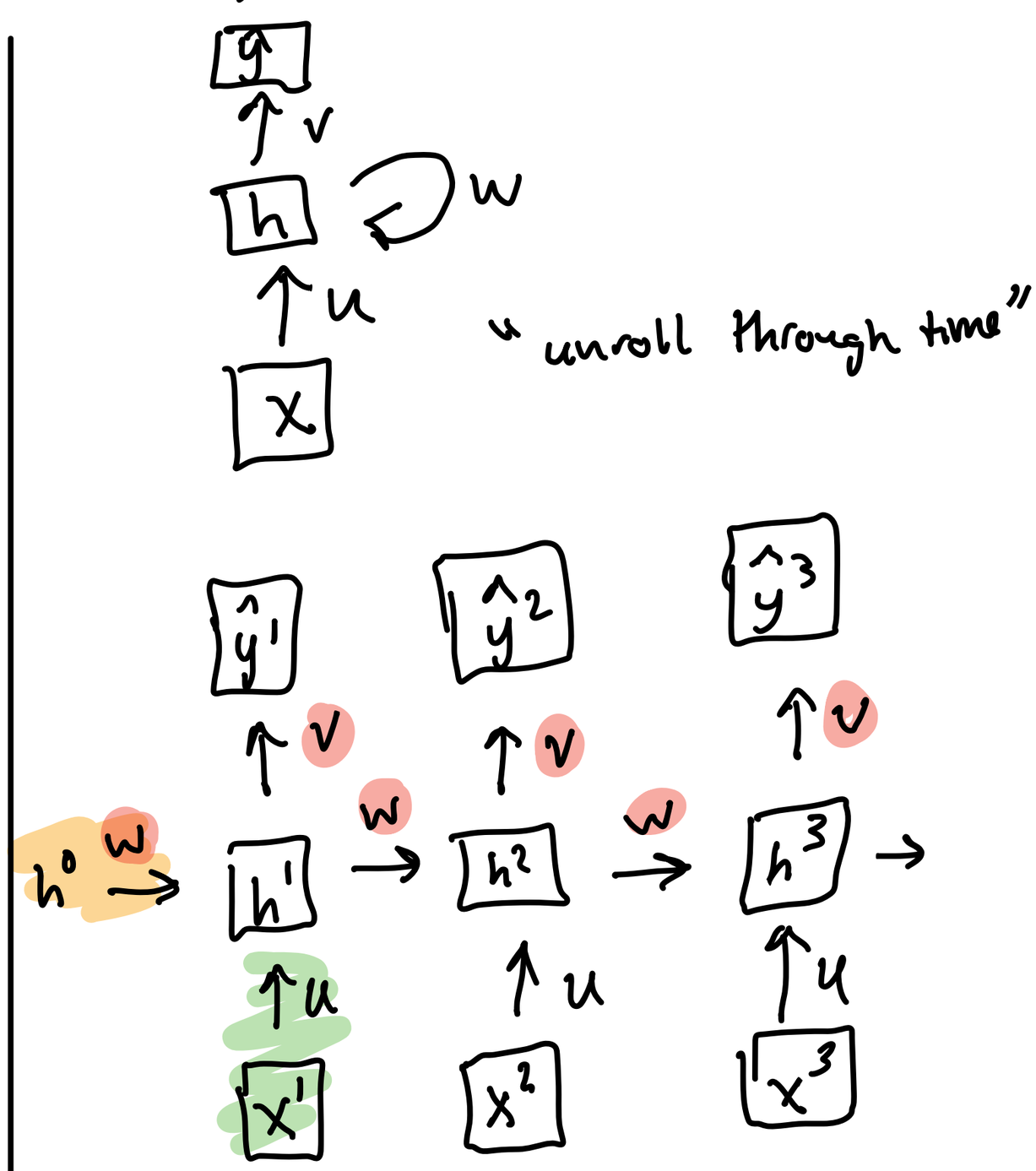
$$\approx f(w^{t-1}, h^{t-1})$$

h^{t-1} encodes information $w^{t-1}, \dots, w^1$

$h \in \mathbb{R}^d$

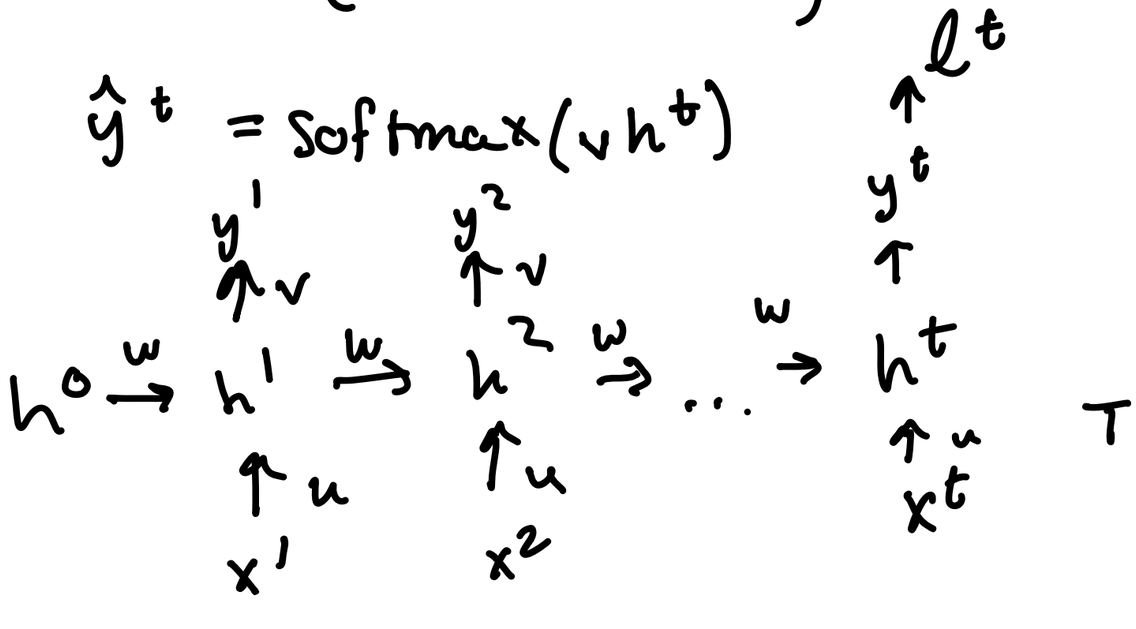
$$h^t = \sigma(\underset{d\times l \;\; l\times 1}{Ux^t} + \underset{d\times d \;\; d\times 1}{Wh^{t-1}})$$

$$\hat{y}^t = \text{softmax}(Vh^t)$$



$$h^t = \sigma(u x^t + w h^{t-1})$$

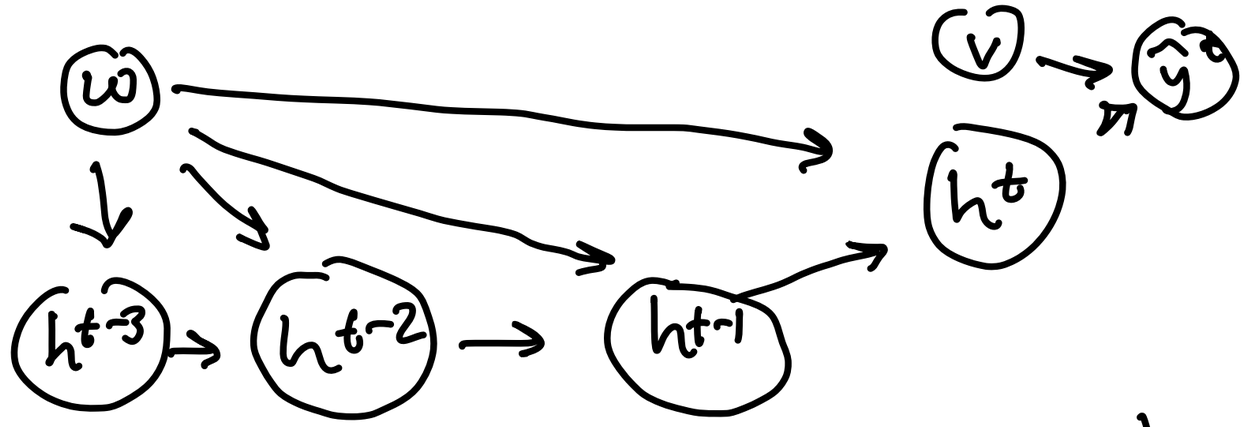
$$\hat{y}^t = \text{Softmax}(v h^t)$$



$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial}{\partial w}\left(\frac{1}{T}\sum_{i=1}^{T} l^t\right)$$

$$= \frac{1}{T}\sum_{i=1}^{T} \frac{\partial l^t}{\partial w}$$

$$\frac{\partial l^t}{\partial y^t} \cdot \frac{\partial y^t}{\partial h^t} \cdot \frac{\partial h^t}{\partial w}$$



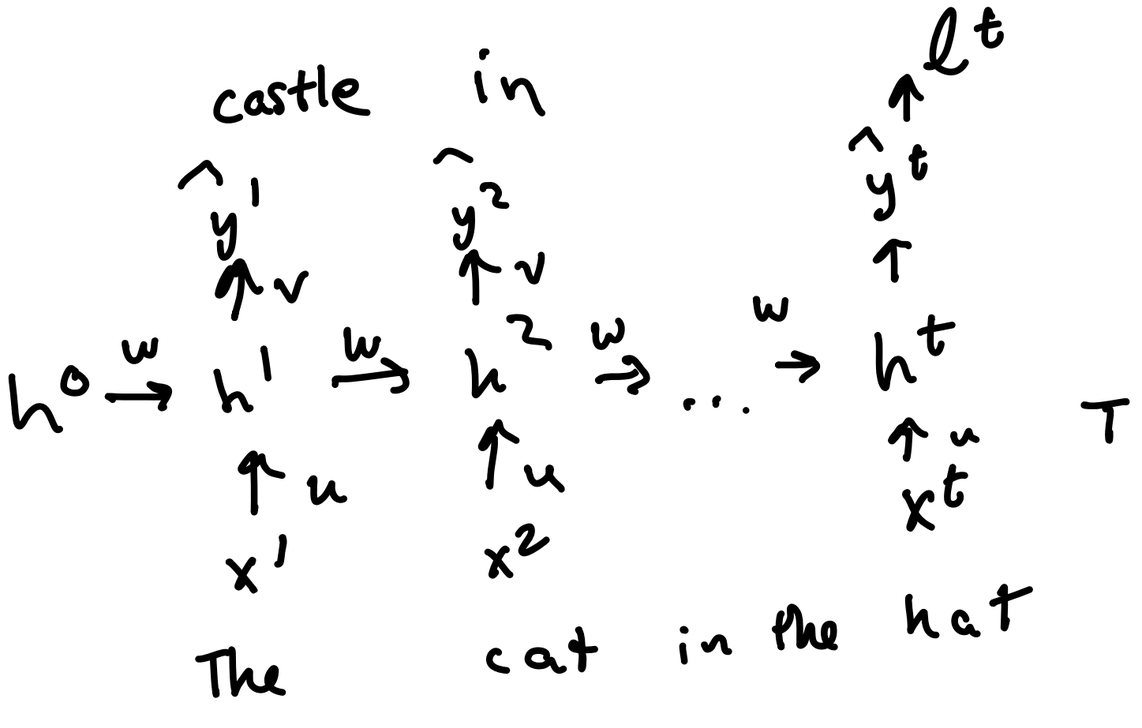
$$\frac{\partial h^t}{\partial w} = \frac{\partial h^t}{\partial w} + \frac{\partial h^t}{\partial h^{t-1}} \cdot \frac{\partial h^{t-1}}{\partial w}$$

$$+ \frac{\partial h^t}{\partial h^{t-1}} \cdot \frac{\partial h^{t-1}}{\partial h^{t-2}} \cdot \frac{\partial h^{t-2}}{\partial w} + \ldots$$

$= a > 1$

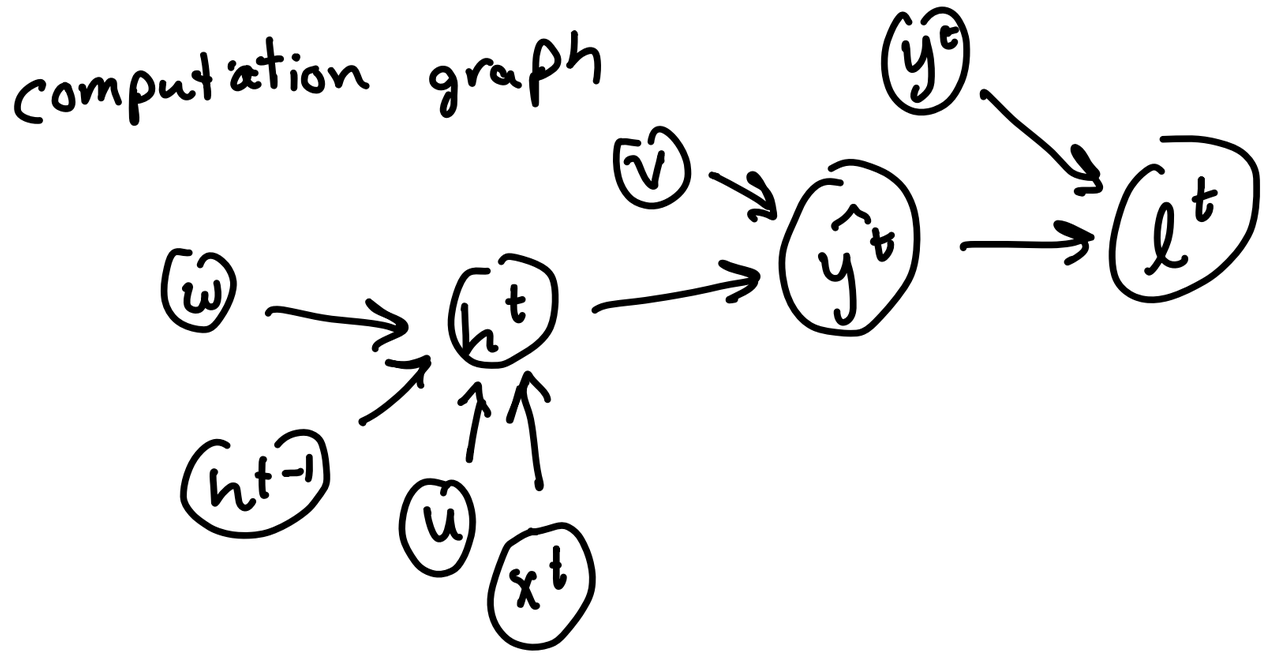
Solutions

- truncate (only last n partial derivatives)
- $\nabla' \mathcal{L}(w) \leftarrow \frac{\nabla \mathcal{L}(w)}{\|\nabla \mathcal{L}(w)\|}$ scaling
- more complicated architectures

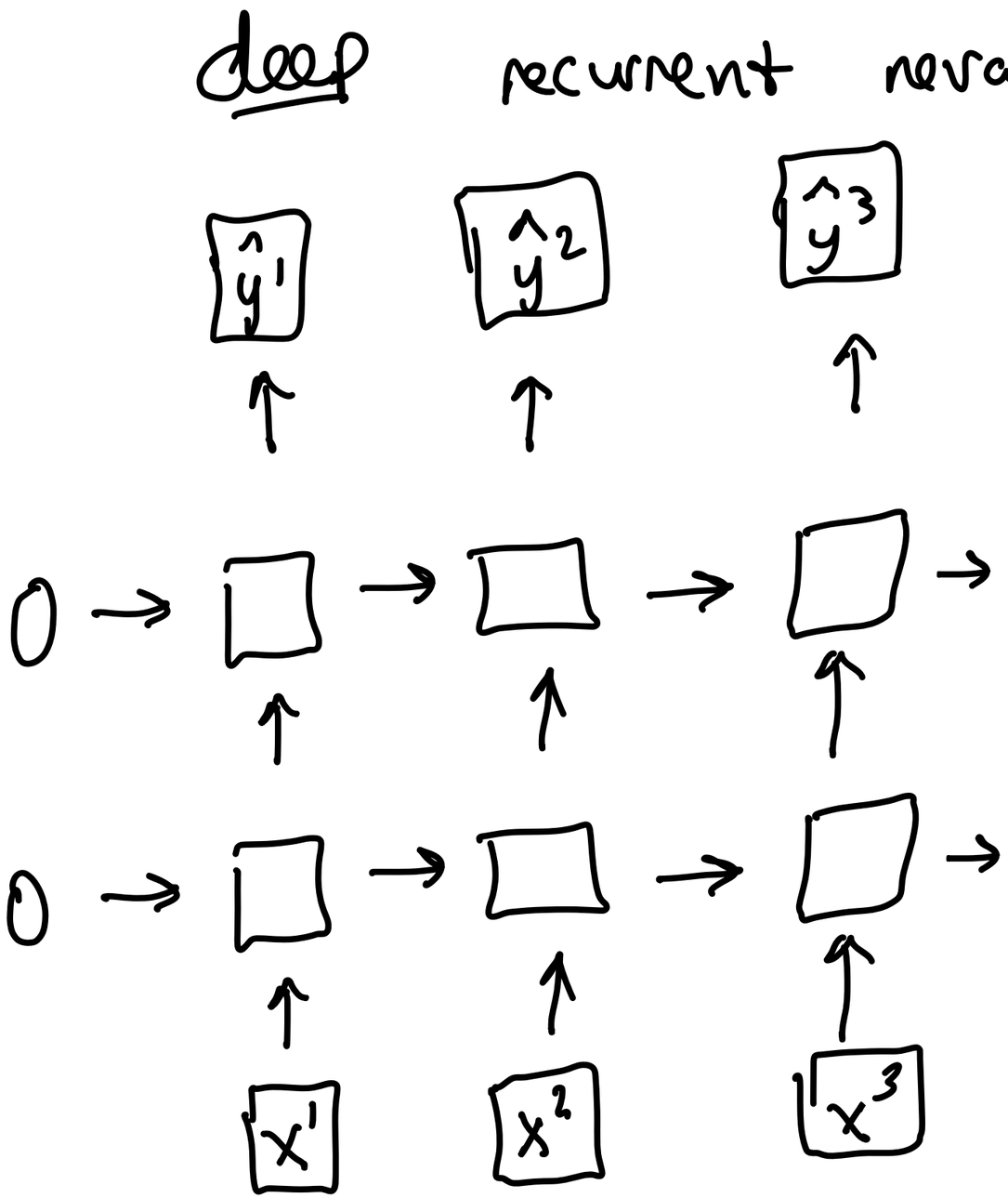


$$h^t = \sigma(u x^t + w h^{t-1})$$

$$\hat{y}^t = \text{Softmax}(v h^t)$$



deep recurrent neural net



Long Short Term Memory