

Recurrent Neural Networks

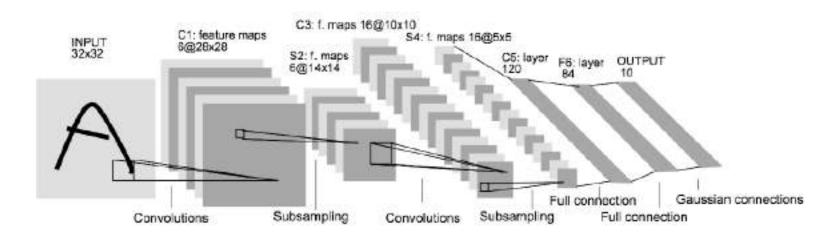
모두의연구소 박은수 Research Director



- 분류기의 구성
 - Score function
 - Loss function
 - Optimization



Convolutional Neural Networks



지난시간 돌아보기



Momentum

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \, \frac{\partial L}{\partial \mathbf{W}}$$
$$\mathbf{W} \leftarrow \mathbf{W} + \mathbf{v}$$

Gradient 이동누적 스러운 방법

업데이트 1)
$$v_1 \leftarrow \alpha * 0 - K_o$$
 : $-K_o$ * 1) $\mathbf{W} \leftarrow \mathbf{W} + v_1$ 업데이트 2) $v_2 \leftarrow \alpha v_1 - K_1$: $-\alpha K_o - K_1$ * 2) $\mathbf{W} \leftarrow \mathbf{W} + v_2$ 업데이트 3) $v_3 \leftarrow \alpha v_2 - K_2$: $-\alpha^2 K_o - \alpha K_1 - K_2$ * 3) $\mathbf{W} \leftarrow \mathbf{W} + v_3$ 업데이트 4) $v_4 \leftarrow \alpha v_3 - K_3$: $-\alpha^3 K_o - \alpha^2 K_1 - \alpha K_2 - K_3$ * 4) $\mathbf{W} \leftarrow \mathbf{W} + v_4$

Adagrad

$$\mathbf{h} \leftarrow \mathbf{h} + \frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}}$$
$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

업데이트 1)
$$\frac{1}{K_0}K_0$$
 업데이트 2) $\frac{1}{K_1^2+K_0^2}K_1$ 업데이트 3) $\frac{1}{K_3^2+K_1^2+K_0^2}K_3$ 업데이트 4) $\frac{1}{K_4^2+K_3^2+K_1^2+K_0^2}K_4$

Gradient Normalization 스러운 방법

1) W ← W + v_i

4) W ← W + v₄

RMSprop

$$\mathbf{h} \leftarrow \alpha \mathbf{h} + (1 - \alpha) \left(\frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}} \right)$$
$$\mathbf{W} - \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

업데이트 1)
$$\mathbf{h}_1 = (1-a)\mathbf{K}_1^2$$

업데이트 2) $\mathbf{h}_2 = a(1-a)\mathbf{K}_1^2 + (1-a)\mathbf{K}_2^2$
업데이트 3) $\mathbf{h}_3 = a^2(1-a)\mathbf{K}_1^2 + a(1-a)\mathbf{K}_2^2 + (1-a)\mathbf{K}_3^2$
업데이트 4) $\mathbf{h}_4 = a^3(1-a)\mathbf{K}_1^2 + a^2(1-a)\mathbf{K}_2^2 + a(1-a)\mathbf{K}_3^2 + (1-a)\mathbf{K}_4^2$



Momentum

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \eta \, \frac{\partial L}{\partial \mathbf{W}}$$
$$\mathbf{W} \leftarrow \mathbf{W} + \mathbf{v}$$

업데이트 1)
$$\mathbf{v}_1 \leftarrow \alpha * 0 - K_o : -K_o$$
 업데이트 2) $\mathbf{v}_2 \leftarrow \alpha \mathbf{v}_1 - K_1 : -\alpha K_o - K_1$ 업데이트 3) $\mathbf{v}_3 \leftarrow \alpha \mathbf{v}_2 - K_2 : -\alpha^2 K_o - \alpha K_1 - K_2$ 업데이트 4) $\mathbf{v}_4 \leftarrow \alpha \mathbf{v}_3 - K_3 : -\alpha^3 K_o - \alpha^2 K_1 - \alpha K_2 - K_3$

Adagrad

$$\mathbf{h} \leftarrow \mathbf{h} + \frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

업데이트 1) $\frac{1}{\sqrt{K_0^2}} K_0$ 업데이트 2) $\frac{1}{\sqrt{K_1^2 + K_0^2}} K_1$ 업데이트 3) $\frac{1}{\sqrt{K_3^2 + K_1^2 + K_0^2}} K_3$

두 방법의 같이쓰자 Adam

. 3) W ← W + va

4) W ← W + v₄

RMSprop

$$\mathbf{h} \leftarrow \alpha \mathbf{h} + (1 - \alpha) \left(\frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}} \right)$$
$$\mathbf{W} - \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

업데이트 4)
$$\frac{1}{K_4^2 + K_3^2 + K_1^2 + K_0^2} K_4$$

업데이트 1)
$$\mathbf{h}_1 = (1-a)\mathbf{K}_1^2$$

업데이트 2) $\mathbf{h}_2 = a(1-a)\mathbf{K}_1^2 + (1-a)\mathbf{K}_2^2$
업데이트 3) $\mathbf{h}_3 = a^2(1-a)\mathbf{K}_1^2 + a(1-a)\mathbf{K}_2^2 + (1-a)\mathbf{K}_3^2$
업데이트 4) $\mathbf{h}_4 = a^3(1-a)\mathbf{K}_1^2 + a^2(1-a)\mathbf{K}_2^2 + a(1-a)\mathbf{K}_3^2 + (1-a)\mathbf{K}_4^2$



Batch Normalization

[loffe and Szegedy, 2015]

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$; Parameters to be learned: γ , β

Output: $\{y_i = BN_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i$$
 // mini-batch mean

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2$$
 // mini-batch variance

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$
 // normalize

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i)$$
 // scale and shift

Note: at test time BatchNorm layer functions differently:

The mean/std are not computed based on the batch. Instead, a single fixed empirical mean of activations during training is used.

(e.g. can be estimated during training with running averages)

Fei-Fei Li & Justin Johnson & Serena Yeung

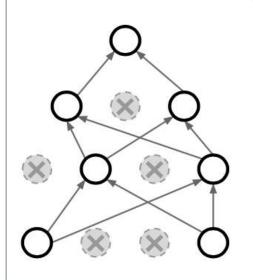
Lecture 6 - 60

April 20, 2017



Regularization: Dropout

How can this possibly be a good idea?



Another interpretation:

Dropout is training a large **ensemble** of models (that share parameters).

Each binary mask is one model

An FC layer with 4096 units has $2^{4096} \sim 10^{1233}$ possible masks! Only $\sim 10^{82}$ atoms in the universe...

Fei-Fei Li & Justin Johnson & Serena Yeung

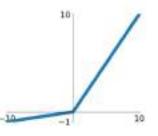
Lecture 7 - 63

April 25, 2017

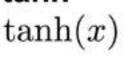
Activation Functions

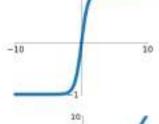
Sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$

Leaky ReLU max(0.1x, x)



tanh



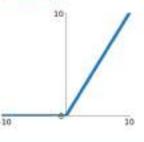


Maxout

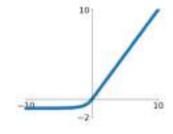
 $\max(w_1^T x + b_1, w_2^T x + b_2)$

ReLU

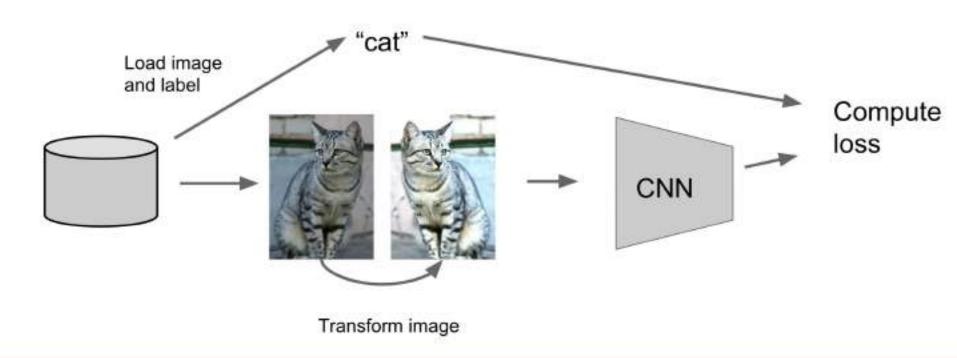
 $\max(0,x)$



ELU $\begin{cases} x & x \ge 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$



Regularization: Data Augmentation



Transfer Learning with CNNs

Train on Imagenet

FC-1000 FC-4096 FC-4096 MaxPool Conv-512 Conv-512 MaxPool Conv-512 Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 Conv-128 MaxPool Conv-64 Conv-64

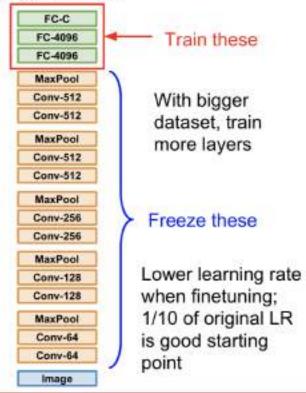
Image

2. Small Dataset (C classes)



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014 Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

Bigger dataset





가 나 다 라 마 () ()



가나다라마사아

ABCDEF()()() ~~~



가 나 다 라 마 사 아

ABCDEFGHI~~~



하 파 타 카 차 자 () () ~~





하 파 타 카 차 자 (아) (사) ~~





하 파 타 카 차 자 (아) (사) ~~

QPONMLKJ()()()~~





하 파 타 카 차 자 (아) (사) ~~

 $QPONMLKJ(I)(H)(G)\sim\sim$



로꾸꺼 가사 ...



(모두) 로꾸거 로꾸거 로꾸거 말해말 로꾸거 로꾸거 로꾸거 말해말

(희철) 아 많다 많다 많다 많다 다 이쁜 이쁜 이쁜이다 여보게 저기 저게보여

(신동) 여보 안경 안보여

(강인) 통술집 술통 소주 만병만 주소 다 이심전삼이다 뻑뻑뻑 아 좋다좋아 수박이 박수

(희철) 다시 합창합시다

(모두) 로꾸거 로꾸거 로꾸거 말해말 로꾸거 로꾸거 로꾸거 말해말

(이특) 니 가는데는 가니 일요일 수이수 수리수리수 몰랑몰랑몰 아 좋다좋아 수박이 박수 다시 한창 한시다 성민) 어제는 거꾸로 오늘도 거꾸로 모두가 거꾸로 돌아가고 있어 내일이 와야해 행복의 시계가 째깍째깍 돌아가겠지

(모두) 째깍 째깍 째깍 원투쓰리포파이브식스 고 로꾸거 로꾸거 로꾸거 말해말 로꾸거 로꾸거 로꾸거 말해맠

(은혁) 하파타카차자아사바마라다나가

(신동) 십구팔칠육오사삼이일땡.

(은혁) 아래서 위로 뒤에서 앞으로

(신동) 모든걸 거꾸로 로꾸거

(신동&은혁) 할아버지 할머니 아저씨 아줌마 남녀노소 짠짠짠 얼씨구 절씨구 빠라빠라 빰빰 모든걸 거꾸로 로꾸거 (이특) 나갔다오나 나오다갔나 아들 딸이 다컸다 이 딸들아

(성민) 다 같은 별은 별은 같다

(은혁) 자꾸만 꿈만 꾸자

(신동) 장가간 가장 시집간 집시 다 된 장국 청국장된다 아 좋다좋아 수박이 박수 다시 합창 합시다

(희철) 어제도 거꾸로 오늘도 거꾸로 모두가 거꾸로 돌아가고 있어 내일이 와야해 행복의 시계가 째깍째깍 돌아가겠지

(모두) 째깍 째깍 째깍 원투쓰리포파이브식스 고 로꾸거 로꾸거 로꾸거 말해말 로꾸거 로꾸거 로꾸거 말해말 아 좋다 좋아 수박이 박수 다시 합창 합시다 로꾸거 로꾸거 로꾸거 로꾸거 로꾸거



비슷한 사례



좋아하는 노래의 후렴구 바로 전을 떠 올려 보세요

비슷한 사례



좋아하는 노래의 후렴구 바로 전을 떠 올려 보세요

혹시 처음부터 시작해서 후렴구 바로 전을 찾지 않으셨나요?



하 파 타 카 차 자 (아) (사) ~~

 $QPONMLKJ(I)(H)(G)\sim\sim$

혹시 처음부터 시작해서 후렴구 바로 전을 찾지 않으셨나요?

21



하 파 타 카 차 자 (아) (사) ~~

 $QPONMLKJ(I)(H)(G)\sim$

혹시 처음부터 시작해서 후렴구 바로 전을 찾지 않으셨나요?

인간은 기억할때 컴퓨터처럼 하드드라이 브에 저장하지 않기 때문입니다

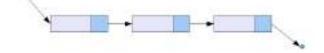


인간은 기억할때 컴퓨터처럼 하드드라이 브에 저장하지 않기 때문입니다

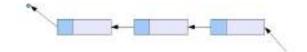


인간은 정보를 Sequence로 학습한다고 합니다

하 파 타 카 차 자 (아) (사) ~~



Q P O N M L K J (I) (H) (G)~~





그렇다고 우리가 A, B, C를 못 외우거나 노래 가사를 잊어 버리고 있는건 아니잖아요?



그렇다고 우리가 A, B, C를 못 외우거나 노래 가사를 잊어 버리고 있는건 아니잖아요?

넵, 단지 더 많은 시간이 걸리는 겁니다.

이런식으로 찾으려 시도한적이 없기 때문이지요 뇌에서 이 정보가 위차하는 곳으로 이어지는 어떤 지도가 딱 존재하지 않는 것 입니다



다니던 길을 찾아 가는 것은 싶게 상상할 수 있 습니다





다니던 길이 아닌 직선 의 길을 상상해서 가는 것은 쉽지 않죠





참고: Anyone Can Learn To Code an LSTM-RNN in Python (한글번역), 유재준. http://jaejunyoo.blogspot.com/2017/06/anyone-can-learn-to-code-LSTM-RNN-Python.html

자주 사용하는 동작은 무의식적으로 발현됨을 기억해 보세요~



우리가 사고하는 방향 혹은 Sequence를 조건부 확률처럼 뉴런이 학습되게 됩니다

또한 이전에 봐왔던 패턴 혹은 기억이 우리의 판단결과에 영향을 미치게 됩니다

참고: Anyone Can Learn To Code an LSTM-RNN in Python (한글번역), 유재준. http://jaejunyoo.blogspot.com/2017/06/anyone-can-learn-to-code-LSTM-RNN-Python.html

Sequence 형태의 데이터를 뉴럴 네 트워크가 학습할 수 있게 하는것?



?

Sequence 형태의 데이터를 뉴럴 네 트워크가 학습할 수 있게 하는것?



Recurrent Neural Networks

뉴럴 네트워크가 이전 정보를 반영하 는 방법



일반적인 NN: input -> hidden -> output

이전 정보를 반영하는 2가지 방법론

(input + **prev_hidden**)->hidden->output **VS**.

(input + prev_input)->hidden->output

Previous input

Previous hidden



Previous hidden

Previous input



Previous hidden

Previous input

```
(input + empty_input) -> hidden -> output
(input + prev_input) -> hidden -> output
(input + prev_input) -> hidden -> output
```



Previous hidden (input + empty_hidden) -> hidden -> output (input + prev_hidden) -> hidden -> output

Previous input (input + empty_input) -> hidden -> output (input + prev_input) -> hidden -> output



4단계 과정비교

Previous hidden

전부다 기억할 수 있음

Previous input

바로 직전만을 기억함

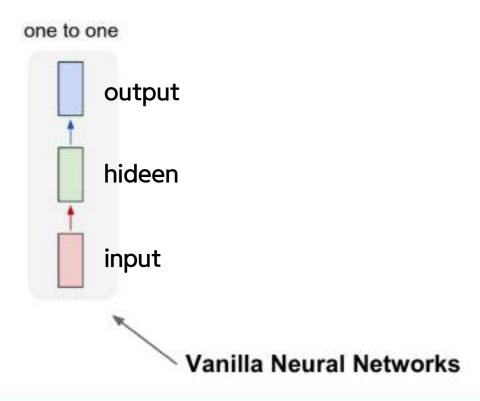
```
(input + empty_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
(input + prev_hidden) -> hidden -> output
```

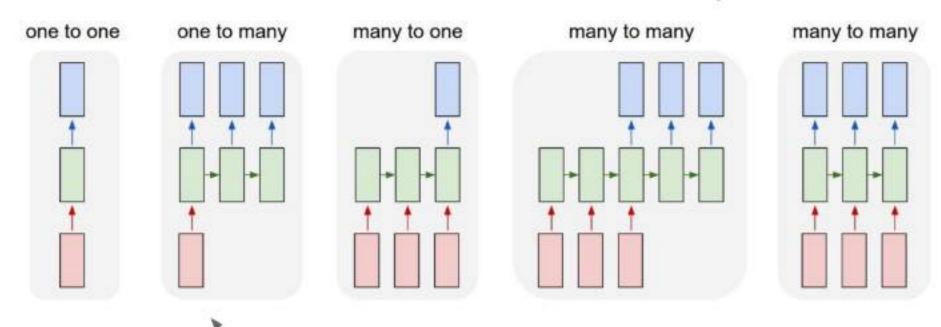
```
(input + empty_input) -> hidden -> output
(input + prev_input) -> hidden -> output
(input + prev_input) -> hidden -> output
(input + prev_input) -> hidden -> output
```



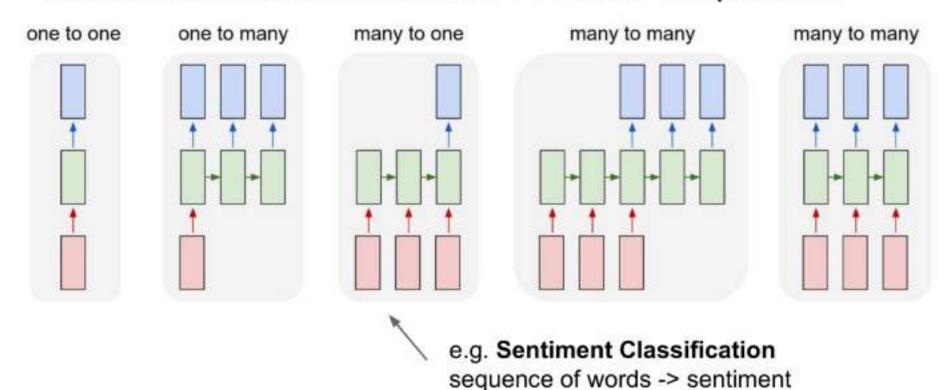
Hidden과 input의 조합으로 네트워크를 구성해 야 긴 sequence를 다룰 수 있겠군요

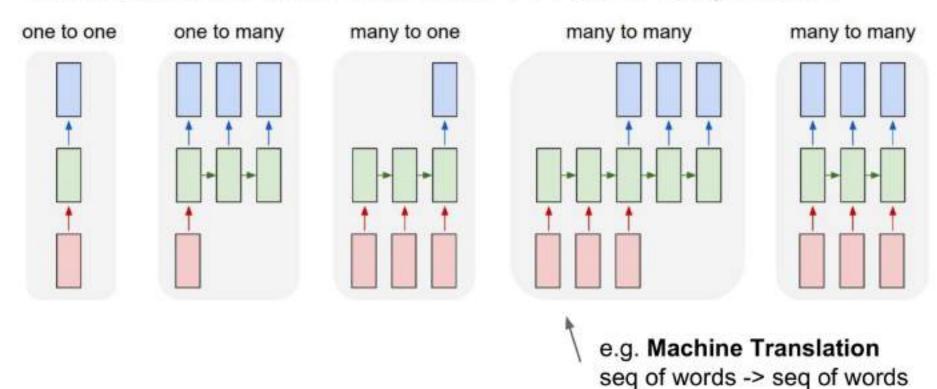
"Vanilla" Neural Network

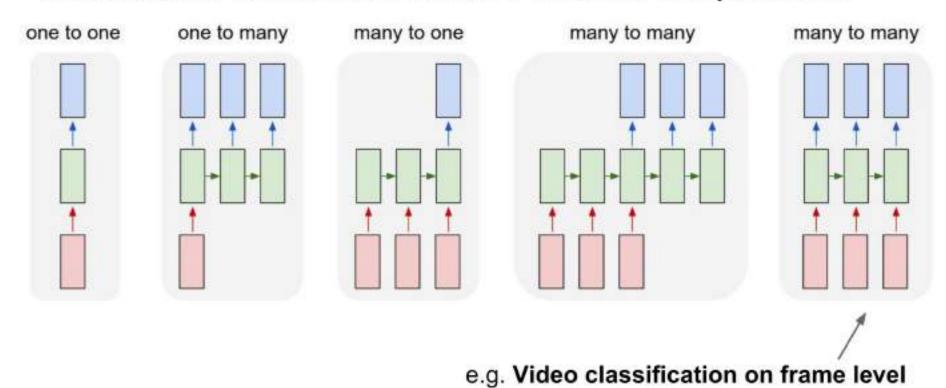


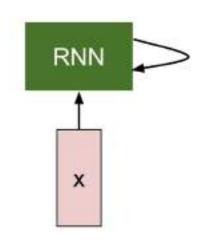


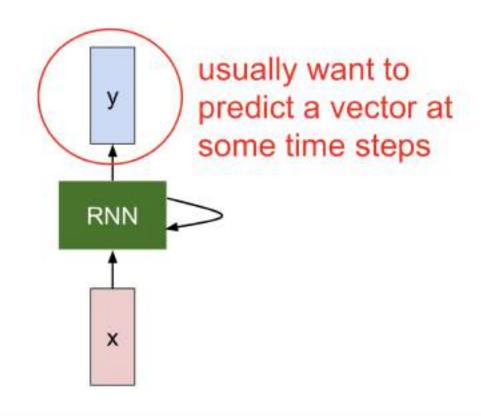
e.g. Image Captioning image -> sequence of words



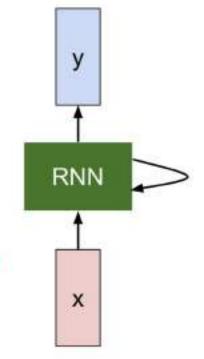








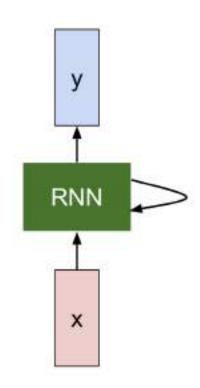
We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:



We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

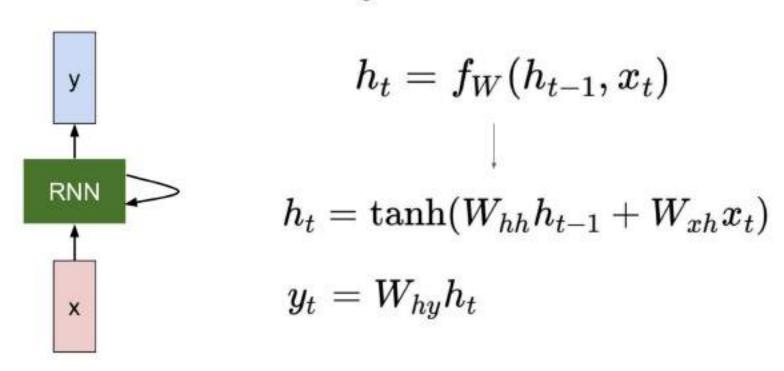
$$h_t = f_W(h_{t-1}, x_t)$$

Notice: the same function and the same set of parameters are used at every time step.



(Vanilla) Recurrent Neural Network

The state consists of a single "hidden" vector h:



그림으로 대략 그려보면

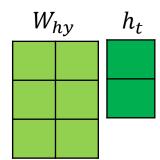


$$h_t = f_W(h_{t-1}, x_t)$$



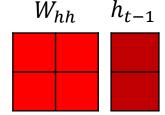
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

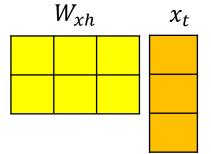
$$y_t = W_{hy}h_t$$

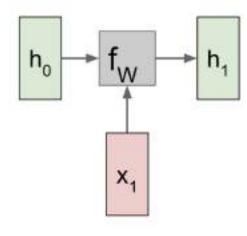


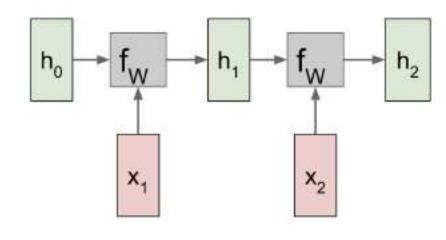
가정

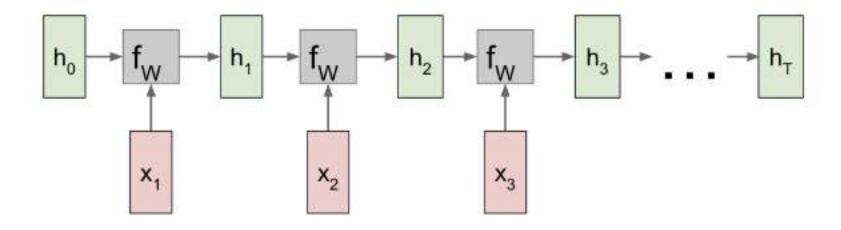
- Input: (3,1)
- Hidden: (2,1)
- Output: (3,1)



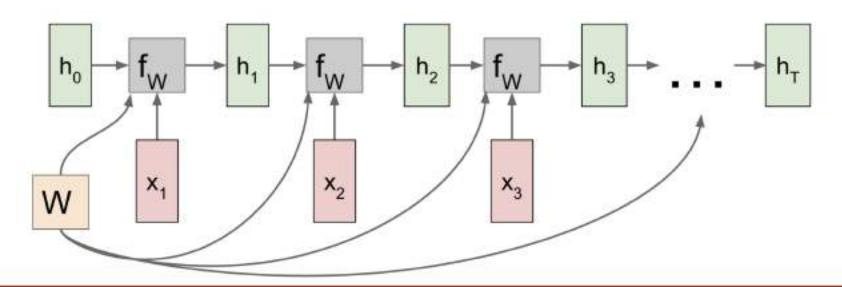




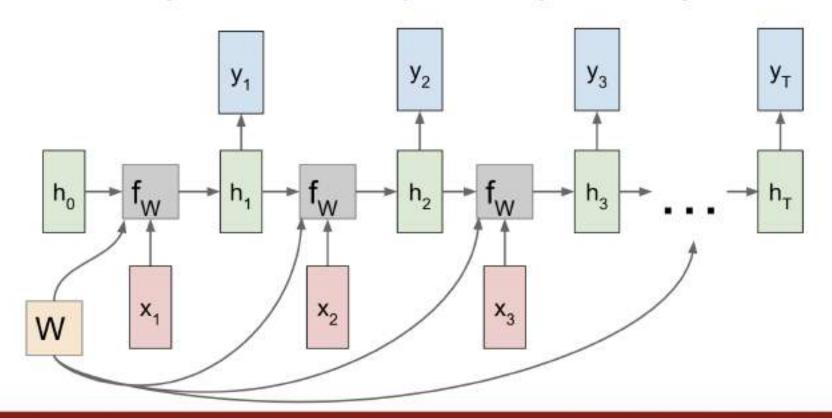




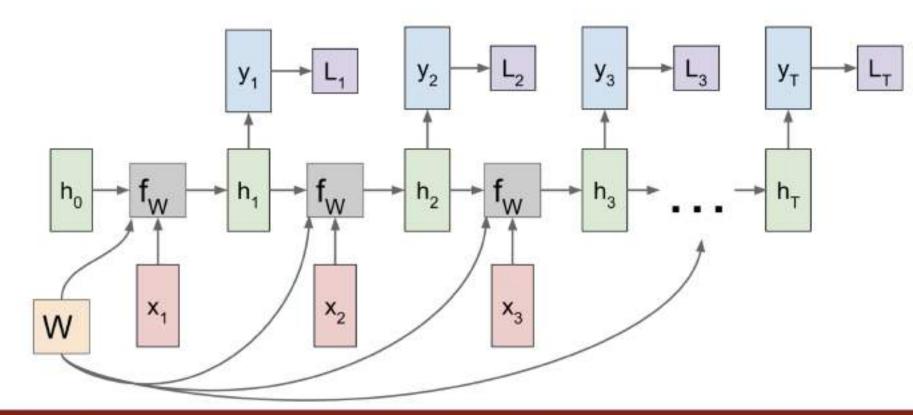
Re-use the same weight matrix at every time-step

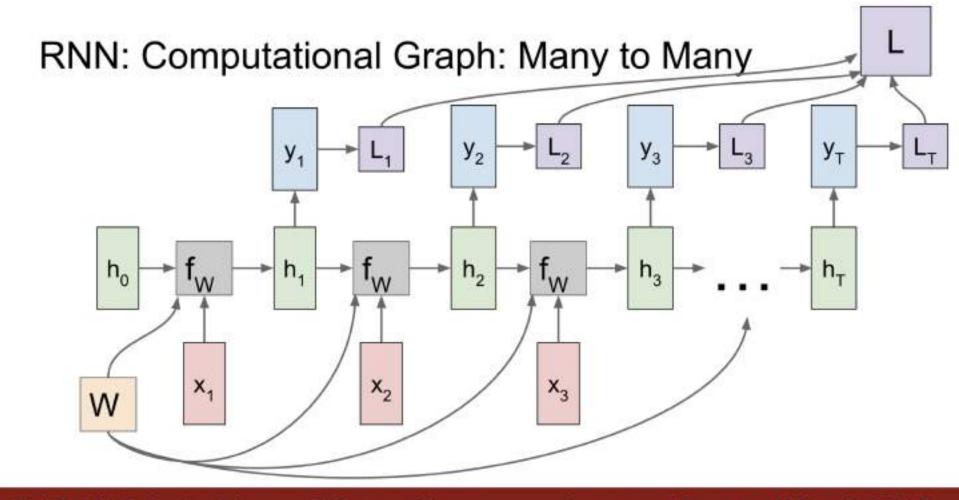


RNN: Computational Graph: Many to Many

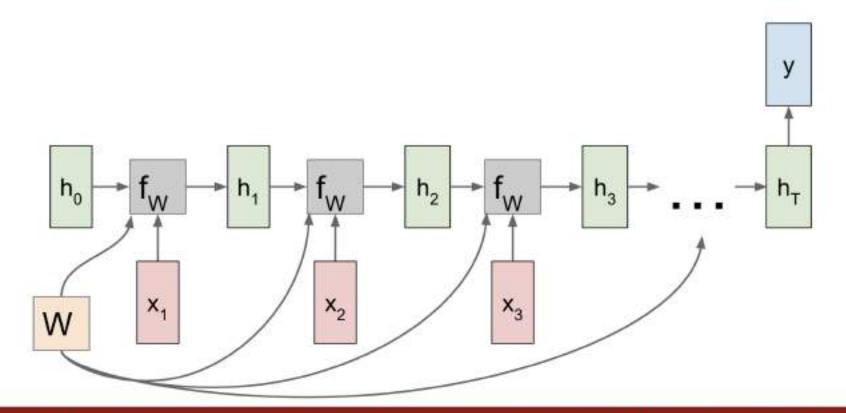


RNN: Computational Graph: Many to Many

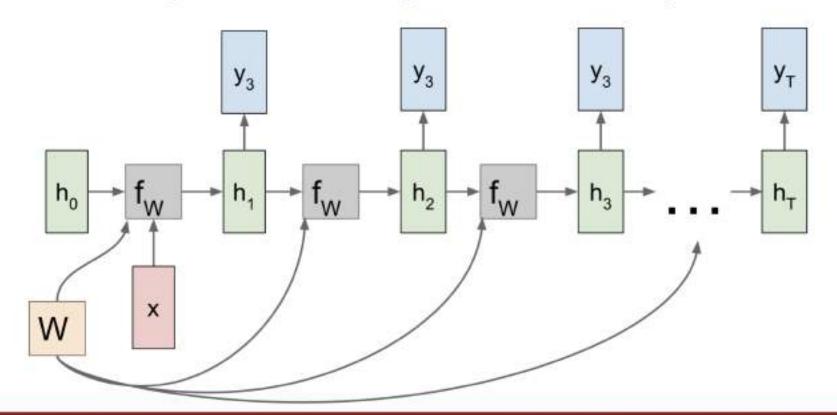




RNN: Computational Graph: Many to One

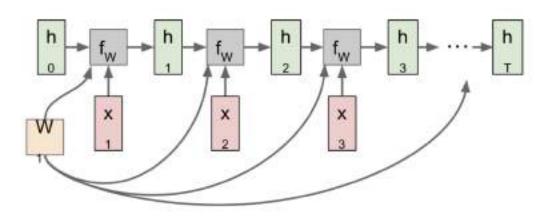


RNN: Computational Graph: One to Many

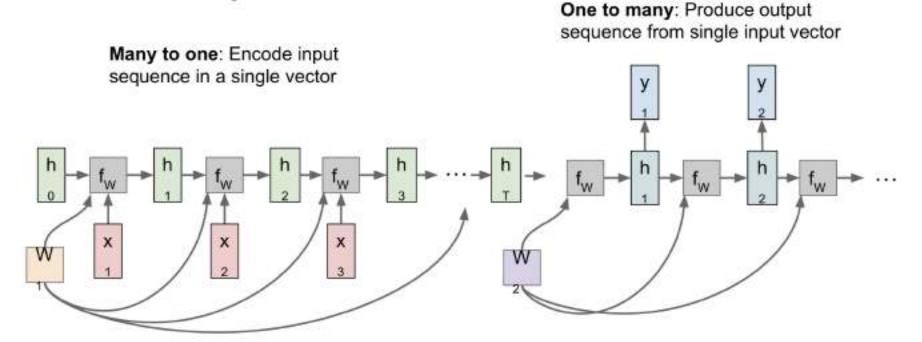


Sequence to Sequence: Many-to-one + one-to-many

Many to one: Encode input sequence in a single vector

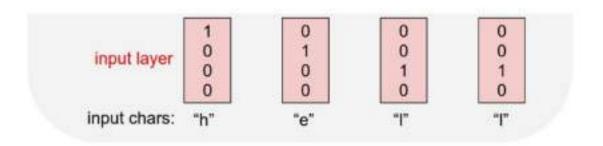


Sequence to Sequence: Many-to-one + one-to-many



Vocabulary: [h,e,l,o]

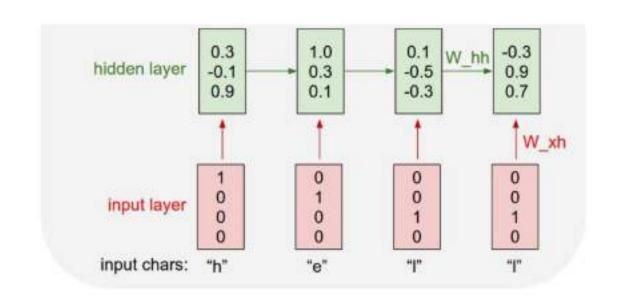
Example training sequence: "hello"



$$h_t = anh(W_{hh}h_{t-1} + W_{xh}x_t)$$

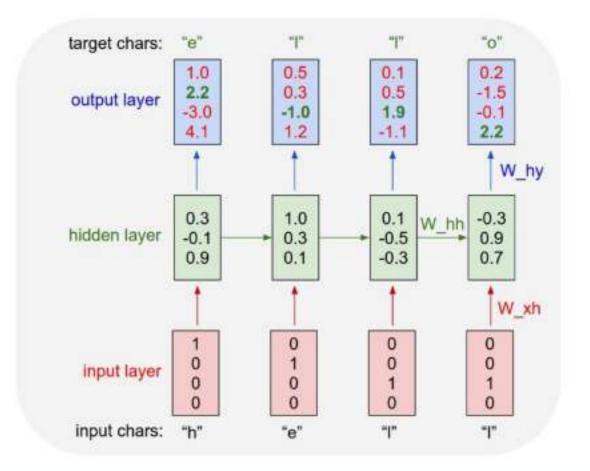
Vocabulary: [h,e,l,o]

Example training sequence: "hello"

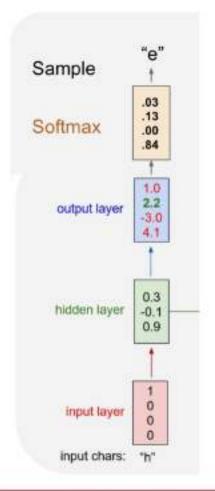


Vocabulary: [h,e,l,o]

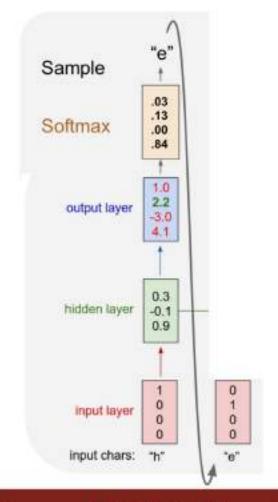
Example training sequence: "hello"



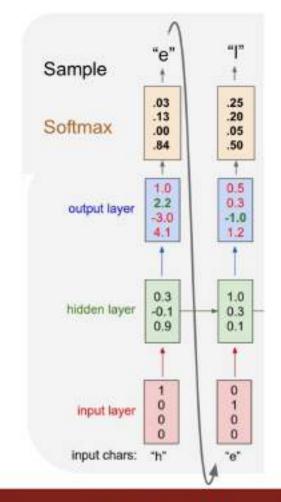
Vocabulary: [h,e,l,o]



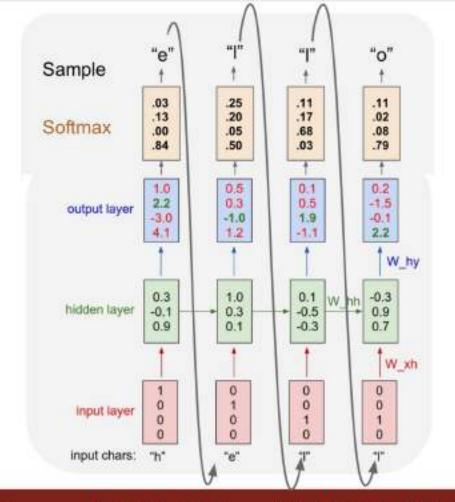
Vocabulary: [h,e,l,o]

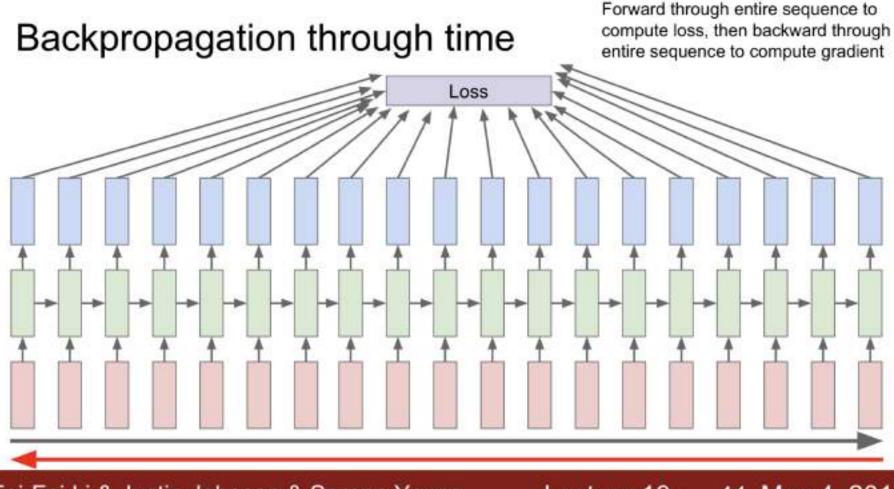


Vocabulary: [h,e,l,o]

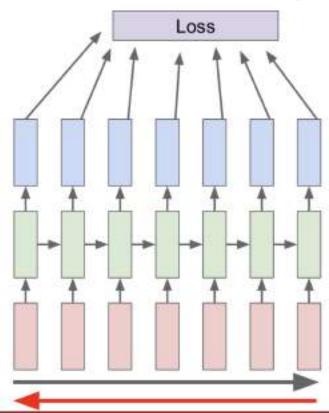


Vocabulary: [h,e,l,o]



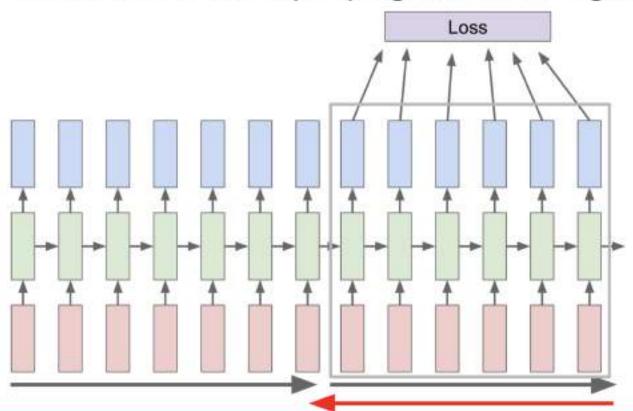


Truncated Backpropagation through time



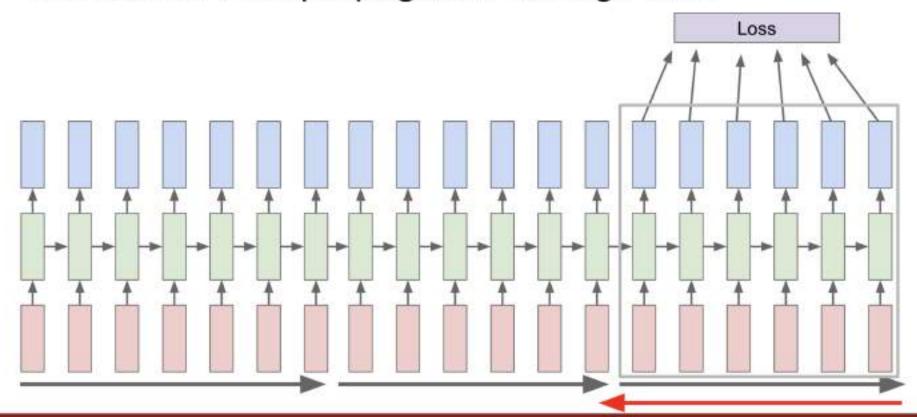
Run forward and backward through chunks of the sequence instead of whole sequence

Truncated Backpropagation through time



Carry hidden states forward in time forever, but only backpropagate for some smaller number of steps

Truncated Backpropagation through time



min-char-rnn.py gist: 112 lines of Python

```
NOW, THERE HAS BELLEVING MICH. ACTOR IN MICH. MICH.
                                     to come
                                     March 1994 41-34
                                 Make a series transfer of the Property of American Street, and the Amer
                                 many is management of
                                 min. dati, see him I believed. I transport
                                 still for he is because, it must be too one, was not
                                 minimum, and it is shown here is not the advancement of the
                                 mobilities in a planting from the selection making a
    - Indian and the last to be a resident to an information
                                 Assistance, Challe in 1971
                        with a proper sensor poor, year question of the con-
and it is a reason assessment of the least proof to the comment
- And I do make continues (4), history algorithm is continued in
            - And Assessed Sedents And Little Control States
    to the control of the same (1) to the con-
                                 and Automotivated Printed Street,
                                                 many course on our last of decades.
                                                 dense to ten urray of Interact dumber trusts.
                                                 returns for too. Administration to been parasisoned, and last hanner stead
                                                 per decide, as 4 fts, 10; 10; 11; 11
                                                 THE RESIDENCE
                                                 bear 1 di
                                                 OH - IN PROPERTY.
                                                                      Table Conclusion and Associated States and Conclusion and Concession and Concessi
                                                                      and a findament of the con-
                                                                          TOTAL CONTRACTOR AND ADDRESS OF THE PROPERTY OF THE PERSON OF THE PERSON
                                                                      add to accomply said to the company of the company 
                                                                          MAY - BORROWS - MORE MANAGED TO COMPANY OF THE PARTY OF T
                                                                          their of the supplicate beginning to within the comment of the
                                                      date, spec, also it is provided and the provided of the provided of
                                                 Mile Start of Artist Control of the 
                                                          profes i spransk (mentant)
                                                                          do not describe the
                                                                  ACCRECATE TO A CONTRACT OF THE PARTY OF THE 
                                                                      Many In the Art Life, Highlines
                                                                  All of the Art Chief Co., Art I Stormer I Section 1 (1977)
                                                                          MONEY TO SHARE A COMPANY OF THE PARTY OF THE
                                                                      mm = 16.441,00m, 4415.41
                                                                      men in the second secon
                                                                      disease I see self-test. It should
                                                 for the sent of liber, bern, day, who will the
                                                              the electricity of the forest of the first party of the contract of the contra
                                                      CHIEF CO. MIC MC. MIC MA. Mr. POLICEMON C.
```

```
THE RESERVE THE PARTY AND ADDRESS OF THE PARTY
      media is become of prospery trips bet weed.
                                                   A 12 WHEN THE WAR TO SEE THE THE THE THE THE THE
                                                To the Advantage of the Life
                                                40 1000 | 14 | 1 | 1
   590 - 11
                                                   that is by the comments of
                                                            5 C No Participal Services, 10 C Government, 45 C Sec.
                                                            21-1 AL-REVAND TR = 301
                                                            at it has enabled it as more has executed
                                                         or 1 of visite feeled made store, and a promising
                                                            of the particular religion process. (1)
                                                            bein parent sol
                                                makes less.
the same with a property of the same property of the same property of the same property of the same of
   the sale in the second platests, he make platests in some lateral for charge
   - mark had to be only from part by proper to the comment
THE PERSON NAMED IN
   and the state of t
                                                            NAME OF BRIDGE OF BRIDGE OF STREET
   THE R. P. LEWIS CO., LANSING, MICHIGAN
   THE RESERVE OF THE PARTY OF THE
                                                   Sergery or principles of the first or excellent content, Serger 122
      and the same of the same of
                                                         weeks pin septembers, medality, but
                                                            Fee 1: " London St. Holl and Total Co. St. Landon St. T.
                                                            prod Section of the Prof. of
- Lot, Str. Str., Str., Str., Str., Str., Str. | September, September, Street, Street,
                                                specificated in traceto black to home could be professional
   and the second of the second o
                                                   for make discover, was in the latter, and other state, and
                                                                                                                                                                                                                                                                         been new sen, sen, dist
                                                                                                                                                                                                                                                                            later, and part, on only
                                                            Marie IV (Maring New Yorks) of carbon - 2018 1 (Sound Street
                                                A reside bearing a series and a series
```

(https://gist.github.com/karpathy/d4dee 566867f8291f086)

Fei-Fei Li & Justin Johnson & Serena Yeung

Lecture 10 - 45 May 4, 2017

https://gist.github.com/karpathy/d4dee566867f8291f086

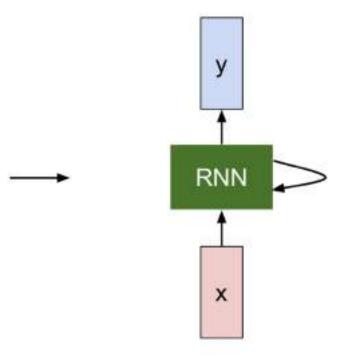


THE SONNETS

by William Shakespeare

From fairest circutures we desire increase,
That thereby brauty's row might never die,
But as the inper should by time decrease,
His sender heir might beer his money:
But thou, contracted in thine own bright eyes.
Feed's thy light's flame with self-substantial fuel,
Making a famine where abundance lies.
Thyself thy foe, to thy awest self too creel:
Those that art ness the world's fresh amariem,
And only hexald to the gasely spring.
Within thise own bud buriest thy content.
And under charf mak'st waste in organizing.
Pity the sortid, or else this glutton be,
To out the world's due, by the grave and thee.

When forty winters shall besiege thy brow,
And dig deep treather in thy beauty's field,
Thy yeath's proad livery so gazed on now,
Will be a satter'd weed of small worth held:
Then being asked, where all thy beauty lies,
Where all the treature of thy larry days,
To say, within thine own deep surices eyes,
Were an all-eating shame, and thriftless praise.
How much more printe deserved thy beauty's use,
If then couldn't asswer. This fair child of mine
Shall sum my count, and make my oid excase,
Proving his beauty by succession thine?
This were to be new made when thou art old,
And see thy brood warm when thou feel'st it cold.



at first:

tyntd-iafhatawiaoihrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e plia tklrgd t o idoe ns,smtt h ne etie h,hregtrs nigtike,aoaenns lng

train more

"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."

train more

Aftair fall unsuch that the hall for Prince Velzonski's that me of her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort how, and Gogition is so overelical and ofter.

train more

"Why do what that day," replied Natasha, and wishing to himself the fact the princess, Princess Mary was easier, fed in had oftened him. Pierre aking his soul came to the packs and drove up his father-in-law women.

PANDARUS:

Alas, I think he shall be come approached and the day When little srain would be attain'd into being never fed, And who is but a chain and subjects of his death, I should not sleep.

Second Senator:

They are away this miseries, produced upon my soul, Breaking and strongly should be buried, when I perish The earth and thoughts of many states.

DUKE VINCENTIO:

Well, your wit is in the care of side and that.

Second Lord:

They would be ruled after this chamber, and my fair nues begun out of the fact, to be conveyed, Whose noble souls I'll have the heart of the wars.

Clown:

Come, sir, I will make did behold your worship.

VIOLA:

I'll drink it.

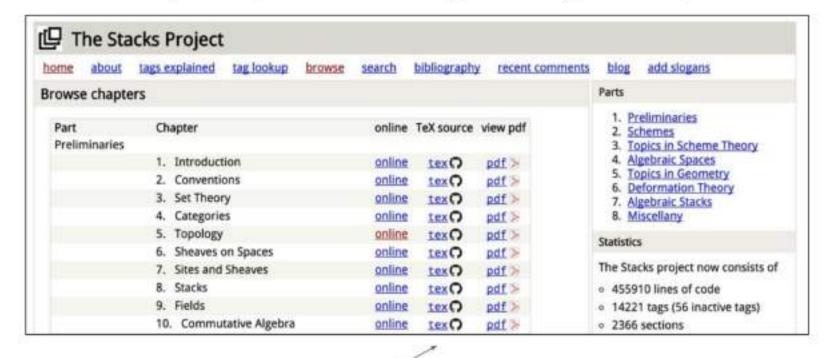
VIOLAT

Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:

O, if you were a feeble sight, the courtesy of your law, Your sight and several breath, will wear the gods With his heads, and my hands are wonder'd at the deeds, So drop upon your lordship's head, and your opinion Shall be against your honour.

The Stacks Project: open source algebraic geometry textbook



Latex source

http://stacks.math.columbia.edu/

The stacks project is licensed under the GNU Free Documentation License

For $\bigoplus_{n=1,...,m}$ where $\mathcal{L}_{m_*} = 0$, hence we can find a closed subset \mathcal{H} in \mathcal{H} and any sets \mathcal{F} on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \operatorname{Spec}(R) = U \times_X U \times_X U$$

and the comparisoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{Ippf} and $U \to U$ is the fibre category of S in U in Section, ?? and the fact that any U affine, see Morphisms, Lemma ??. Hence we obtain a scheme S and any open subset $W \subset U$ in Sb(G) such that $Spec(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_t is of finite presentation over S. We claim that $\mathcal{O}_{X,x'}$ is a scheme where $x,x',s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma ?? we can define a map of complexes $\operatorname{GL}_{S'}(x'/S'')$ and we win.

To prove study we see that $\mathcal{F}|_{U}$ is a covering of \mathcal{X}' , and \mathcal{T}_{i} is an object of $\mathcal{F}_{X/S}$ for i > 0 and \mathcal{F}_{p} exists and let \mathcal{F}_{i} be a presheaf of \mathcal{O}_{X} -modules on \mathcal{C} as a \mathcal{F} -module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^{\bullet} = I^{\bullet} \otimes_{\operatorname{Spec}(V)} \mathcal{O}_{S,s} - i_{s}^{-1} \mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

Arrows =
$$(Sch/S)_{fpof}^{npp}$$
, $(Sch/S)_{fpof}$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \operatorname{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets.

The result for prove any open covering follows from the less of Example ??. It may replace S by $X_{spaces,diale}$ which gives an open subspace of X and T equal to S_{Zar} , see Descent, Lemma ??. Namely, by Lemma ?? we see that R is geometrically regular over S. Lemma 0.1. Assume (3) and (3) by the construction in the description.

Suppose $X = \lim |X|$ (by the formal open covering X and a single map $\underline{Proj}_X(A) = \operatorname{Spec}(B)$ over U compatible with the complex

$$Set(A) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X})$$
.

When in this case of to show that $Q \to C_{Z/X}$ is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition ?? (without element is when the closed subschemes are catenary. If T is surjective we may assume that T is connected with resulue fields of S. Moreover there exists a closed subspace $Z \subset X$ of X where U in X' is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem

f is locally of finite type. Since S = Spec(R) and Y = Spec(R).

Preof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,...,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$.

The following lemma surjective restrocomposes of this implies that $F_{\nu_0} = F_{x_0} = F_{X_1...D}$.

Lemma 0.2. Let X be a locally Noetherian scheme over S, $E = \mathcal{F}_{X/S}$. Set $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}_n$. Since $\mathcal{I}^n \subset \mathcal{I}^n$ are nonzero over $i_0 \leq \mathfrak{p}$ is a subset of $\mathcal{J}_{n,0} \circ \overline{A}_2$ works.

Lemma 0.3. In Situation ??. Hence we may assume q' = 0.

Proof. We will use the property we see that p is the mext functor (??). On the other hand, by Lemma ?? we see that

$$D(\mathcal{O}_{X^{+}}) = \mathcal{O}_{X}(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S.

Proof. Omitted.

Lemma 0.1. Let C be a set of the construction.

Let C be a gerber covering. Let F be a quasi-coherent sheaves of O-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

Proof. This is an algebraic space with the composition of sheaves F on $X_{étale}$ we have

$$O_X(\mathcal{F}) = \{morph_1 \times_{O_X} (\mathcal{G}, \mathcal{F})\}\$$

where G defines an isomorphism $F \to F$ of O-modules.

Lemma 0.2. This is an integer Z is injective.

Proof. See Spaces, Lemma ??.

Lemma 0.3. Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $U \subset X$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b: X \to Y' \to Y \to Y \to Y' \times_X Y \to X$$
.

be a morphism of algebraic spaces over S and Y.

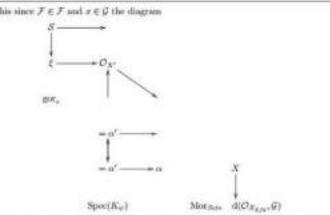
Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let \mathcal{F} be a quasi-coherent sheaf of O_X -modules. The following are equivalent

- F is an algebraic space over S.
- (2) If X is an affine open covering.

Consider a common structure on X and X the functor $O_X(U)$ which is locally of finite type.

This since $F \in F$ and $\sigma \in G$ the diagram

51 May 4, 2017



is a limit. Then G is a finite type and assume S is a flat and F and G is a finite type f_{\bullet} . This is of finite type diagrams, and

the composition of G is a regular sequence,

Lecture 10 -

. Ox- is a sheaf of rings.

Proof. We have see that $X = \operatorname{Spec}(R)$ and F is a finite type representable by algebraic space. The property F is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U.

Proof. This is clear that G is a finite presentation, see Lemmas 77.

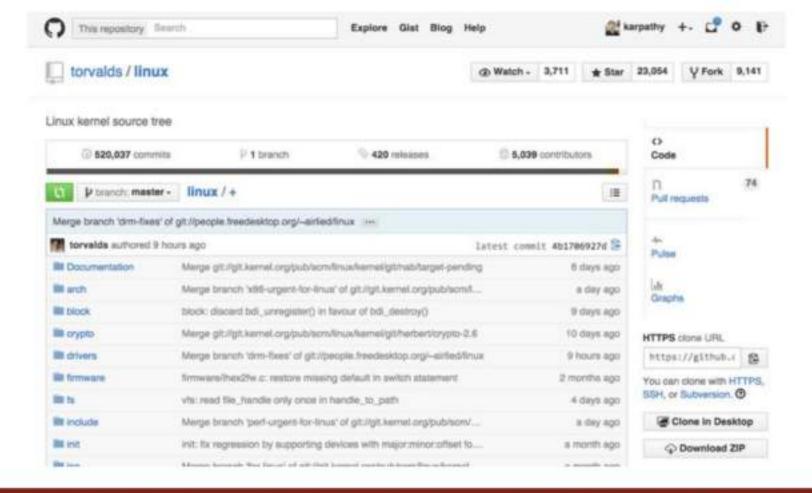
A reduced above we conclude that U is an open covering of C. The functor F is a "Beld"

$$\mathcal{O}_{X,a} \longrightarrow \mathcal{F}_{\overline{a}} \rightarrow \mathcal{I}(\mathcal{O}_{X_{Out}}) \longrightarrow \mathcal{O}_{X_{t}}^{-1}\mathcal{O}_{X_{t}}(\mathcal{O}_{X_{t}}^{p})$$

is an isomorphism of covering of O_{X_i} . If F is the unique element of F such that Xis an isomorphism.

The property F is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme O_X -algebra with F are opens of finite type over S. If \mathcal{F} is a scheme theoretic image points.

If F is a finite direct sum O_{X_k} is a closed immersion, see Lemma ??. This is a sequence of F is a similar morphism.

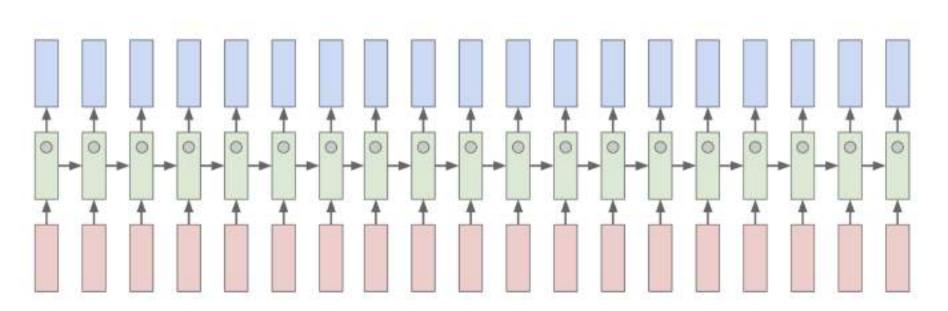


```
static void do command(struct seq file *m, void *v)
  int column = 32 << (cmd[2] & 0x80);
  if (state)
    cmd = (int)(int state ^ (in 8(&ch->ch flags) & Cmd) ? 2 : 1);
  else
    seq = 1;
  for (i = 0; i < 16; i++) (
   if (k & (1 << 1))
      pipe = (in use & UMXTHREAD UNCCA) +
        ((count & 0x0000000ffffffff8) & 0x000000f) << 8;
    if (count == 0)
      sub(pid, ppc md.kexec handle, 0x20000000);
    pipe set bytes(i, 0);
  /* Free our user pages pointer to place camera if all dash */
  subsystem info = &of changes[PAGE SIZE];
 rek controls(offset, idx, &soffset);
  /* Now we want to deliberately put it to device */
  control check polarity(&context, val, 0);
  for (i = 0; i < COUNTER; i++)
    seq puts(s, "policy ");
```

Generated C code

```
Copyright (c) 2006-3010, Intel Mobile Communications. All rights reserved.
    This program is free software; you can redistribute it and/or modify it
 * under the terms of the GMU General Public License version 2 as published by
 * the Pree Software Foundation.
         This program is distributed in the hope that it will be useful,
 * but WITHOUT ANY WARRANTY; without even the implied warranty of
     MERCHARTABILITY OF FITNESS FOR A FARTICULAR PURPOSE. See the
   GMV General Public Livense for more details.
     You should have received a copy of the GMM General Public License
     along with this program; if not, write to the Free Software Foundation,
   Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
 01
finelude inug/house.h>
#include sur/errou.h>
#include nux/io.b>
Windlads Sur/platform device.h>
#include inux/multi.h>
#include inux/ckevent.h>
#include <mm/io.b>
#include <mem/prom.k>
Finclude <pen/e820.ho
#include <asm/system_info.h>
#include <nem/setme.h>
finclads <asm/pgproto.b>
```

```
#include <anm/io.b>
#include <nsm/prom.h>
#include <asm/e820.h>
#include <asm/system info.h>
#include <asm/setew.h>
#include <msm/pgproto.h>
#define REG FQ wess slot addr pack
#define PFM NOCOMP AFER(0, load)
#define STACK_DDR(type) (func)
#define SWAP_ALLOCATE(nr)
                            (m)
#define smulste sigs() arch get unaligned child()
#define access rw(TST) asm volatile("movd %tesp, %0, %3" : : "r" (0)); \
 if (_type & DO_READ)
static void stat PC SEC read mostly offsetof(struct seq argsqueue, \
         pC>[1]);
static void
os prefix(unsigned long sys)
#ifdef CONFIG PREEMPT
  PUT PARAM RAID(2, sel) = get state state();
  set pid sum((unsigned long)state, current state str(),
           (unsigned long)-1->1r full; low;
```



```
/* Unpack a filter field's string representation from user-space
   buffer. */
char *audit_unpack_string(void *bufp, size_t *remain, size_t len)
{
   char *str;
   if (!*bufp || (len == 0) || (len > *remain))
     return ERR_PTR(-EINVAL);
/* Of the currently implemented string fields, PATH_HAX
     defines the longest valid length.
   */
```

```
"You mean to imply that I have nothing to eat out of... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."
```

quote detection cell

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.
```

line length tracking cell

```
static int __dequeue_signal(struct signending 'pending, sigset_t 'mask,
    siginfo_t 'info)

int sig = next_signal(pending, mask);

if (sig)

if (surrent->notifier) {
    if (sigismember(current->notifier_mask, sig)) {
        if (l(current->notifier)(current->notifier_data)) {
            clear_thread_flag(TIF_SIGPENDING);
            return 0;
        }

}

collect_signal(sig, pending, info);
}

return sig;
}
```

if statement cell

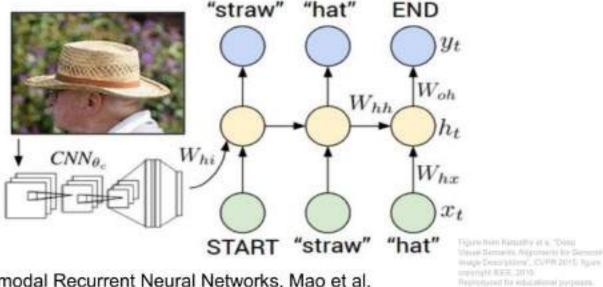
```
Cell that turns on inside comments and quotes:
                               quote/comment cell
```

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)

int i;
if (classes[class]) {
  for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
   if (mask[i] & classes[class][i])
   return B;
}
return 1;</pre>
```

code depth cell

Image Captioning

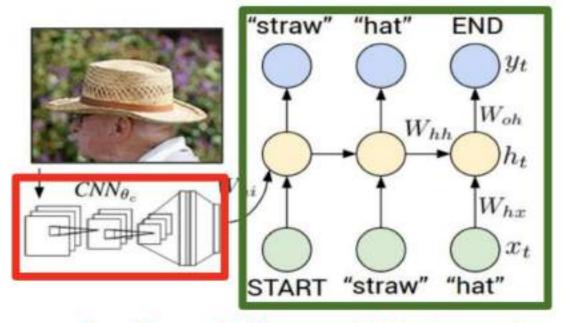


Explain Images with Multimodal Recurrent Neural Networks, Mao et al.

Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei Show and Tell: A Neural Image Caption Generator, Vinyals et al.

Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al. Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick

Recurrent Neural Network



Convolutional Neural Network



test image

This triage is GCR public detroits

90

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096

FC-1000

softmax



test image

image

conv-64

conv-64

maxpool

conv-128

conv-128

maxpool

conv-256

conv-256

maxpool

conv-512

conv-512

maxpool

conv-512

conv-512

maxpool

FC-4096

FC-4096



test image



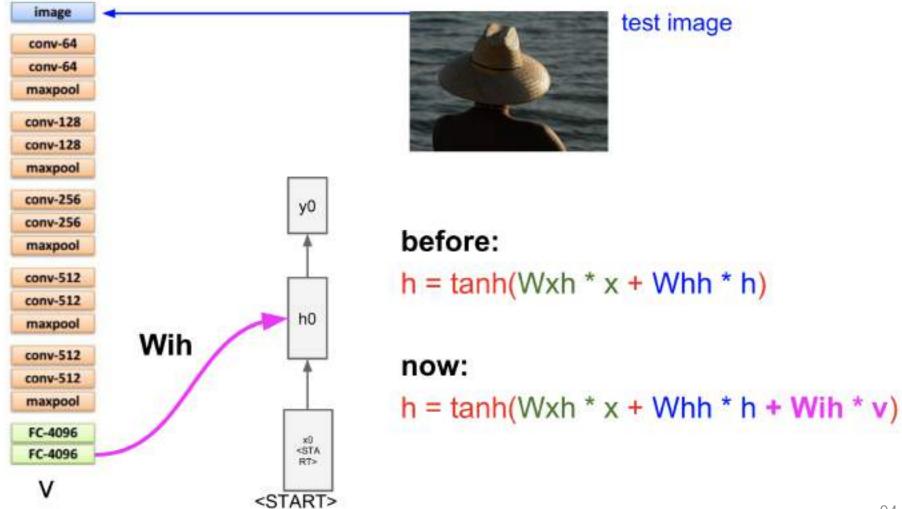
image conv-64 conv-64 maxpool conv-128 conv-128 maxpool conv-256 conv-256 maxpool conv-512 conv-512 maxpool conv-512 conv-512 maxpool FC-4096

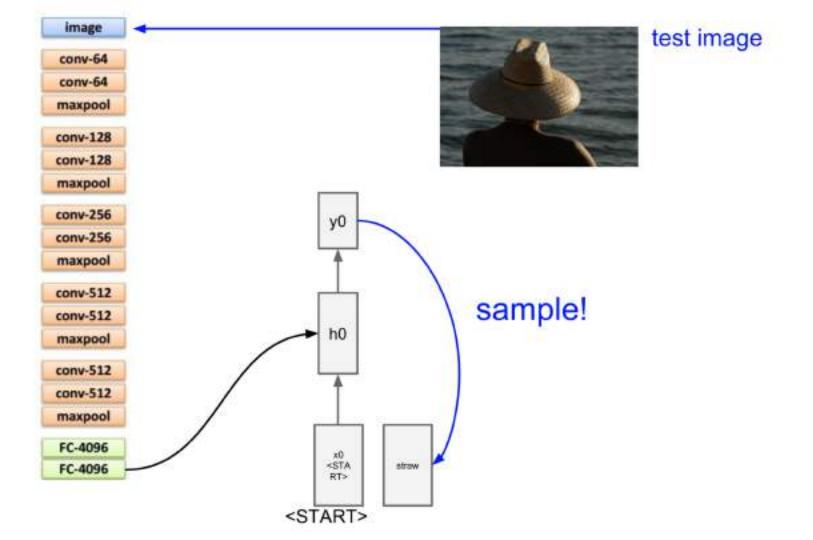
FC-4096

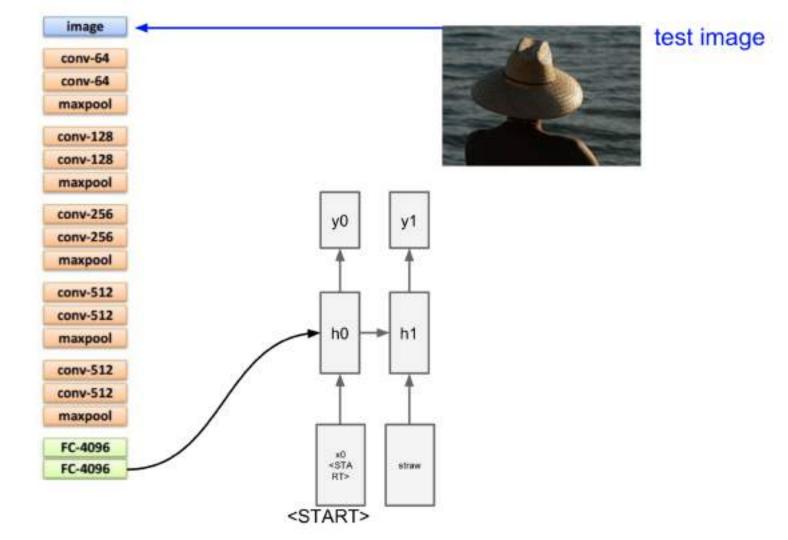


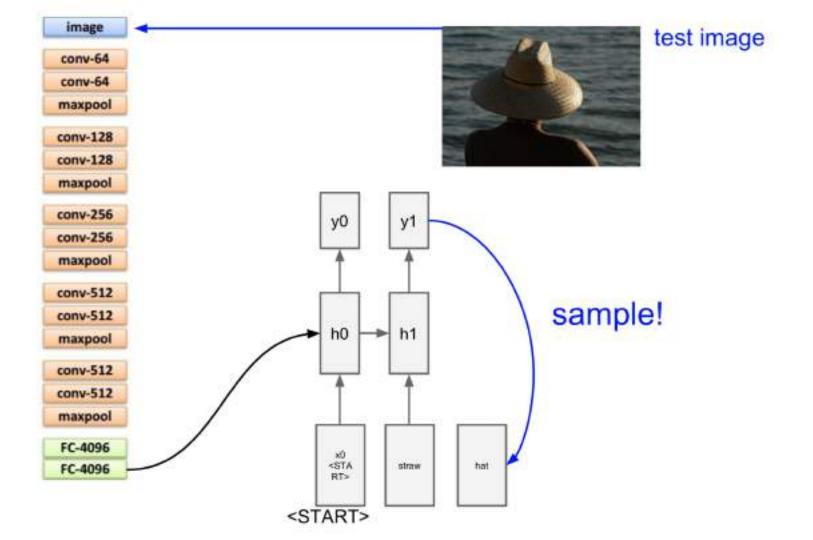
test image

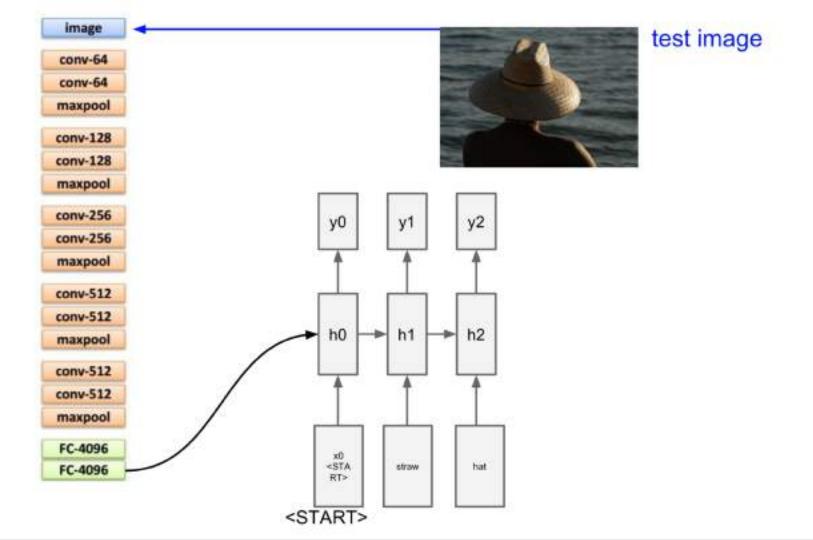


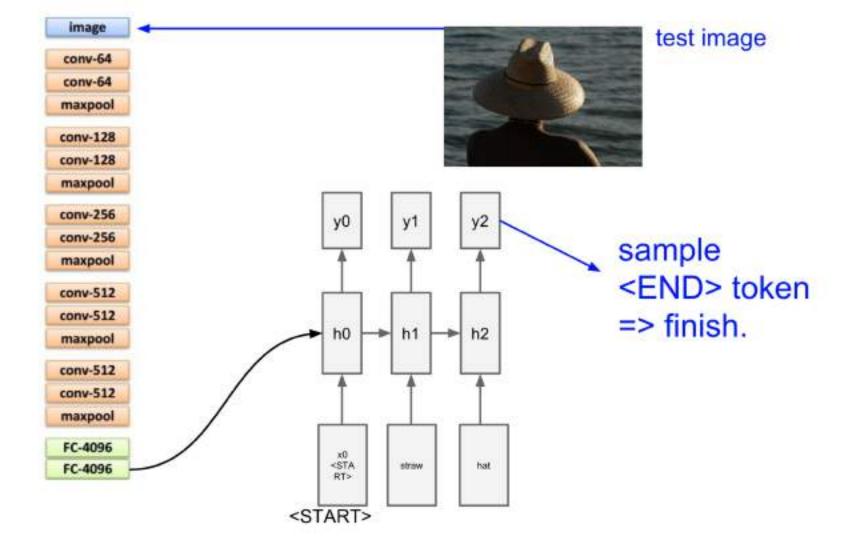














A cat sitting on a suitcase on the floor



A cat is sitting on a tree branch



A dog is running in the grass with a frisbee



A white teddy bear sitting in the grass



Two people walking on the beach with surfboards



A tennis player in action on the court



Two giraffes standing in a grassy field



A man riding a dirt bike on a dirt track

Image Captioning: Failure Cases



A woman is holding a cat in her hand



A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard

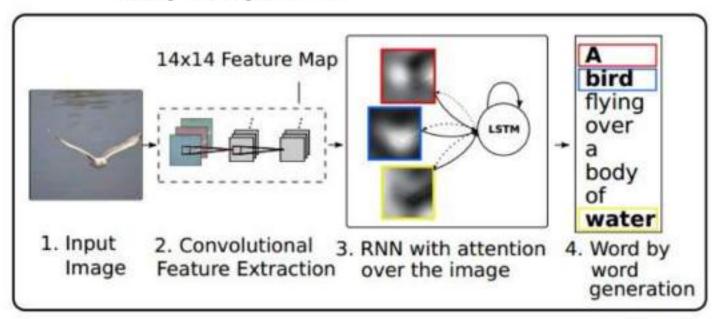


A bird is perched on a tree branch

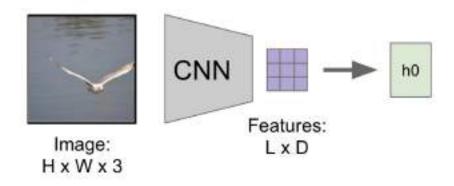


A man in a baseball uniform throwing a ball

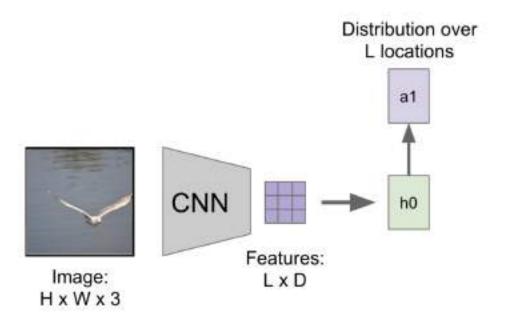
RNN focuses its attention at a different spatial location when generating each word



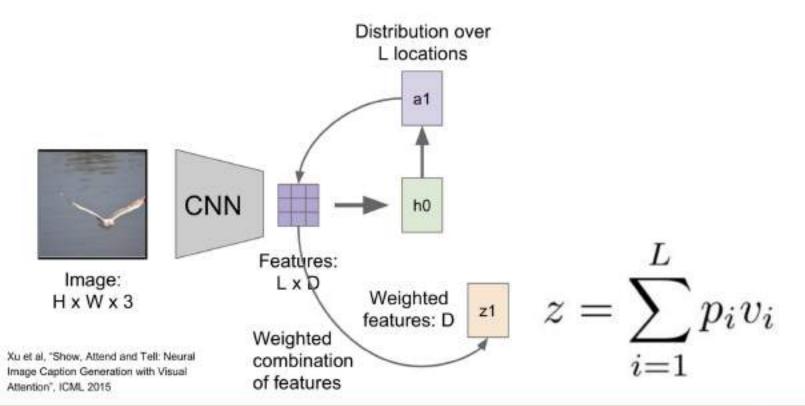
Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015
Figure copyright Kelvin Xu, Jimmy Lei Ba, Jamie Kiros, Kyunghyun Cho, Aaron Courville, Rustan Salakhutdinov, Richard S. Zemel, and Yoshua Benchio. 2016. Reproduced with permission

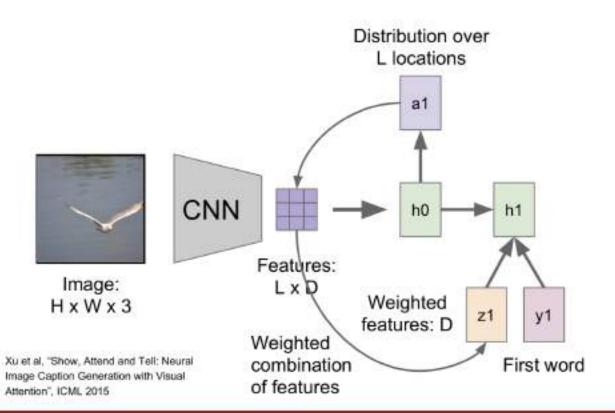


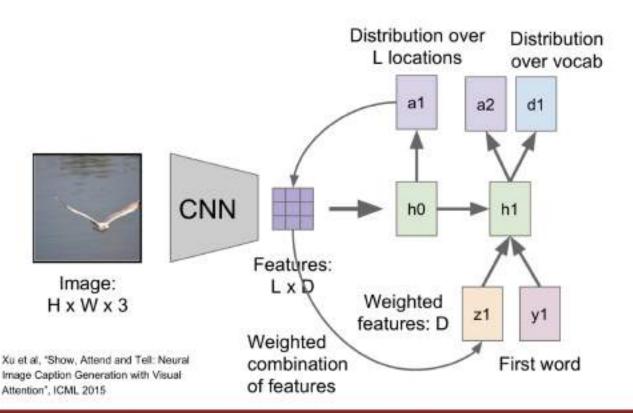
Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015







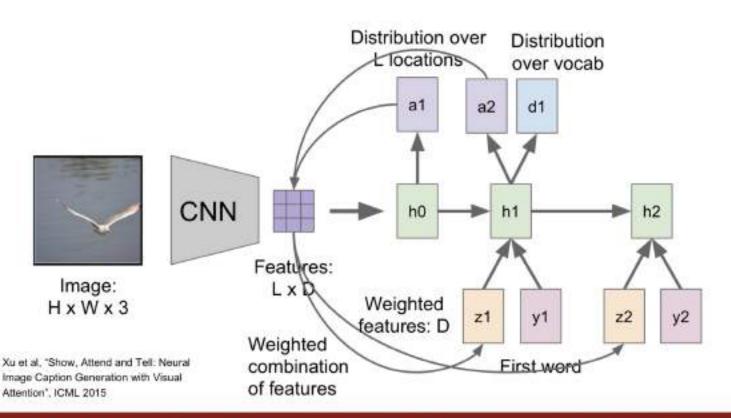


Image Captioning with Attention

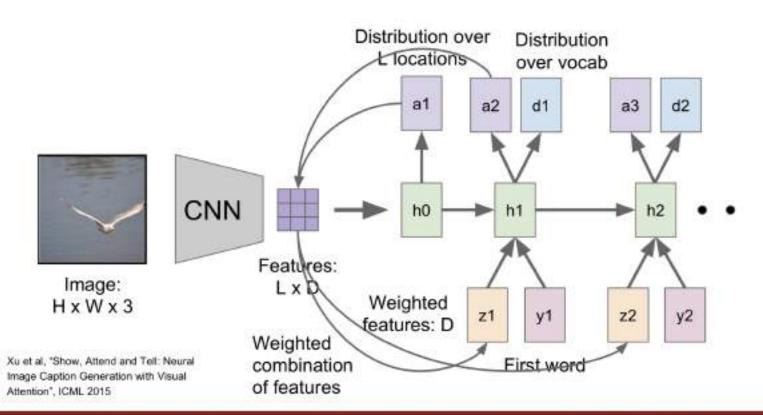
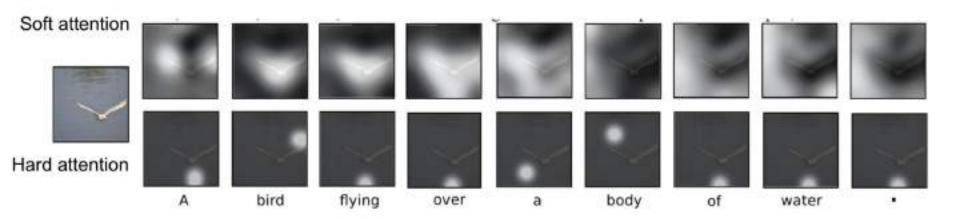


Image Captioning with Attention



Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015
Figure copyright Kelvin Xu. Jimmy Lei Ba, Jamie Kros. Kyunghyun Cho, Aaron Counville, Rustan Salakhutdinov. Richard S. Zemel, and Yoshua Banchio, 2015. Reproduced with permission.

Image Captioning with Attention



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of <u>people</u> sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Xu et al, "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015
Figure copyright Kelvin Xu, Jimmy Lei Ba, Jamie Kiros, Kyunghyun Cho, Aaron Countile, Rusian Salakhutdinox, Richard S. Zemel, and Yoshua Benchio, 2015. Reproduced with permission.

Visual Question Answering



Q: What endangered animal is featured on the truck?

- A: A bald eagle.
- A: A sparrow.
- A: A humming bird.
- A: A raven.



Q: Where will the driver go if turning right?

- A: Onto 24 % Rd.
- A: Onto 25 % Rd.
- A: Onto 23 % Rd.
- A: Onto Main Street.



Q: When was the picture taken?

- A: During a wedding.
- A: During a bar mitzvah.
- A: During a funeral.
- A: During a Sunday church



Q: Who is under the umbrella?

- A: Two women.
- A: A child
- A: An old man.
- A: A husband and a wife.

Agrawal et al, "VQA: Visual Question Answering", ICCV 2015
Zhu et al, "Visual 7W: Grounded Question Answering in Images", CVPR 2016
Figure from Zhu et al, copyright IEEE 2016. Reproduced for educational purposes.

Multilayer RNNs

$$h_t^l = \tanh W^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$h \in \mathbb{R}^n \quad W^l \ [n \times 2n]$$

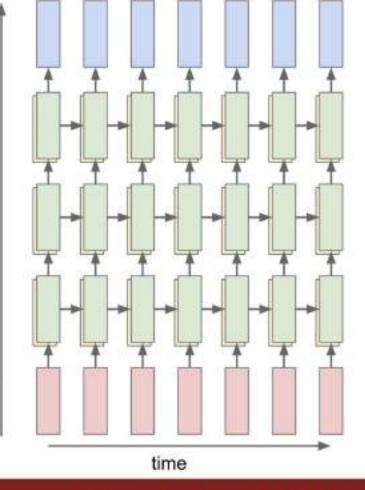
LSTM:

$$W^{l} [4n \times 2n]$$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{pmatrix} W^{l} \begin{pmatrix} h_{t}^{l-1} \\ h_{t-1}^{l} \end{pmatrix}$$

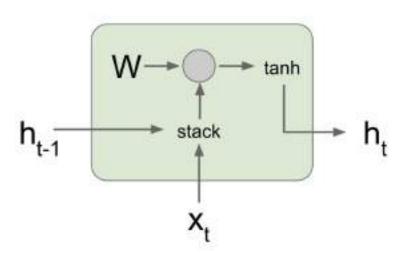
$$c_{t}^{l} = f \odot c_{t-1}^{l} + i \odot g$$

$$h_{t}^{l} = o \odot \tanh(c_{t}^{l})$$



depth

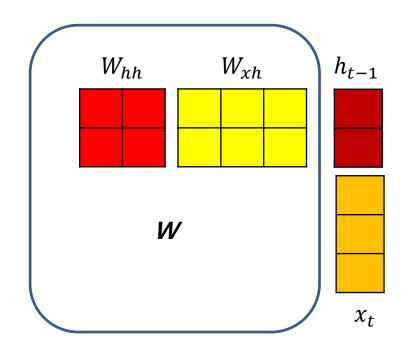
Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013

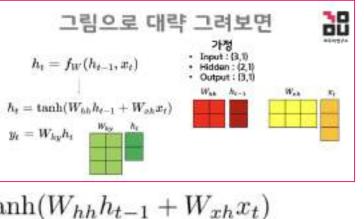


$$h_{t} = \tanh(W_{hh}h_{t-1} + W_{xh}x_{t})$$

$$= \tanh\left(\left(W_{hh} \quad W_{hx}\right) \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$

$$= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$





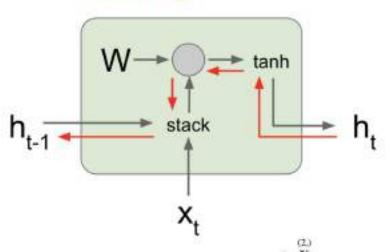
$$h_{t} = \tanh(W_{hh}h_{t-1} + W_{xh}x_{t})$$

$$= \tanh\left(\left(W_{hh} \quad W_{hx}\right) \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$

$$= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013

Backpropagation from h, to h, multiplies by W (actually W, T)



(2, 3)

(3,)

$$h_{t} = \tanh(W_{hh}h_{t-1} + W_{xh}x_{t})$$

$$= \tanh\left((W_{hh} \quad W_{hx}) \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$

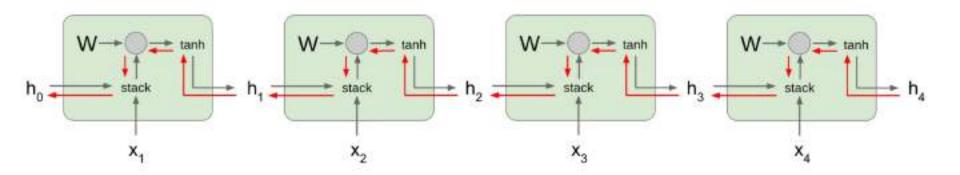
$$= \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_{t} \end{pmatrix}\right)$$

$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^{\mathrm{T}}$$
$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^{\mathrm{T}} \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

Fei-Fei Li & Justin Johnson

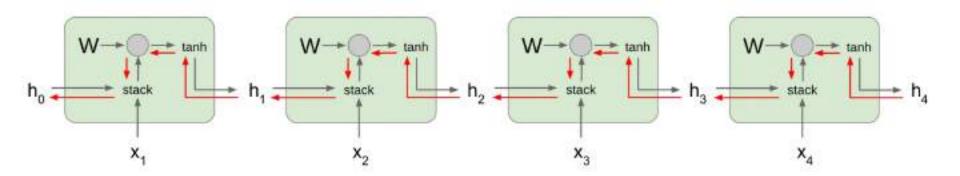
Lecture 10 - 91 May 4, 2017

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Computing gradient of h₀ involves many factors of W (and repeated tanh)

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013

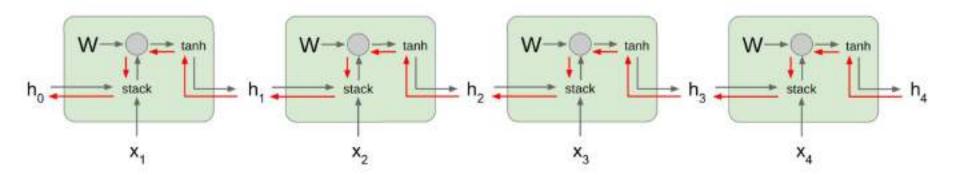


Computing gradient of h₀ involves many factors of W (and repeated tanh)

Largest singular value > 1: Exploding gradients

Largest singular value < 1: Vanishing gradients

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



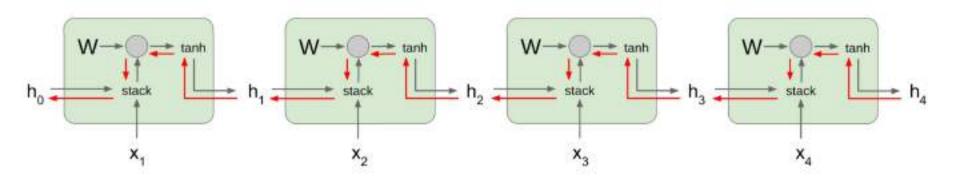
Computing gradient of h₀ involves many factors of W (and repeated tanh)

Largest singular value > 1: Exploding gradients

Largest singular value < 1: Vanishing gradients Gradient clipping: Scale gradient if its norm is too big

```
grad_norm = np.sum(grad * grad)
if grad_norm > threshold:
    grad *= (threshold / grad_norm)
```

Bengio et al, "Learning long-term dependencies with gradient descent is difficult", IEEE Transactions on Neural Networks, 1994 Pascanu et al, "On the difficulty of training recurrent neural networks", ICML 2013



Computing gradient of h₀ involves many factors of W (and repeated tanh)

Largest singular value > 1: Exploding gradients

Largest singular value < 1: Vanishing gradients

Change RNN architecture

Long Short Term Memory (LSTM)

Vanilla RNN

$$h_t = \tanh\left(W\begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

LSTM

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

Hochreiter and Schmidhuber, "Long Short Term Memory", Neural Computation 1997

Long Short Term Memory (LSTM)

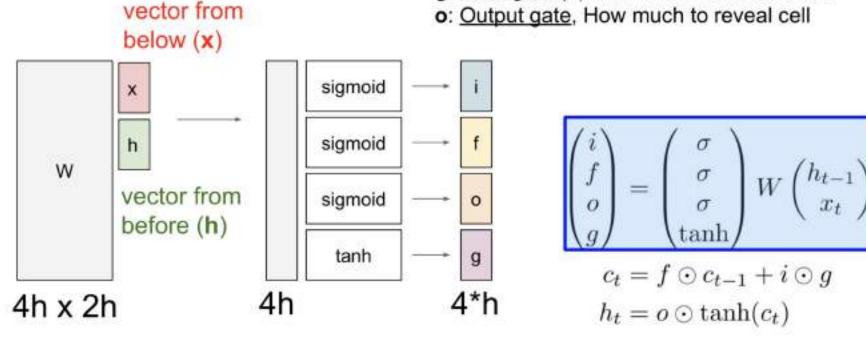
[Hochreiter et al., 1997]

f: Forget gate, Whether to erase cell

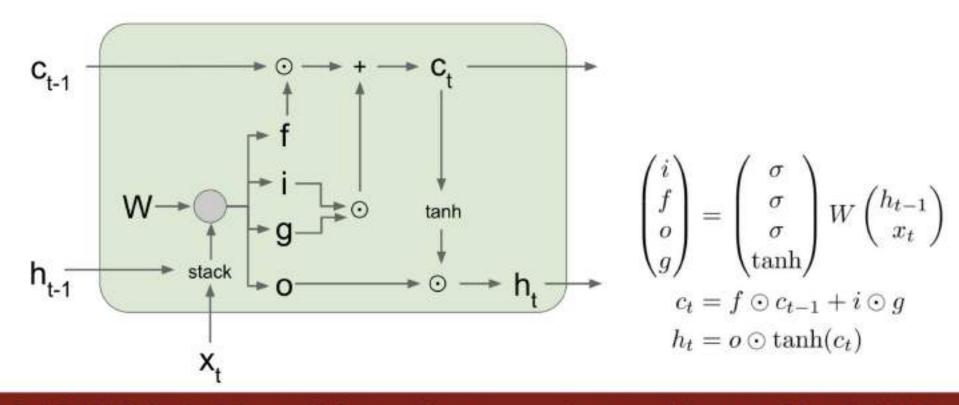
i: Input gate, whether to write to cell

g: Gate gate (?), How much to write to cell

o: Output gate, How much to reveal cell

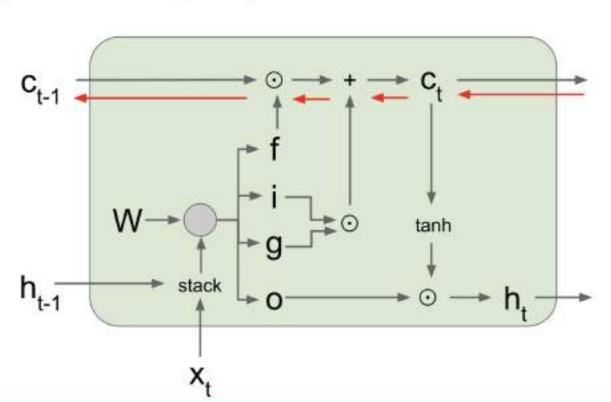


Long Short Term Memory (LSTM) [Hochreiter et al., 1997]



Long Short Term Memory (LSTM): Gradient Flow

[Hochreiter et al., 1997]

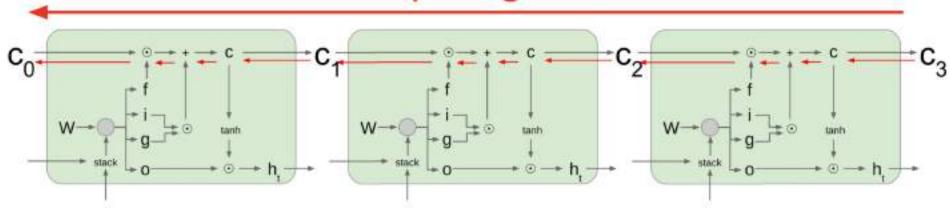


Backpropagation from c_t to c_{t-1} only elementwise multiplication by f, no matrix multiply by W

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$
$$c_t = f \odot c_{t-1} + i \odot g$$
$$h_t = o \odot \tanh(c_t)$$

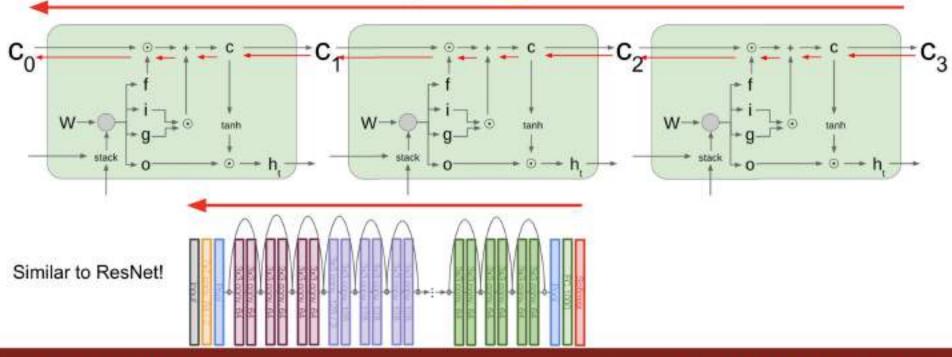
Long Short Term Memory (LSTM): Gradient Flow [Hochreiter et al., 1997]

Uninterrupted gradient flow!



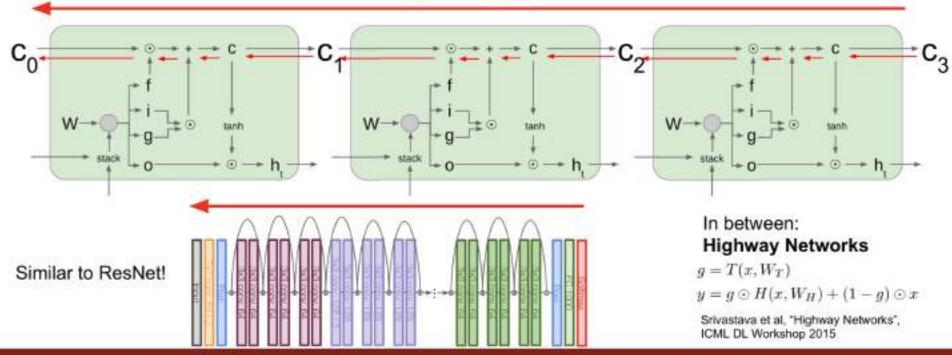
Long Short Term Memory (LSTM): Gradient Flow [Hochreiter et al., 1997]

Uninterrupted gradient flow!



Long Short Term Memory (LSTM): Gradient Flow [Hochreiter et al., 1997]

Uninterrupted gradient flow!



Other RNN Variants

GRU [Learning phrase representations using rnn encoder-decoder for statistical machine translation, Cho et al. 2014]

$$r_{t} = \sigma(W_{xr}x_{t} + W_{hr}h_{t-1} + b_{r})$$

$$z_{t} = \sigma(W_{xz}x_{t} + W_{hz}h_{t-1} + b_{z})$$

$$\tilde{h}_{t} = \tanh(W_{xh}x_{t} + W_{hh}(r_{t} \odot h_{t-1}) + b_{h})$$

$$h_{t} = z_{t} \odot h_{t-1} + (1 - z_{t}) \odot \tilde{h}_{t}$$

[LSTM: A Search Space Odyssey, Greff et al., 2015] [An Empirical Exploration of Recurrent Network Architectures, Jozefowicz et al., 2015]

```
MUTI:
       z = \operatorname{sigm}(W_{xx}x_t + b_x)
           = \operatorname{sigm}(W_{xr}x_t + W_{hr}h_t + b_r)
   h_{t+1} = \tanh(W_{hh}(r \odot h_t) + \tanh(x_t) + b_h) \odot z
            + h<sub>t</sub> ⊙ (1-z)
MUT2:
             = \operatorname{sigm}(W_{xx}x_t + W_{bx}h_t + b_x)
             = \operatorname{sigm}(x_t + W_{br}h_t + b_r)
    h_{t+1} = \tanh(W_{hh}(r \odot h_t) + W_{xh}x_t + b_h) \odot z
              + h<sub>t</sub> ⊙ (1-z)
MUT3:
             = \operatorname{sigm}(W_{xx}x_t + W_{hx} \tanh(h_t) + b_s)
            = \operatorname{sigm}(W_{xr}x_t + W_{hr}h_t + b_r)
    h_{t+1} = \tanh(W_{hh}(r \odot h_t) + W_{xh}x_t + b_h) \odot z
```

+ h, @ (1-z)

Summary

- RNNs allow a lot of flexibility in architecture design
- Vanilla RNNs are simple but don't work very well
- Common to use LSTM or GRU: their additive interactions improve gradient flow
- Backward flow of gradients in RNN can explode or vanish.
 Exploding is controlled with gradient clipping. Vanishing is controlled with additive interactions (LSTM)
- Better/simpler architectures are a hot topic of current research
- Better understanding (both theoretical and empirical) is needed.





时 七午 Research Director

E-mail: es.park@modulabs.co.kr