Recurrent Neural Networks (RNNs)

dayche.com . گروه دایکه

دای‌کِه DAYCHE

Data types

- Tabular data

Owner	Country	File_Date	IPC_Class
Company A	US	6/18/2008	H05H13
Company A	EP	1/30/1998	A61N5
Company A	EP	1/30/1998	A61N5
Company A	EP	1/30/1998	A61N5
Company A	JP	8/28/1997	A61N5
Company A	JP	10/4/2002	A61N5
Company A	JP	1/27/2003	A61N5
Company A	JP	4/14/2003	A61N5
Company A	JP	5/13/2011	A61N5
Company B	JP	4/2/1998	G12B13
Company B	JP	4/2/1998	G12B13
Company B	JP	5/28/1997	A61N5
Company B	JP	11/12/1997	A61N5
Company B	JP	2/29/2000	A61N5
Company B	JP	4/30/2002	A61N5

هدف اصلی در توسعه یک الگوریتم یادگیر، مدل کردن توزیع توام $P(x,y)$ است که فرض می‌شود داده‌گانی که در دست هستند از این توزیع، به صورت مستقل از یکدیگر نمونه‌برداری شده‌اند.

- هر داده‌گانی که قابلیت نمایش به صورت جدولی مقابل را داشته باشد ازین قاعده مستثنی نیست.

این فرض برای برخی از داده‌های صادق نیست.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Spatially dependent data

در یک داده‌ی تصویری، هر پیکسل به عنوان ورودی در نظر گرفته می‌شود
جابجا کردن این پیکسل‌ها موجب از دست رفتن معنا در داخل تصویر خواهد شد
پیکسل‌ها بر اساس یک وابسته مکانی یک مفهوم بینایی را منتقل می‌کنند.

- برای مدل‌سازی اینگونه داده‌ها چه باید کرد؟

تولید محتوا: وحید محمدزاده ایوقی

Temporally dependent data

یادگیری ماشین قصد دارد تا بخشی از مغز انسان را که مربوط به یادگیری است مدل کند.

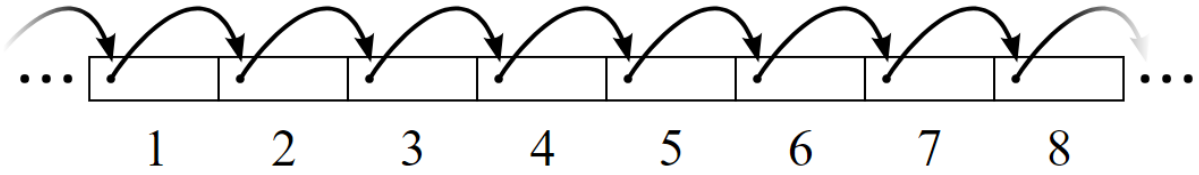
مدل ماشین مغز قصد دارد تا مربوط به یادگیری انسان مغز از است کند را بخشی که.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Sequential data

- پدیده‌های پیرامون ما پیوسته در حال تغییر هستند.

$$x_{1:T} = x_1, x_2, \ldots, x_T, \qquad T > 1$$



یک ترتیب زمانی و یا مکانی در داده‌ها وجود دارد.

- داده‌های سری زمانی
- داده‌های ویدیویی
- داده‌های متنی
- داده‌های تصویری

- روند امتیاز دهی و تغییر عقیده درباره یک پدیده
- پیش‌بینی بازار بورس
- پیش‌بینی آب و هوا
- زلزله و پس‌لرزه
- و ...

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Biological interpretation

- اتفاقاتی که به صورت پیوسته در فریم‌های مختلف یک ویدیو رخ می‌دهد یک فیلم را روایت می‌کند.
- طبق تحقیقات انجام شده، در طول تماشای یک فیلم انسان به طور متوسط 15 دقیقه یک فیلم 2 ساعته را از دست می‌دهد. درک روایت فیلم چطور ممکن است؟

- از دیدگاه ریاضی نوع داده‌های ویدیویی به صورت یک دنباله است.
- عملکرد مغز در مواجه با این داده‌ها به گونه‌ایست که قادراست ارتباط بین فریم‌های مختلف را capture کند.
- از دیدگاه بیولوژیکی، شبکه‌های عصبی بازگشتی بیشترین تطابق را با مغز دارند.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Stock market prediction

اگر قیمت روزهای قبل یک سهم را داشته باشیم، چگونه می‌توان قیمت روز بعد را پیش‌بینی کرد؟

$$y_t \sim P(y_t | y_{t-1}, \dots, y_1)$$

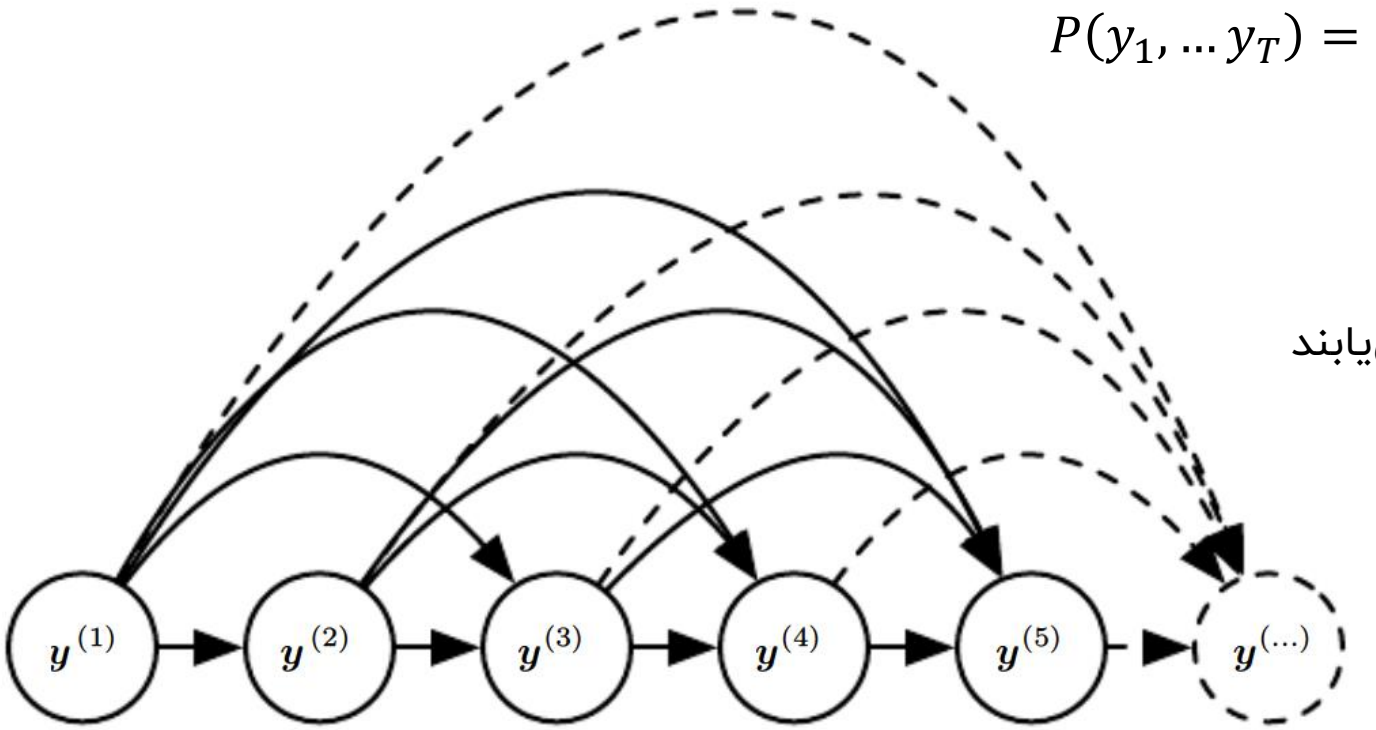
مسئله در حالت کلی به شکل فوق خواهد بود، اما امکان حل دقیق آن وجود دارد. زیرا تعداد پارامترهایی که این مدل احتمالاتی شرطی دارد بسیار زیاد است و در حالت کلی با افزایش طول دنباله افزایش می‌یابد.

پیش‌بینی قیمت یک سهم بر اساس تاریخچه سهم

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Directed graphical models

گراف اتصال کامل



$$P(y_1, \dots y_T) = P(y_1)P(y_2|y_1)P(y_3|y_1, y_2) \dots P(y_t|y_{t-1}, \dots, y_1)$$

$$= \prod_{t=1}^{T} P(y_t|y_1, \dots, y_{t-1})$$

با گذر زمان تعداد متغیرهایی که بهم ارتباط دارند افزایش می‌یابند

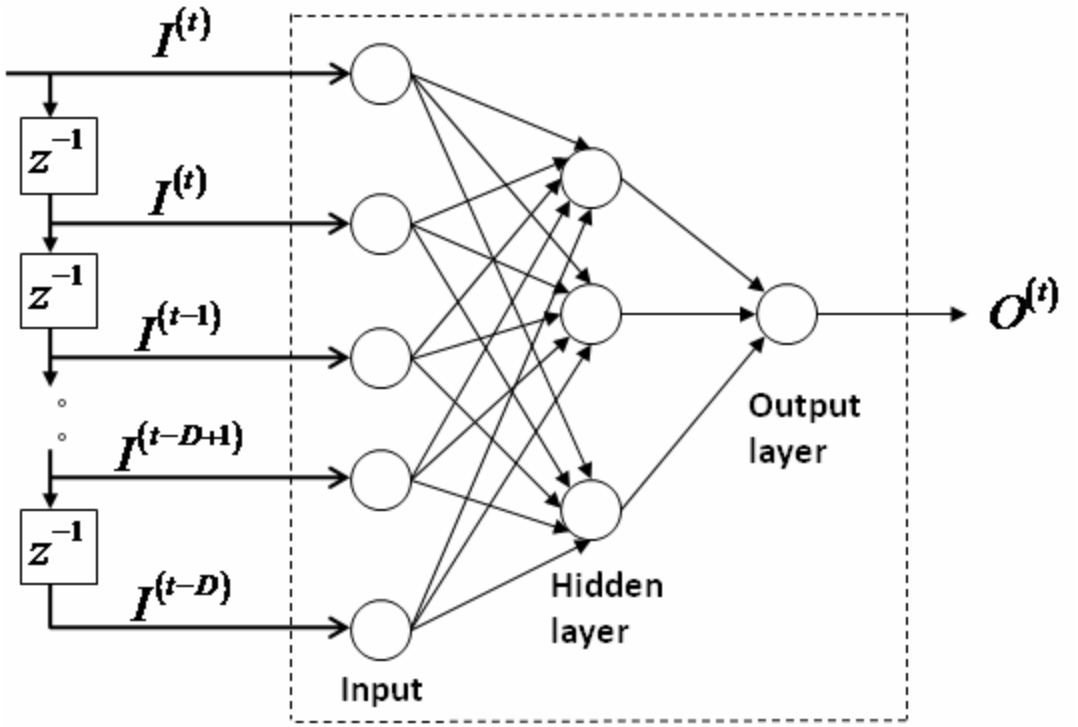
- تعداد ورودی‌های مدل وابسته به مدل است
- امکان استفاده از شبکه‌های عصبی معمولی وجود ندارد

به منظور مدل کردن این توزیع احتمال، نیاز است تا فرضیات ساده‌سازی به مسئله تحمیل شود

Markov assumption (fixed window)

$$P(y_1, \dots y_T) = P(y_1)P(y_2|y_1)P(y_3|y_1, y_2) \dots P(y_t|y_{t-1}, \dots, y_1) = \prod_{t=1}^{T} P(y_t|y_1, \dots, y_{t-1})$$

$$y_t \sim P(y_t|y_{t-1}, \dots, y_{t-\tau})$$

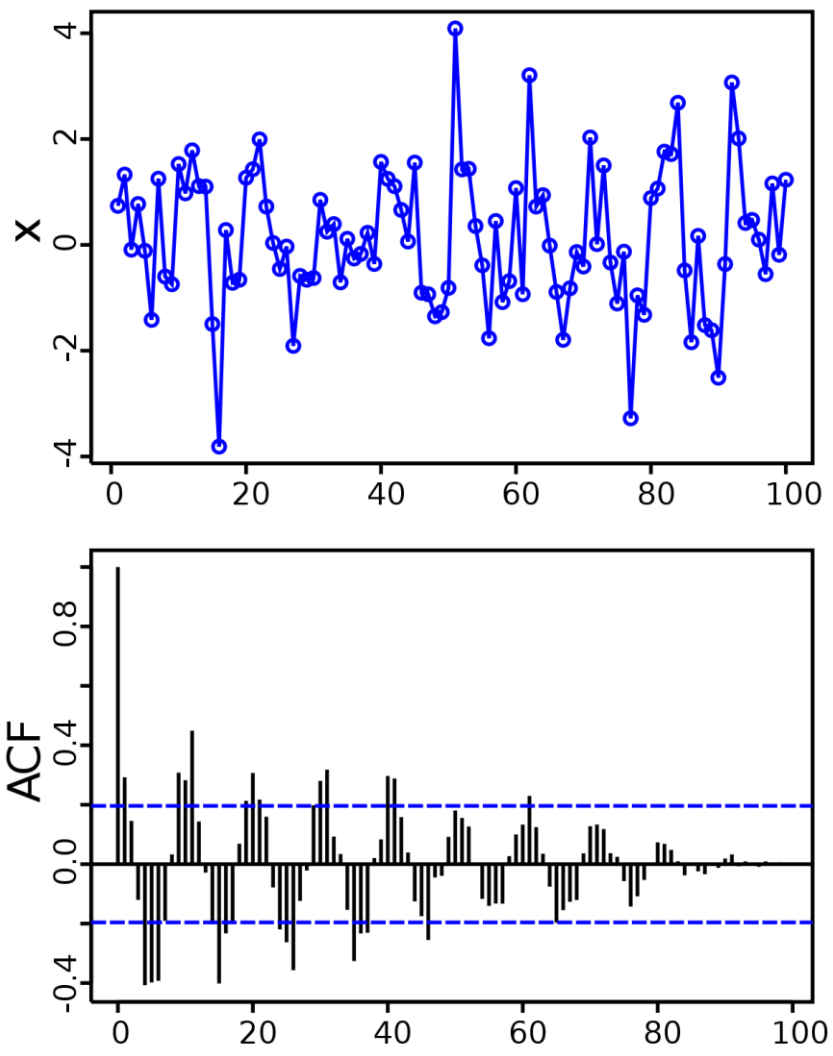


$$y_t = f(y_{t-1}, \dots, y_{t-\tau}; W)$$

چالش اساسی: مقدار τ چگونه تعیین می‌شود؟

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Correlation analysis



- پیدا کردن عمق وابستگی (عمق دینامیک)
  - دینامیک‌های خطی

$$R_{yy}(\tau) = E(y(t)y(t-\tau)^T)$$

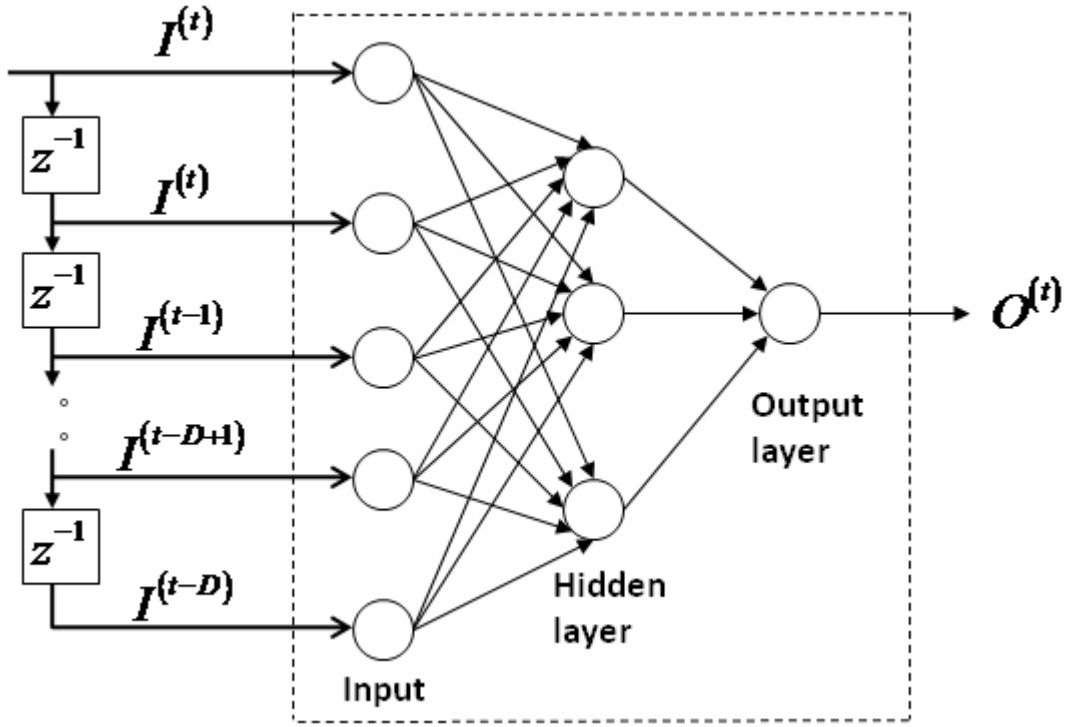
$$\widehat{R}_{yy}(\tau) = \frac{1}{N}\sum_{t=\tau+1}^{N} x(t)x(t-\tau)$$

$$\widehat{R}_{yy}(\tau) = \frac{1}{N-\tau}\sum_{t=\tau+1}^{N} x(t)x(t-\tau)$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Drawbacks of fixed window approach

- محدودیت پارامترها
  - تعداد پارامترهای افزایش خواهد یافت چون تعداد ورودی‌های مسئله بزرگ شده است.



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Drawbacks of fixed window approach

- وابستگی‌های با طول بزرگ
  - به منظور جلوگیری از کاهش تعداد پارامترها، ناگزیریم که وابستگی را در عمق کمتری در نظر بگیریم – وابستگی با طول بزرگ چه می‌شود؟

"**France** is where I grew up, but I now live in Boston. I speak fluent ____."

پیش‌بینی مدل: English

پیش‌بینی مورد انتظار: France

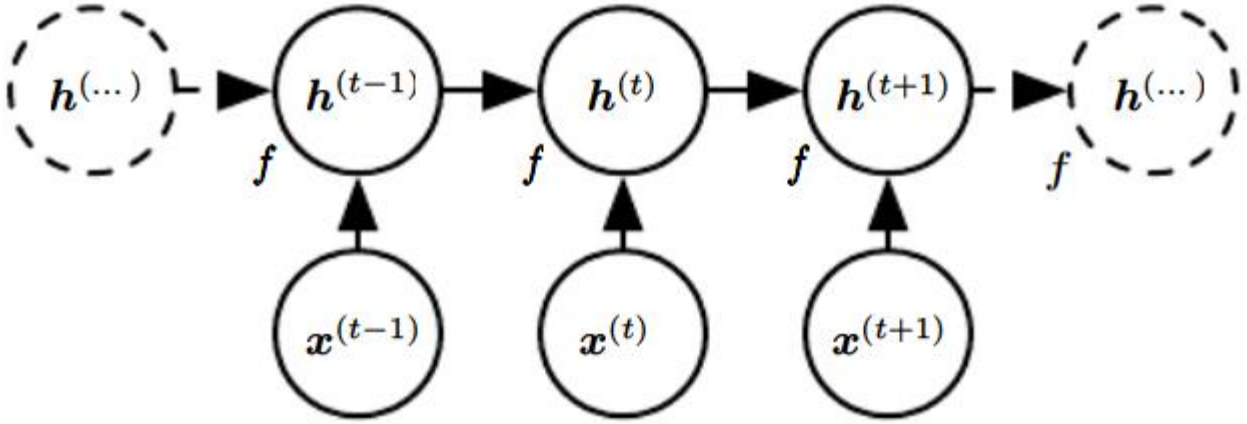
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Introduction to latent variable models

- گاهی اوقات نیاز است تا وابستگی‌هایی با طول بزرگتر را به منظور مدل‌سازی صحیح در نظر بگیریم.

$$P(y_t|y_{t-1}, \dots, y_1) \approx P(y_t|h_{t-1})$$

متغیر پنهان (حالت) – حاوی اطلاعات گذشته از وابستگی بین ورودی‌ها



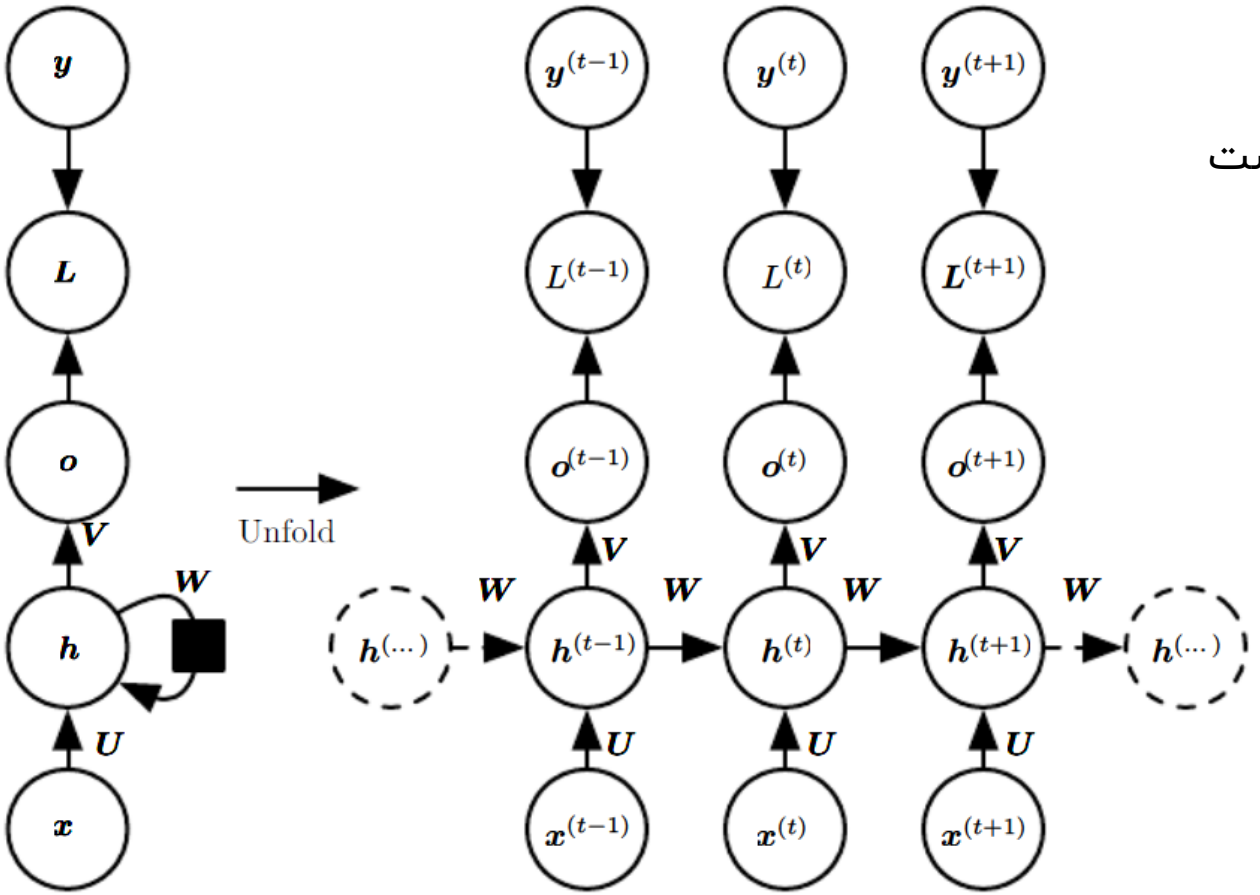
$$h_t = f(h_{t-1}, x_t; W)$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Recurrent neural networks

- ساختارهای متفاوتی برای شکل‌گیری یک مدل بازگشتی وجود دارد.

پرکاربردترین ساختار-Hidden to hidden



- باید این پارامترها آموزش ببینند
- روند آموزش مشابه آموزش یک شبکه عصبی معمولی است

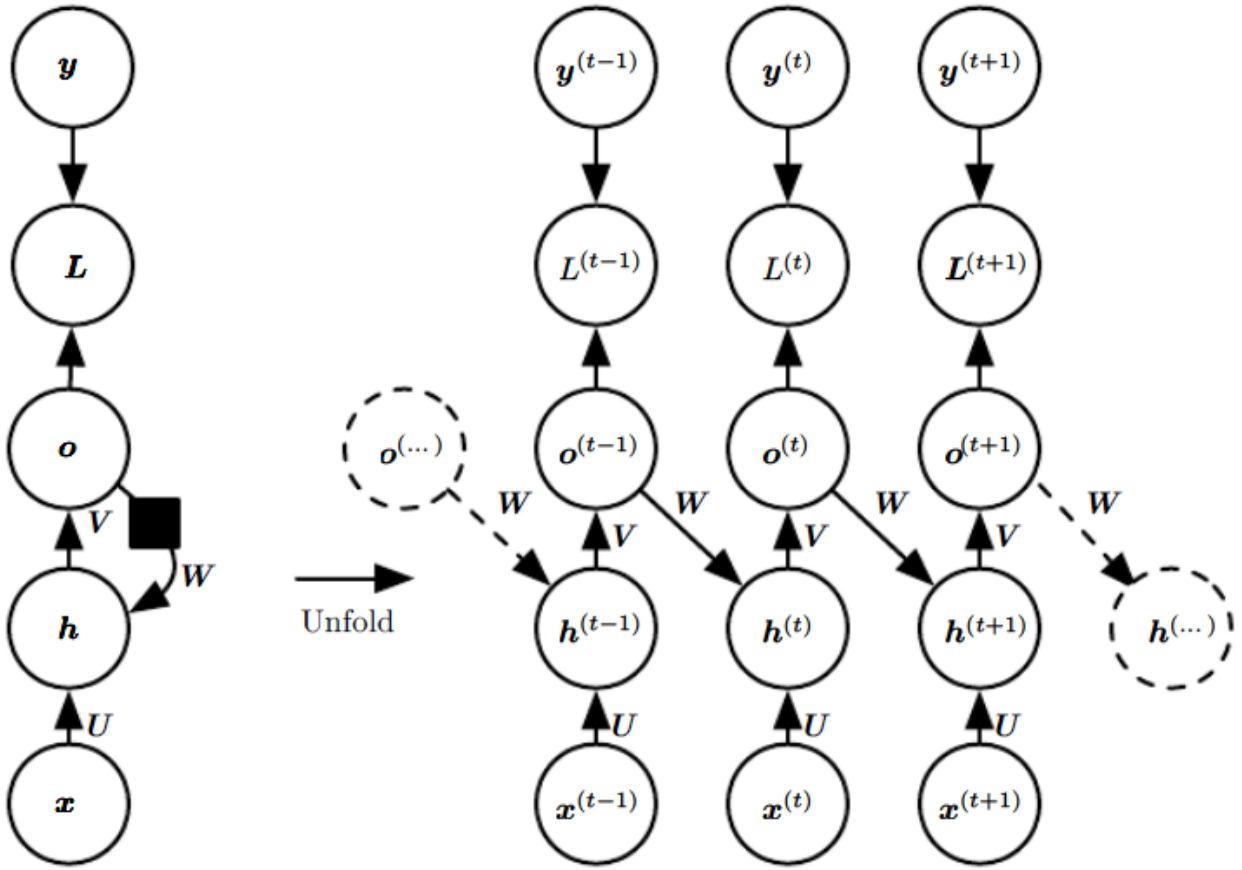
$$H_t = \phi_1(W_{xh}X_t + W_{hh}H_{t-1} + b_h)$$

$$y_t = \phi_2(W_{hy}H_t + b_y)$$

$$\theta = \{W_{xh}, W_{hh}, W_{hy}, b_h, b_y\}$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Recurrent neural networks



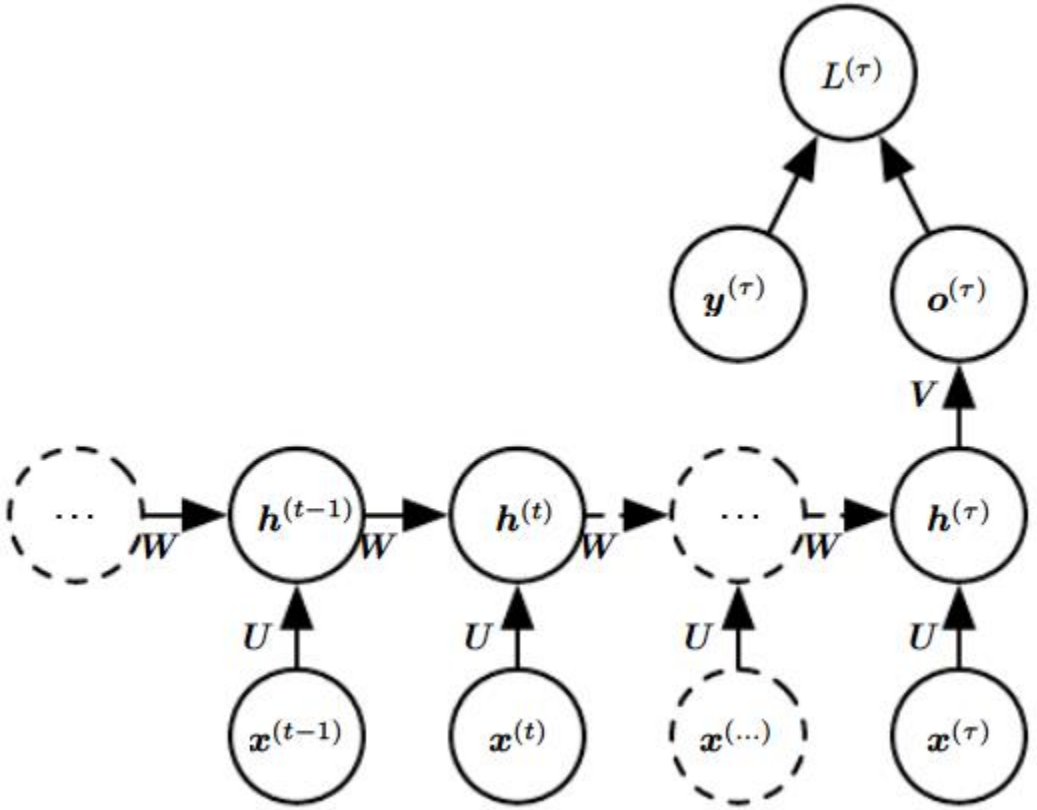
تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

dayche.com | گروه دایکه

Sequence to scaler – sentiment analysis



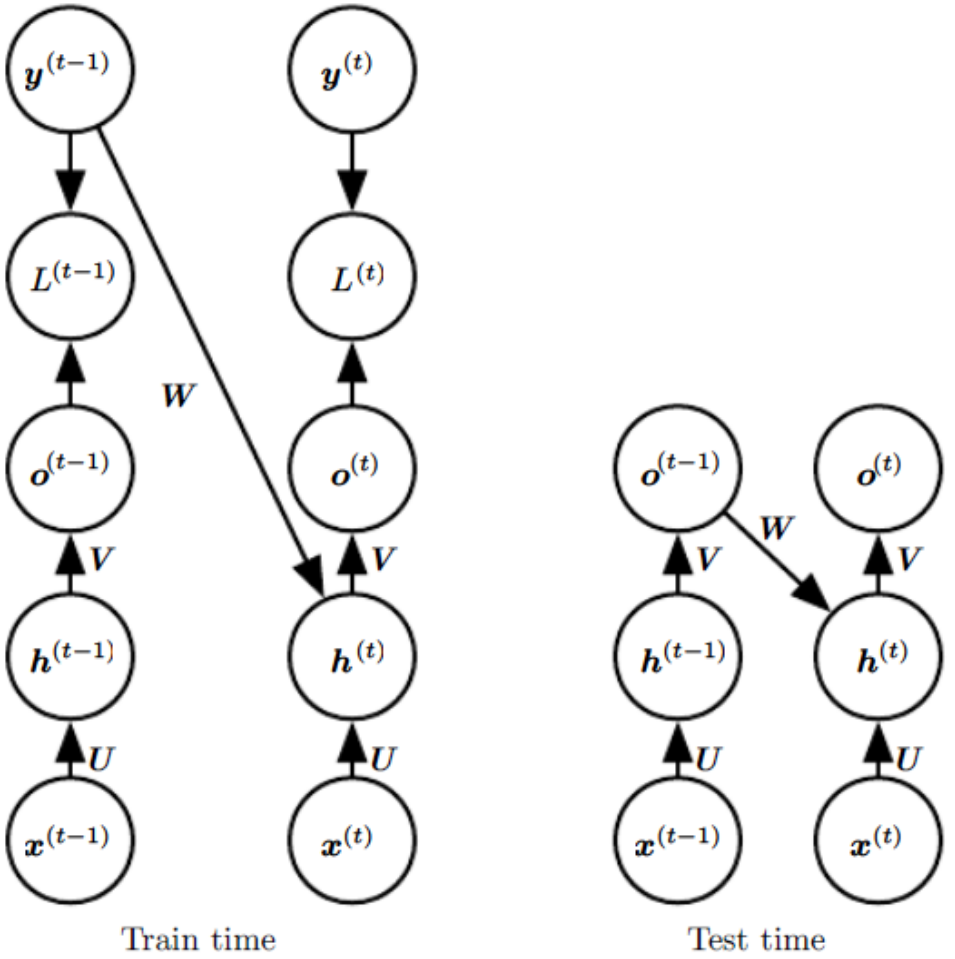
تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

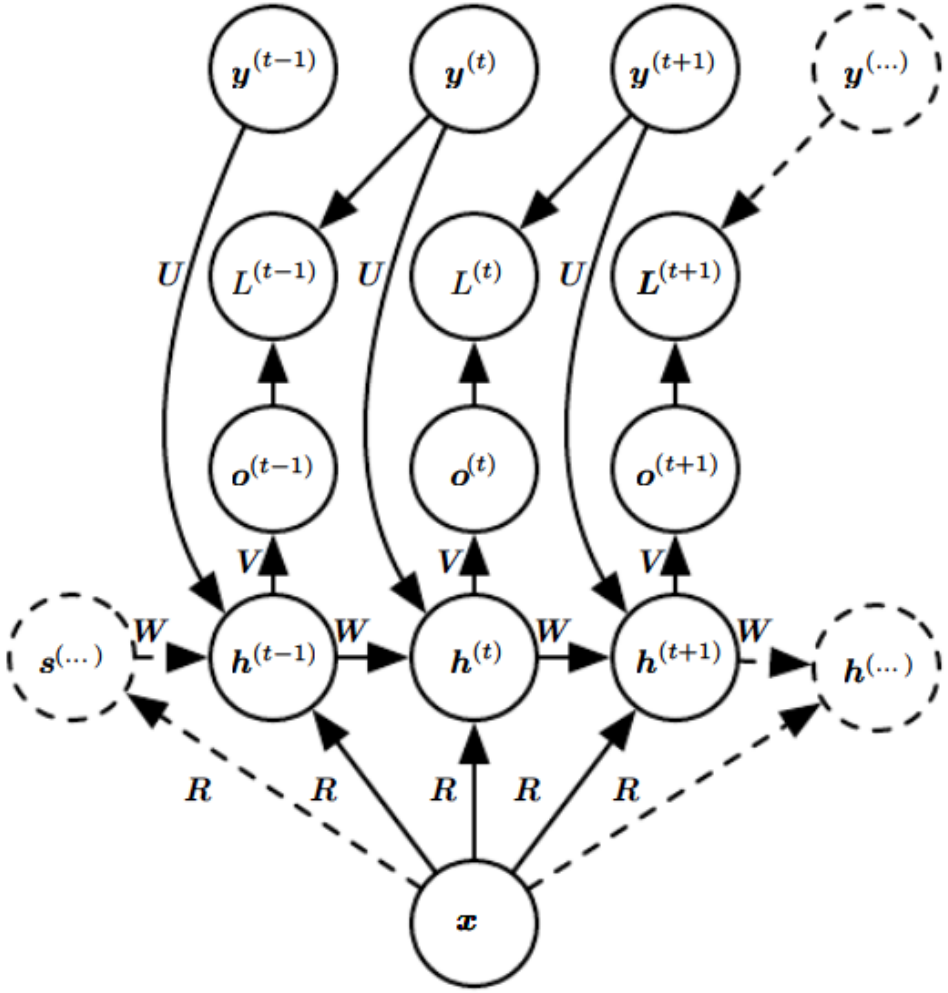
dayche.com | گروه دایکه

Teacher forcing network



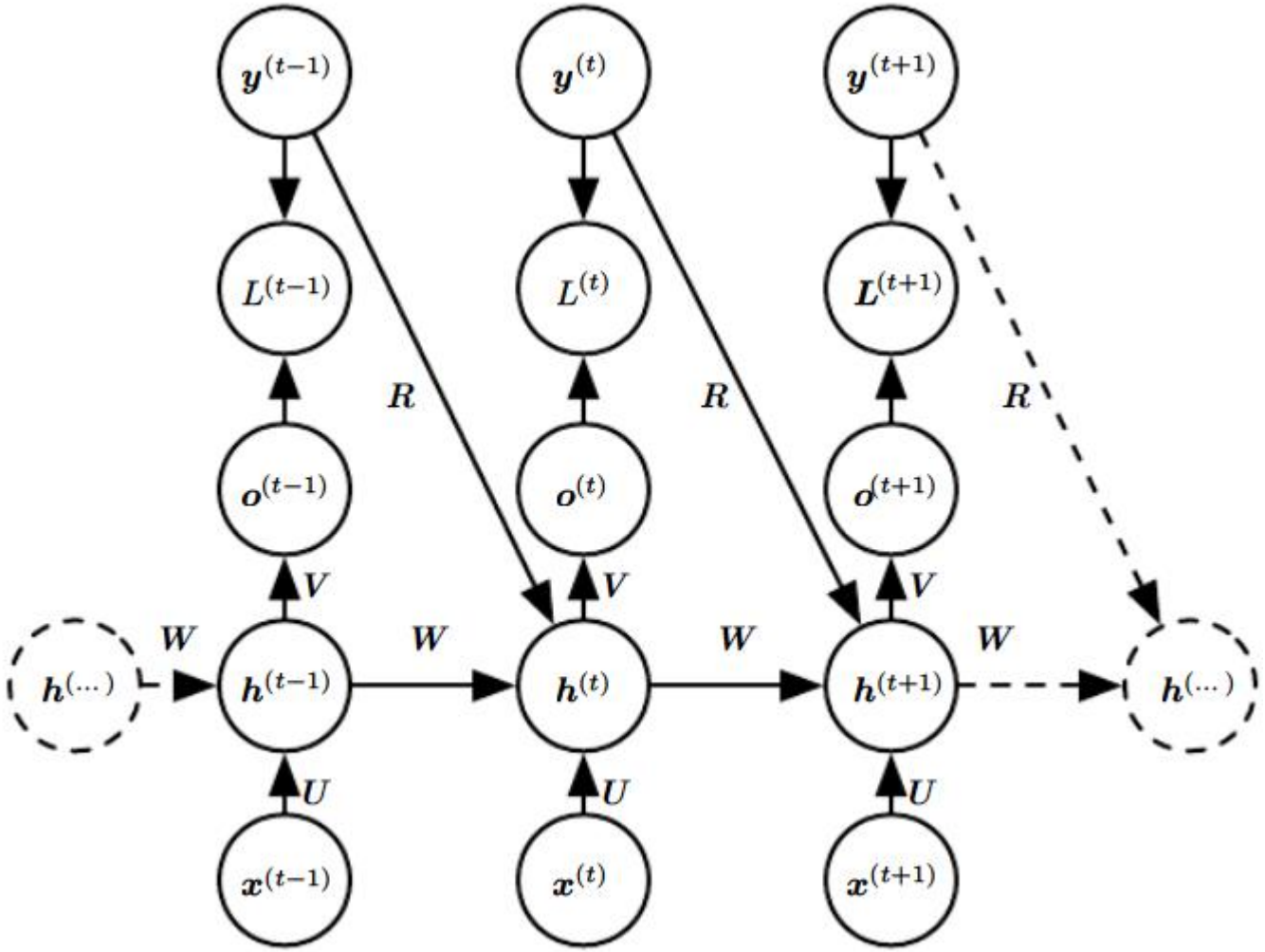
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Vector to sequence – image captioning



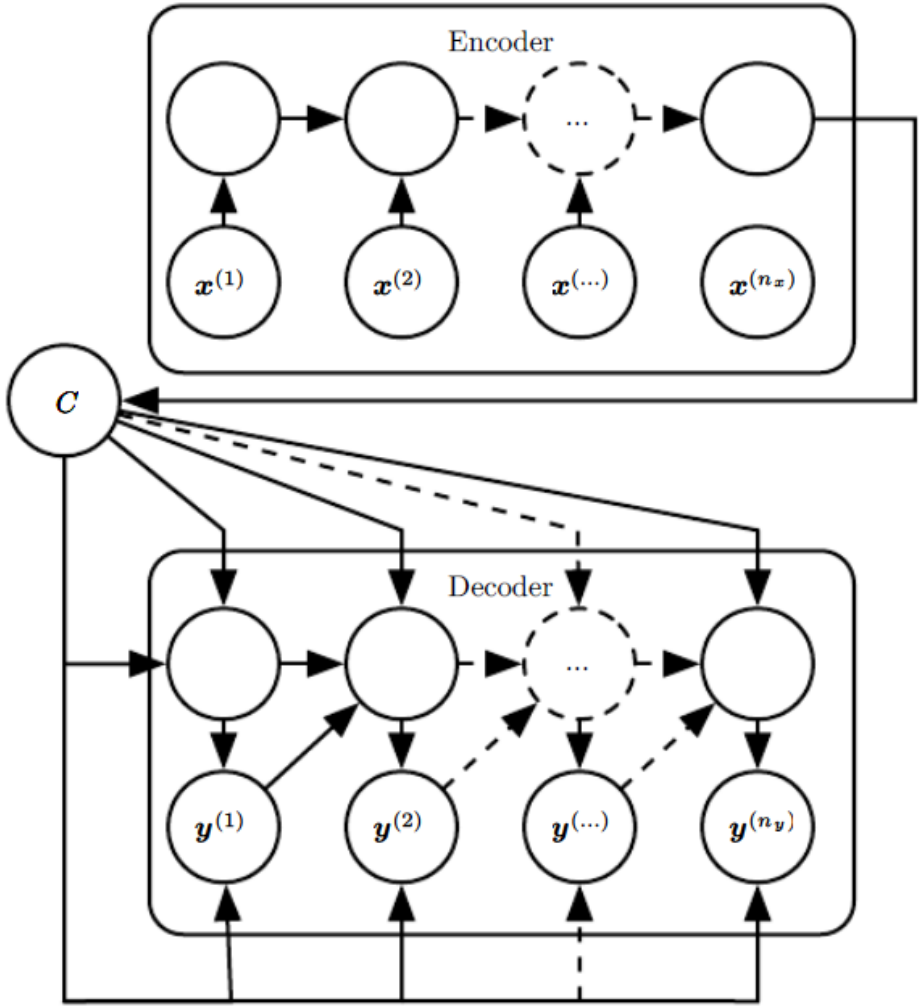
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Sequence to sequence - speech recognition

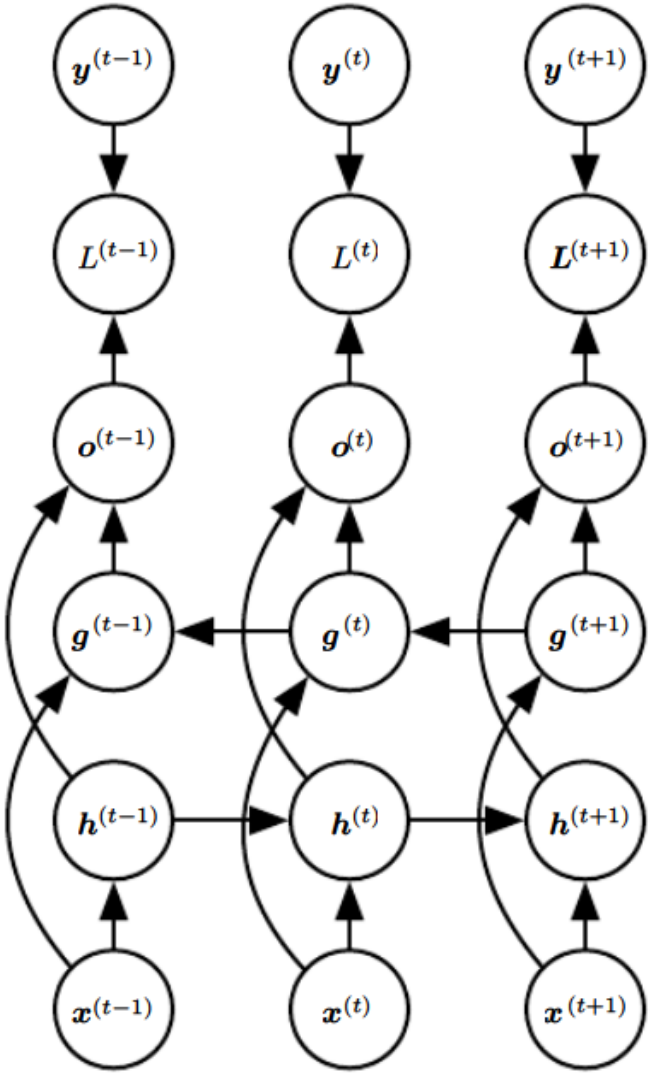


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Encoder-decoder RNN – Machine translation



Bidirectional RNN – handwritten recognition



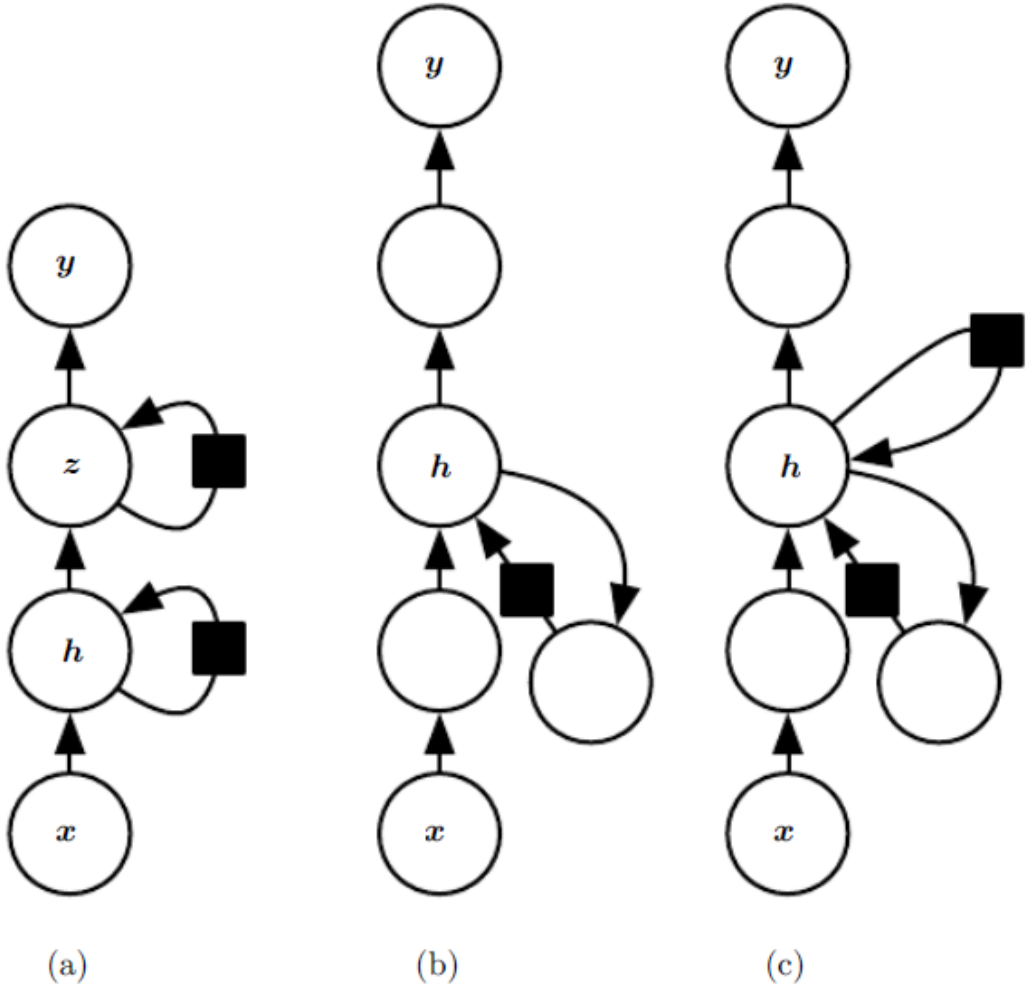
تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

dayche.com | گروه دایکه

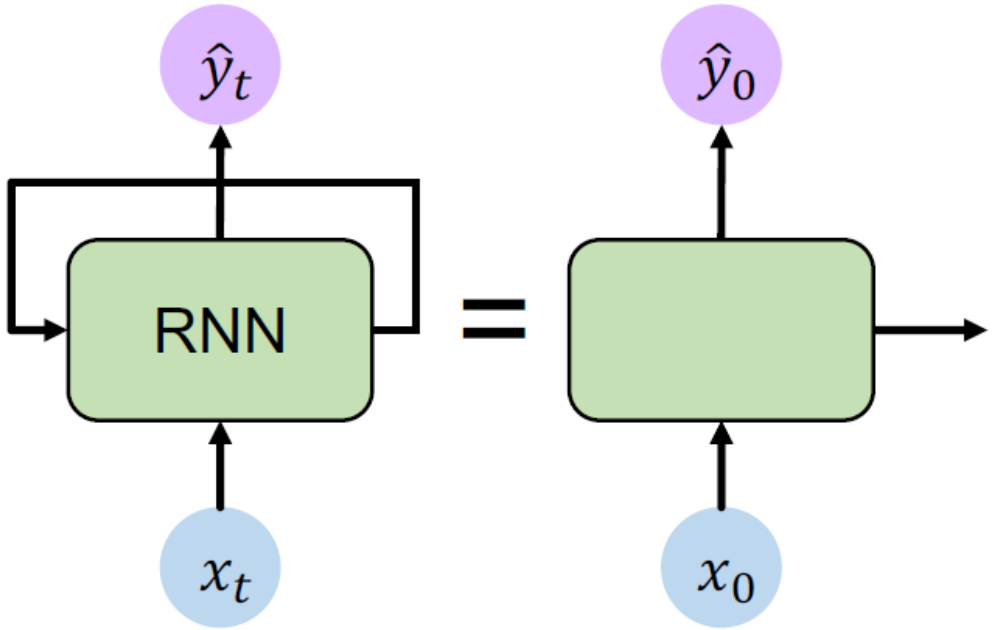
Deep recurrent network



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

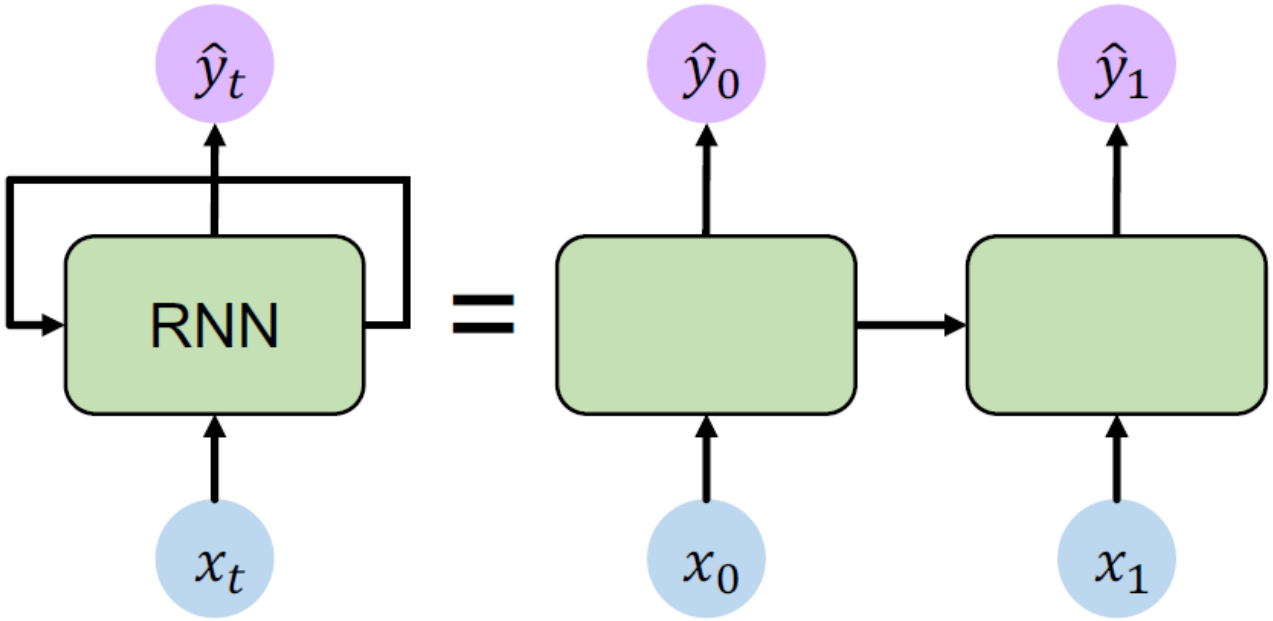
RNN training – Forward path

محاسبه گرادیان خطا توسط روش پس انتشار ممکن نیست، چرا؟
برای امکان استفاده از الگوریتم پس انتشار خطا شبکه به صورت زیر unfold می‌شود.



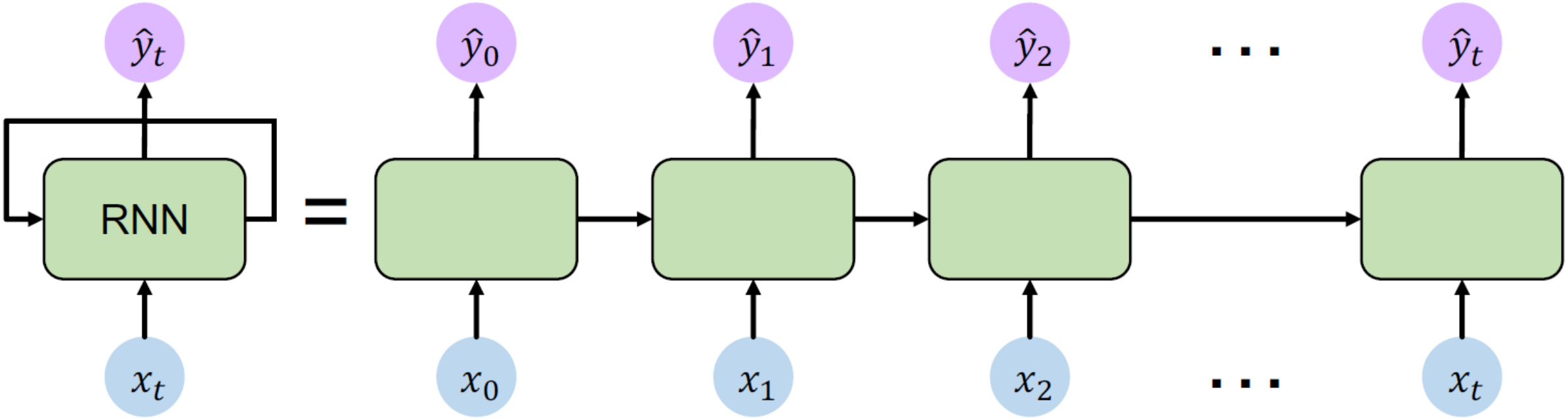
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

RNN training – Forward path



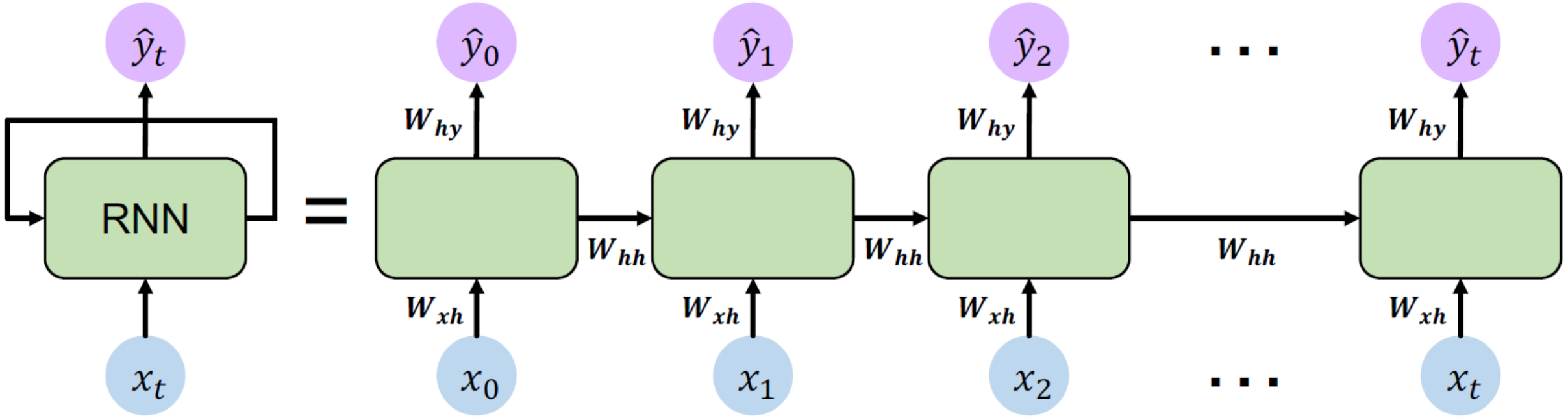
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

RNN training – Forward path



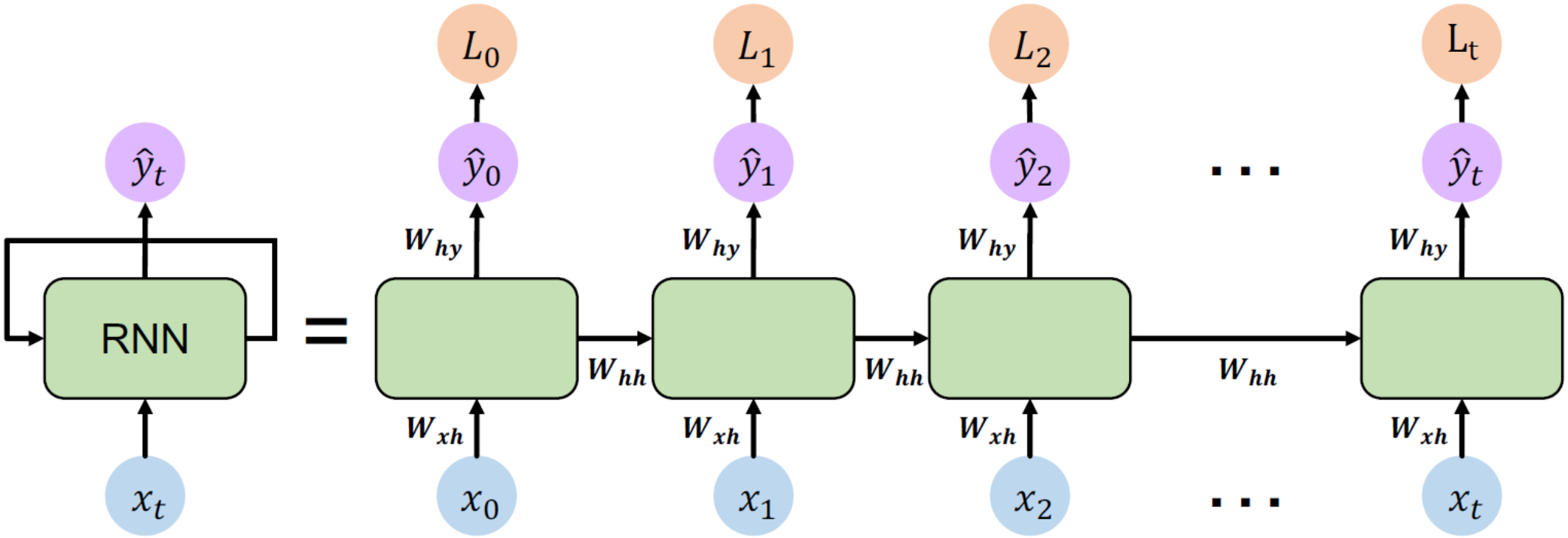
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

RNN training – Forward path



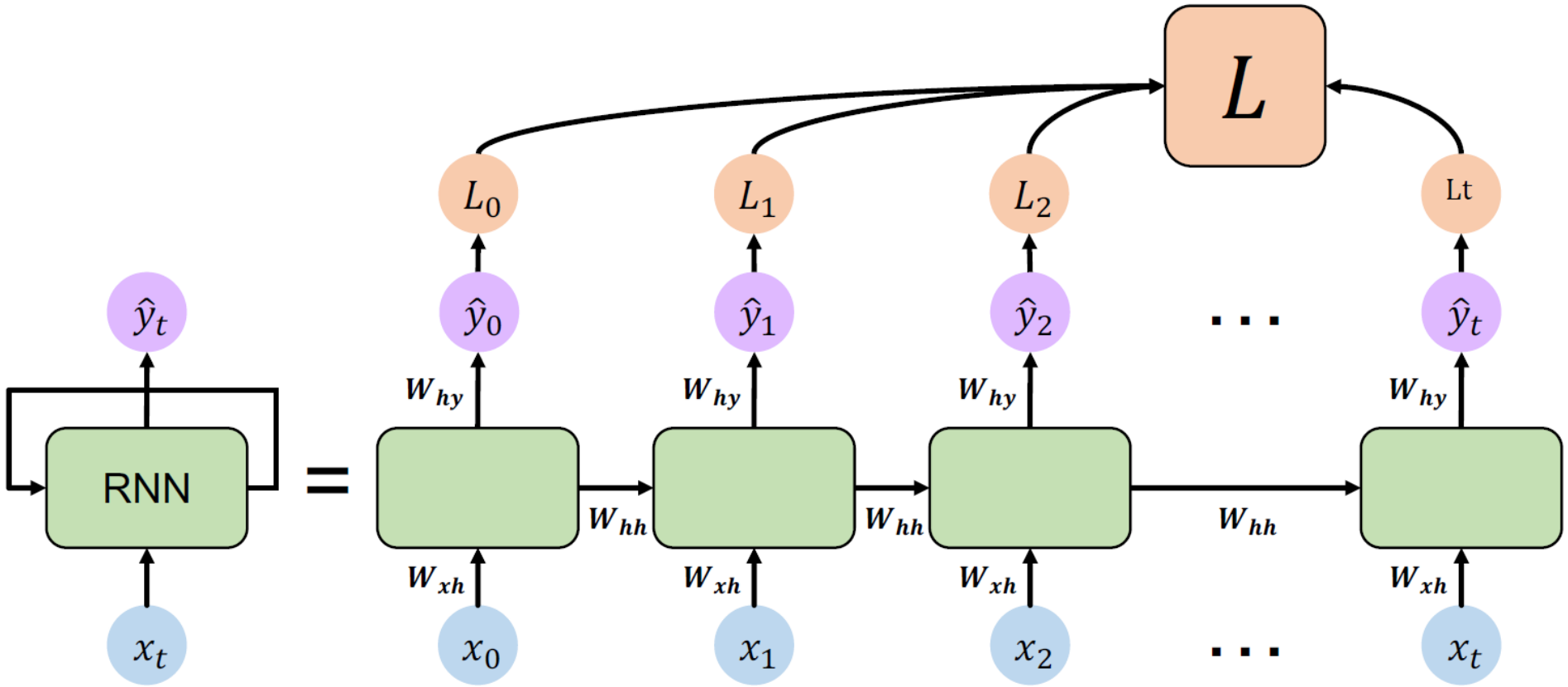
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

RNN training – Forward path



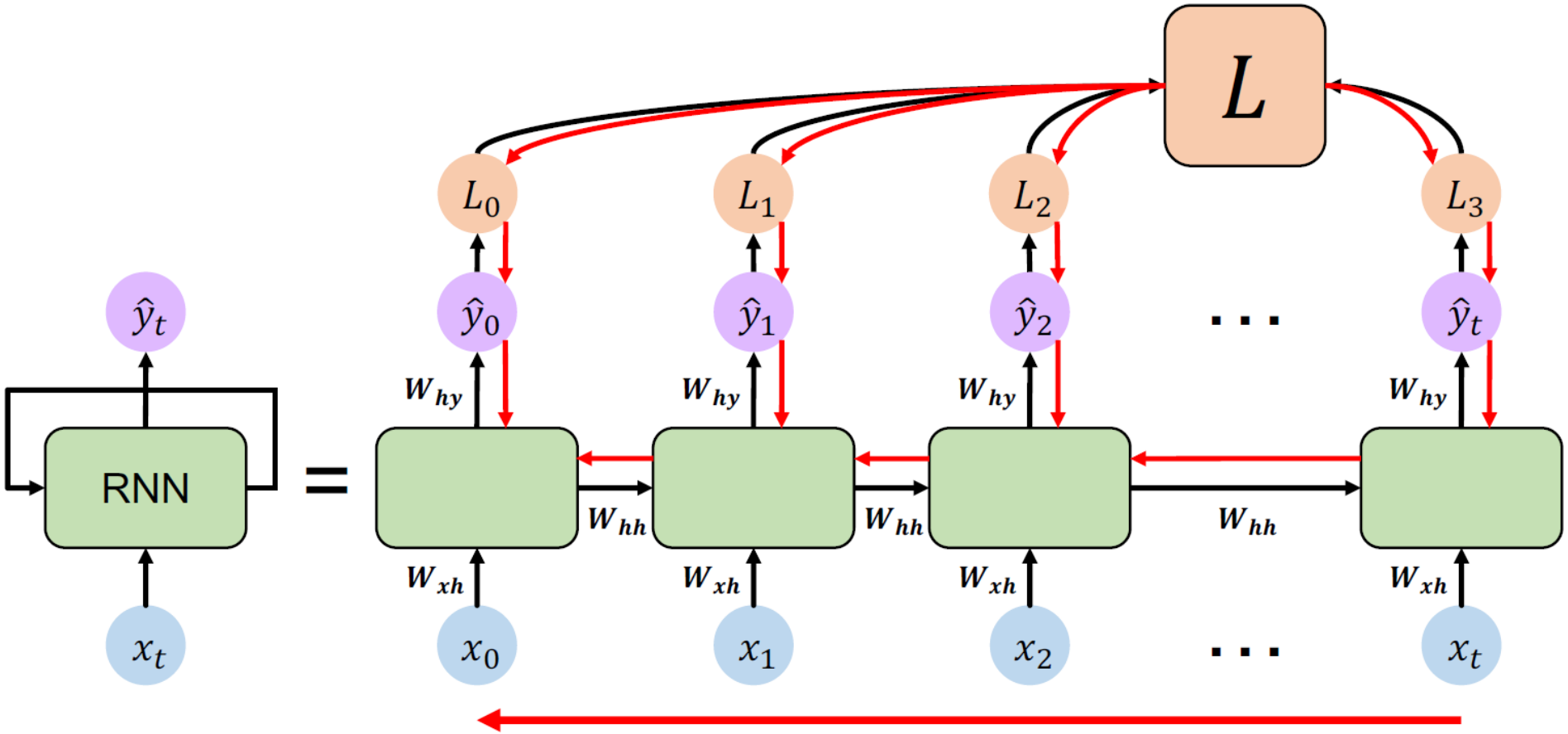
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

RNN training – Forward path



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

RNN training – Backward path

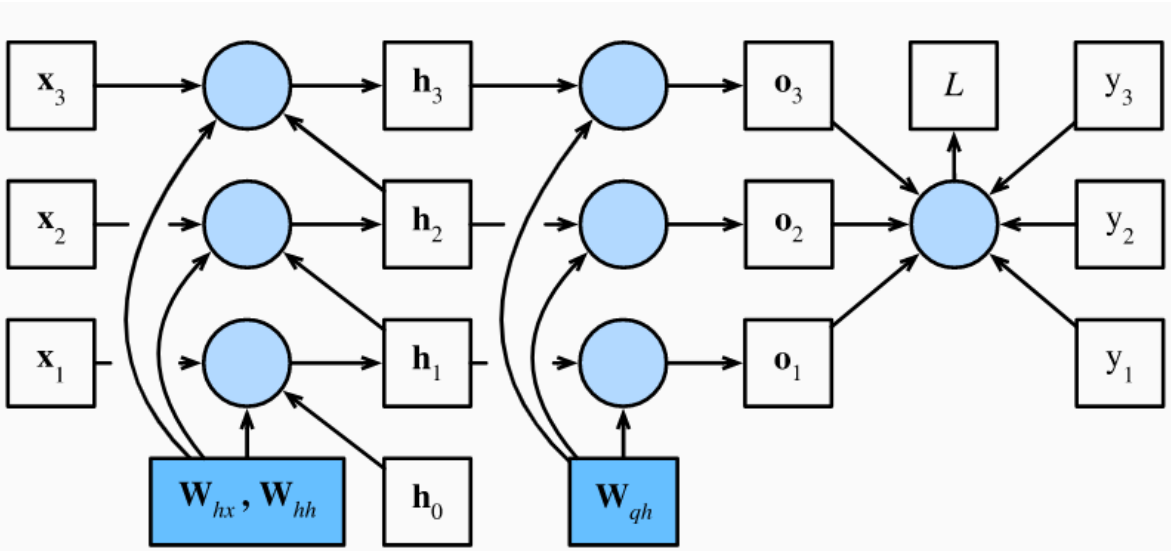


تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Backpropagation through time (BPTT)

پس انتشار خطا در طول زمان، کاربرد الگوریتم پس انتشار خطا برای یادگیری پارامترهای یک نگاشت دینامیکی است.

Computational graph



$$h_t = f(h_{t-1}, x_t; W)$$
$$o_t = g(h_t; W_o)$$

$$L = \frac{1}{T}\sum_{t=1}^{T} L_t(o_t, y_t)$$

$$\frac{\partial L}{\partial W_o} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial L_t(o_t, y_t)}{\partial o_t} \times \frac{\partial o_t}{\partial W_o}$$

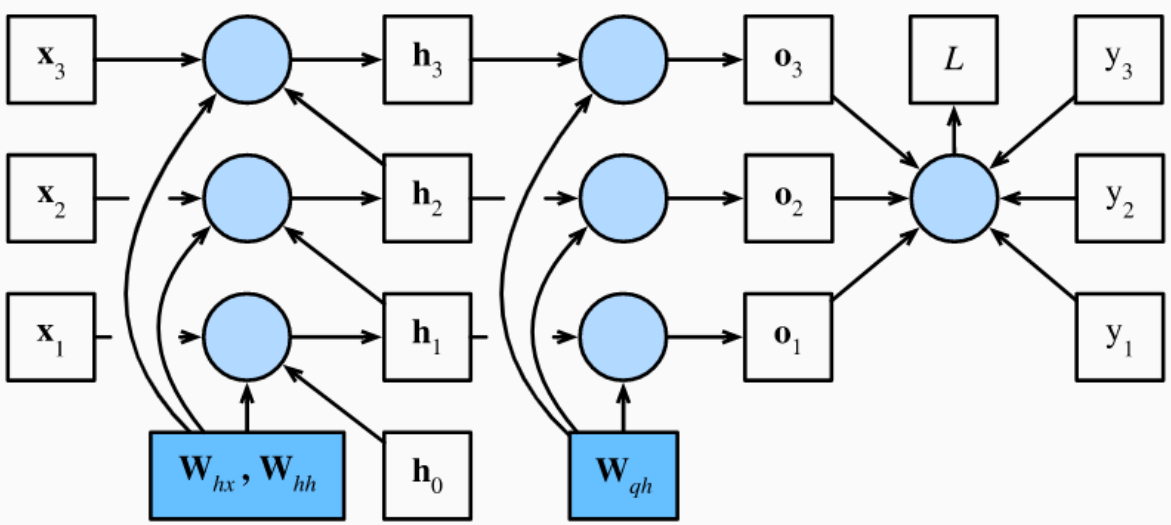
با فرض خطی بودن لایه خروجی خواهیم داشت:

$$\frac{\partial L}{\partial W_o} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial L_t(o_t, y_t)}{\partial o_t} h_t$$

تولید محتوا: وحید محمدزاده ایوقی

Backpropagation through time (BPTT)

Computational graph



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

$$o_t = W_o h_t$$

$$L = \frac{1}{T}\sum_{t=1}^{T} L_t(o_t, y_t)$$

برای محاسبه گرادیان خطا نسبت به وزن‌های لایه مخفی باید حساسیت تابع هزینه را نسبت به خروجی لایه مخفی محاسبه کنیم.

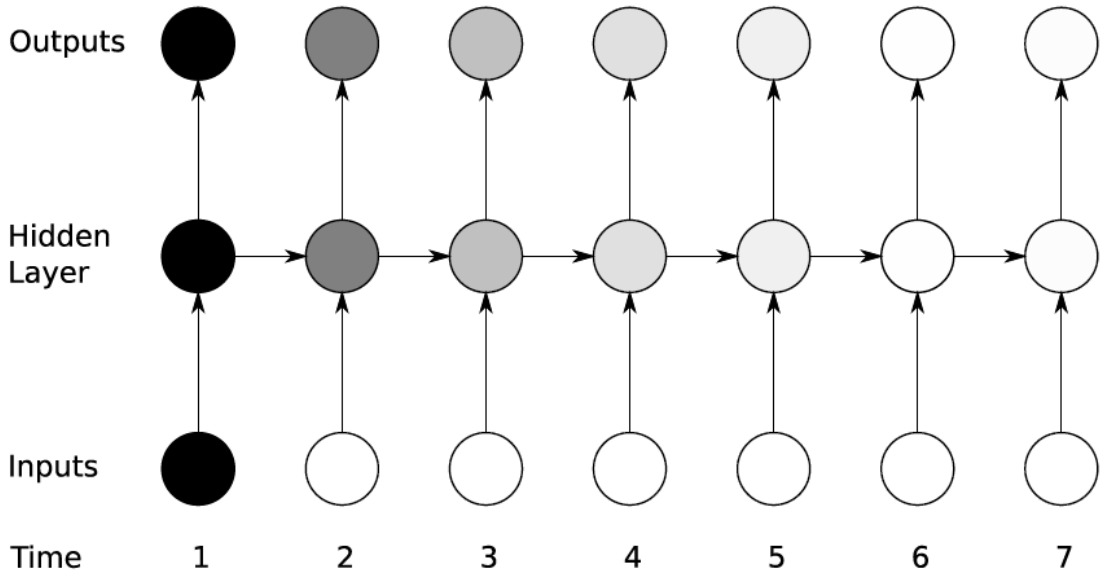
$$L = \frac{1}{T}\{L_1(o_1, y_1) + \cdots L(o_t, y_t) + \cdots + L(o_T, y_T)\}$$

$$\frac{\partial L}{\partial h_T} = \frac{\partial L}{\partial o_T} \times \frac{\partial o_T}{\partial h_T} = W_0^T \frac{\partial L}{\partial o_T}$$

$$\frac{\partial L}{\partial h_{T-1}} = \frac{\partial L}{\partial o_T} \times \frac{\partial o_T}{\partial h_T} \times \frac{\partial h_T}{\partial h_{T-1}} + \frac{\partial L}{\partial o_{T-1}} \times \frac{\partial o_{T-1}}{\partial h_{T-1}} = W_{hh}^T W_0^T \frac{\partial L}{\partial o_T} + W_0^T \frac{\partial L}{\partial o_{T-1}}, \ldots\ldots$$

Gradient vanishing

$$h_t = W_{hh}^T h_{t-1} \rightarrow h_t = \left(W_{hh}^T\right)^t h_0 = Q^T \lambda^t Q h_0$$



با گذر زمان، اثر ورودی‌ها بر روی خروجی از بین می‌رود و مدل قادر به لحاظ کردن وابستگی خروجی به ورودی‌های قبلی نیست.

راه حل چیست؟

- نوع توابع فعالساز
- مقداردهی اولیه
- تغییر ساختار مدل

تولید محتوا: وحید محمدزاده ایوقی

daychegroup

daychegroup

dayche.com | گروه دایکه

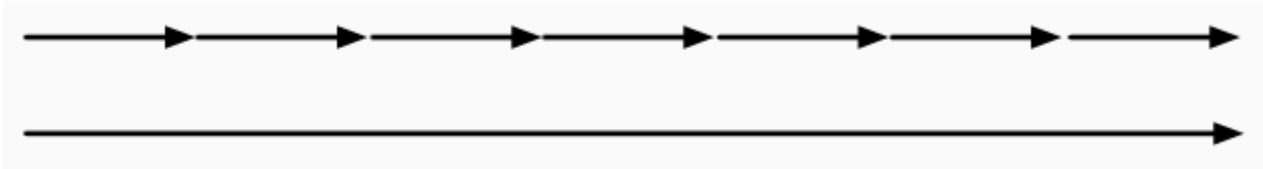
Truncated BPTT

$$\frac{\partial L}{\partial h_t} = \sum_{i=t}^{T} (W_{hh})^{T-i} W_o^T \cdot \frac{\partial L}{\partial O_{T+t-i}}$$

- اگر مقادیر ویژه ماتریس وزن‌ها کوچکتر از یک باشد، این عبارت با بزرگ شدن طول دنباله، برای گام‌های اولیه زمانی به صفر میل خواهد کرد.
- اگر مقادیر ویژه ماتریس وزن‌ها بزرگت از یک باشد، این عبارت با بزرگ شدن طول دنباله، برای گام‌های اولیه زمانی به بینهایت میل خواهد کرد.
- در حالت کلی با افزایش طول دنباله امکان محاسبه گرادیان، حساسیت خروجی مدل نسبت به لایه مخفی، وجود ندارد زیرا بار محاسباتی بسیار بالاست.

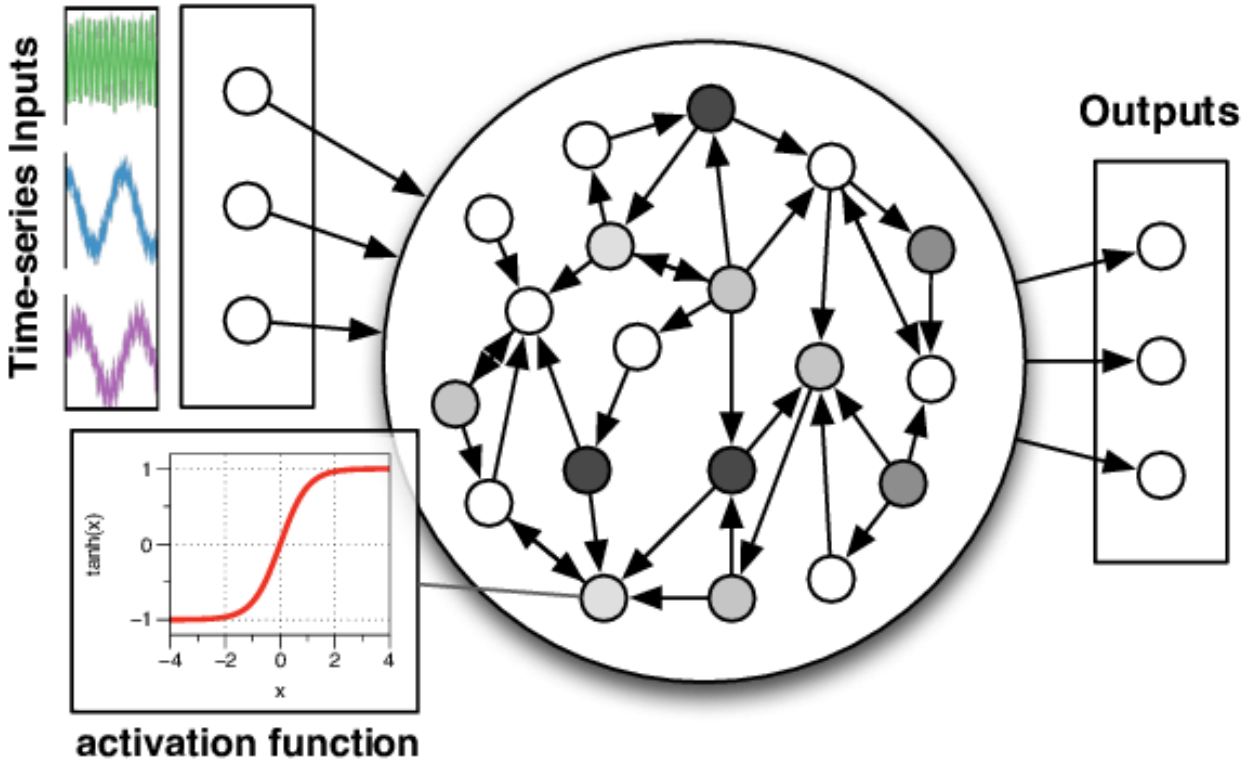
راه حل: برای یک طول مشخص و از پیش تعیین شده گرادیان را محاسبه و وزن‌ها را به روز می‌کنیم.

- دیگر قادر به ردیابی وابستگی‌های بلندمدت نیستیم، چرا؟



تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Echo State Network (ESN)



- وزن‌های ورودی و لایه مخفی به صورت تصادفی در نظر گرفته می‌شوند و آموزش نمی‌بینند.
- ورودی به یک فضای با بعد بالا به صورت تصادفی نگاشت داده می‌شود.
- ورژن بدون فیدبک این شبکه، مدل Extreme learning machine است.
- مقاوم در برابر چندشاخگی و مناسب برای سری‌های زمانی آشوبناک

هیچ کنترلی بر روی نوع نمایش بدست آمده نیست!

تولید محتوا: وحید محمدزاده ایوقی

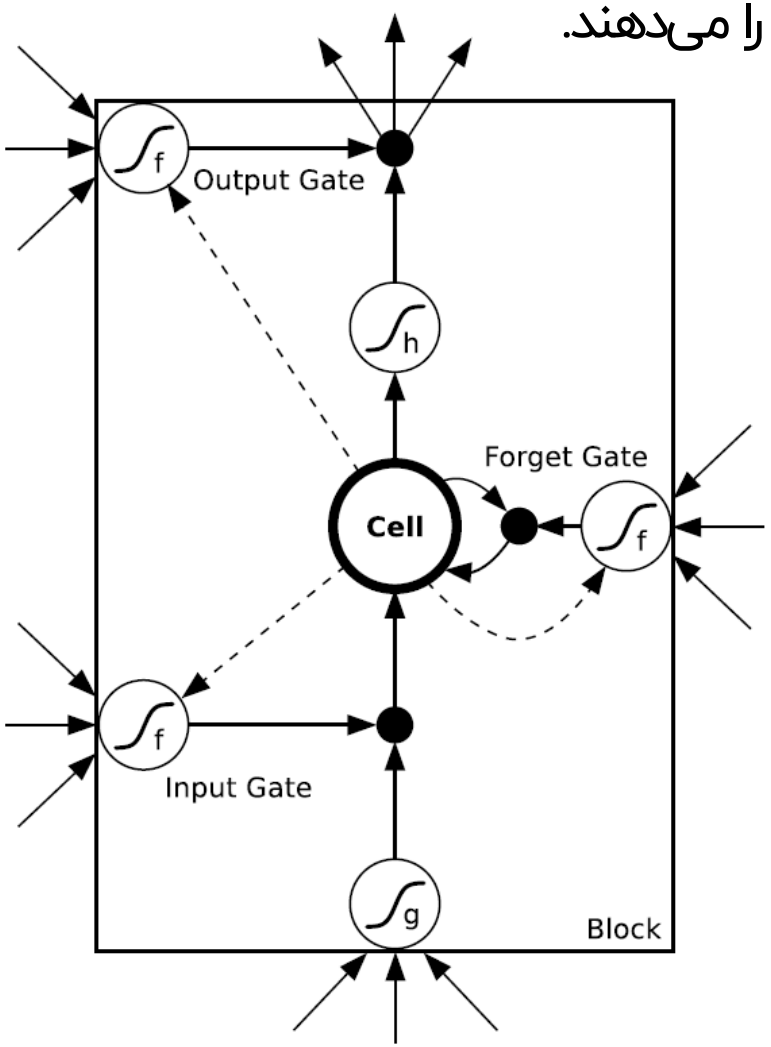
Practical issues with gradient vanishing

- مشکلاتی که محو شدگی گرادیان می‌تواند ایجاد کند، به صورت جزئی، شامل چه مواردی می‌شود؟
  - در برخی از تسک‌ها، اهمیت اولین المان یک دنباله، و یا یک المان به خصوص، از سایر المان‌ها در پیش‌بینی و یا تصمیم‌گیری یک کمیت بالا باشد (Memory).
  - در برخی تسک‌ها نیاز است اطلاعات غیرمفید و کم‌اثر بر روی پیش‌بینی خروجی حذف شود (Forgetting).
  - در برخی تسک‌ها نیاز است تا به صورت کل ارتباط منطقی بین دو دنباله متفاوت از بین برود (Resetting).

سه موضوع یاد شده اصلی‌ترین مشکلاتی هستند که محوشدگی گرادیان در عمل به مسئله تحمیل می‌کند. بنابراین به منظور اطمینان از یک آموزش صحیح لازم است تا مدل طوری اصلاح شود تا موارد ذکر شده را شامل گردد

Long-short term memory (LSTM)

- چند شبکه به صورت بازگشتی بهم متصل شده و تشکیل یک بلوک حافظه را می‌دهند.

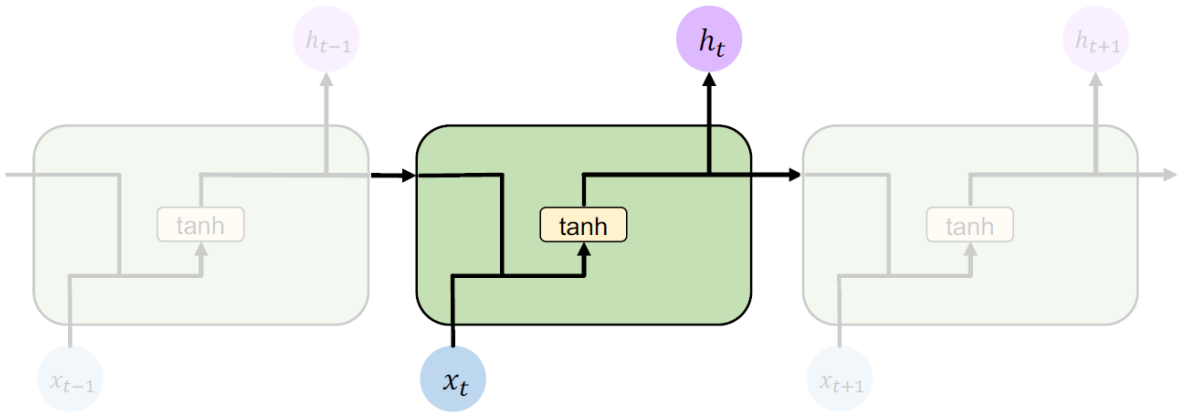


- یک یا چندین سلول حافظه به منظور نگهدار اطلاعات
- سه گیت ورودی، خروجی و فراموشی به منظور خواندن، نوشتن و پاک کردن
- پیاده‌سازی تراشه‌های حافظه کامپیوترهای دیجیتال

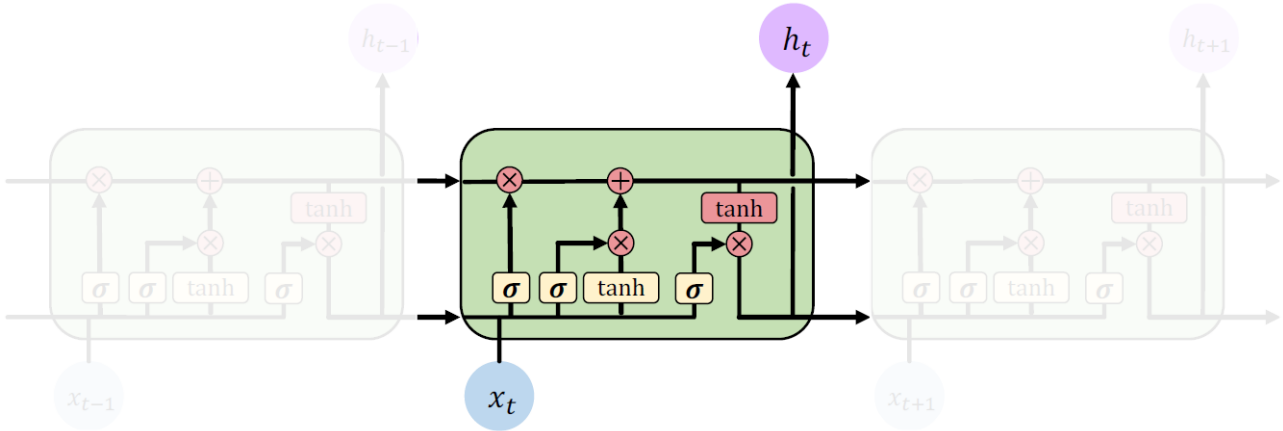
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

LSTM cell

ساختار یک سلول برگشتی استاندارد (Vanilla recurrent)



ساختار یک سلول LSTM



تولید محتوا: وحید محمدزاده ایوقی

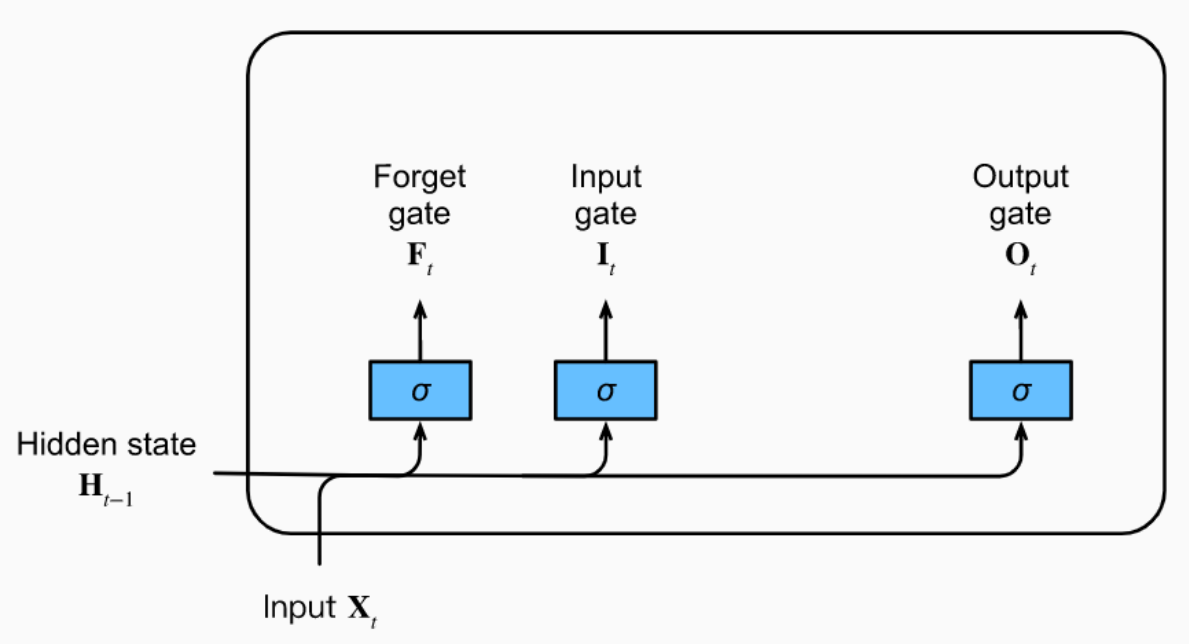
daychegroup

daychegroup

dayche.com | گروه دایکه

How does LSTM work?

سه شبکه عصبی با اتصال کامل وظیفه کنترل و ارزیابی کیفیت اطلاعات را دارند.

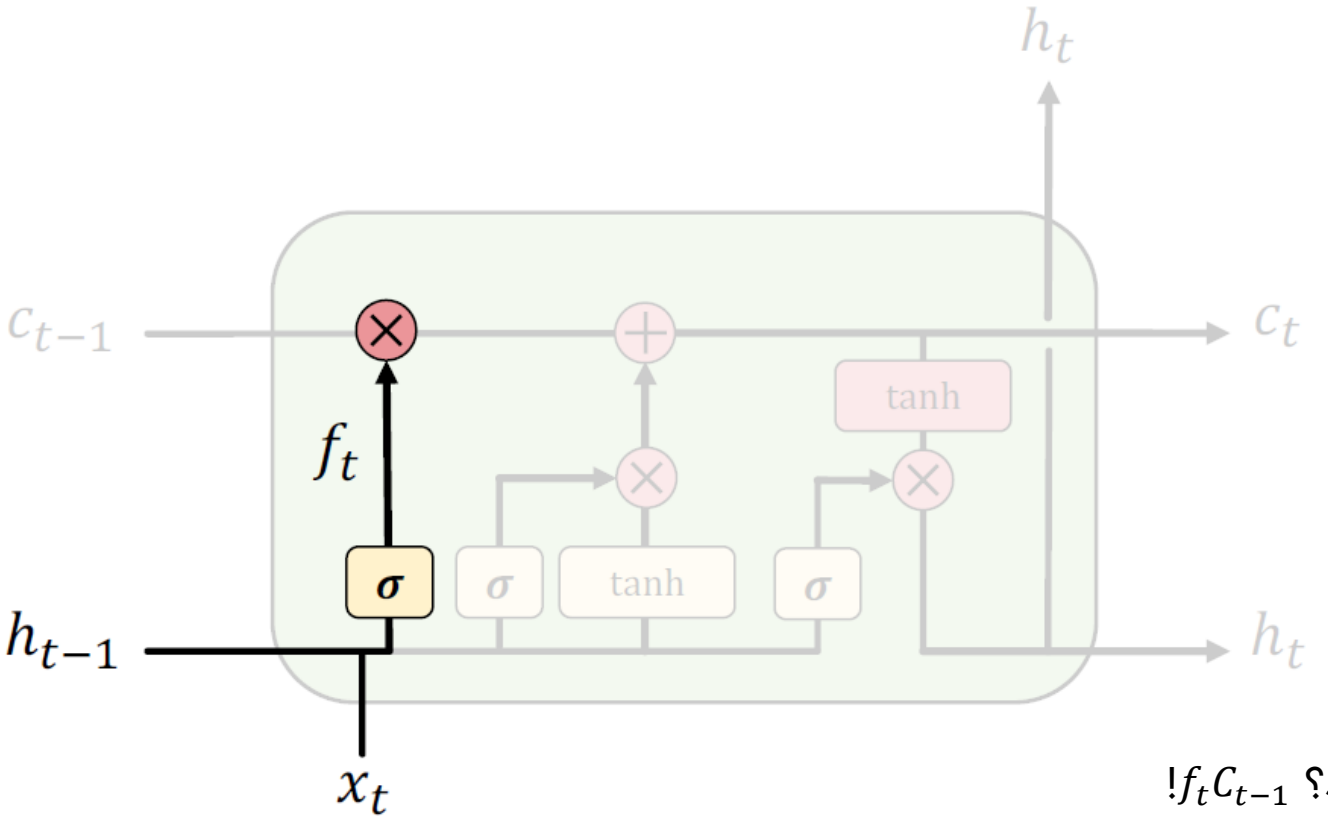


- هر نرون از لایه مخفی (بردار حالت) سطحی از وابستگی میان دنباله ورودی را یاد می‌گیرد.
- به ازای هر نرون، سه گیت یاد شده باید تعریف شود. بنابراین ابعاد خروجی هر گیت متناسب با ابعاد بردار حالت است.

بنابراین هر سلول LSTM دارای سه گام فراموشی، به روزرسانی و خروجی است.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Forget gate



- گیت فراموشی

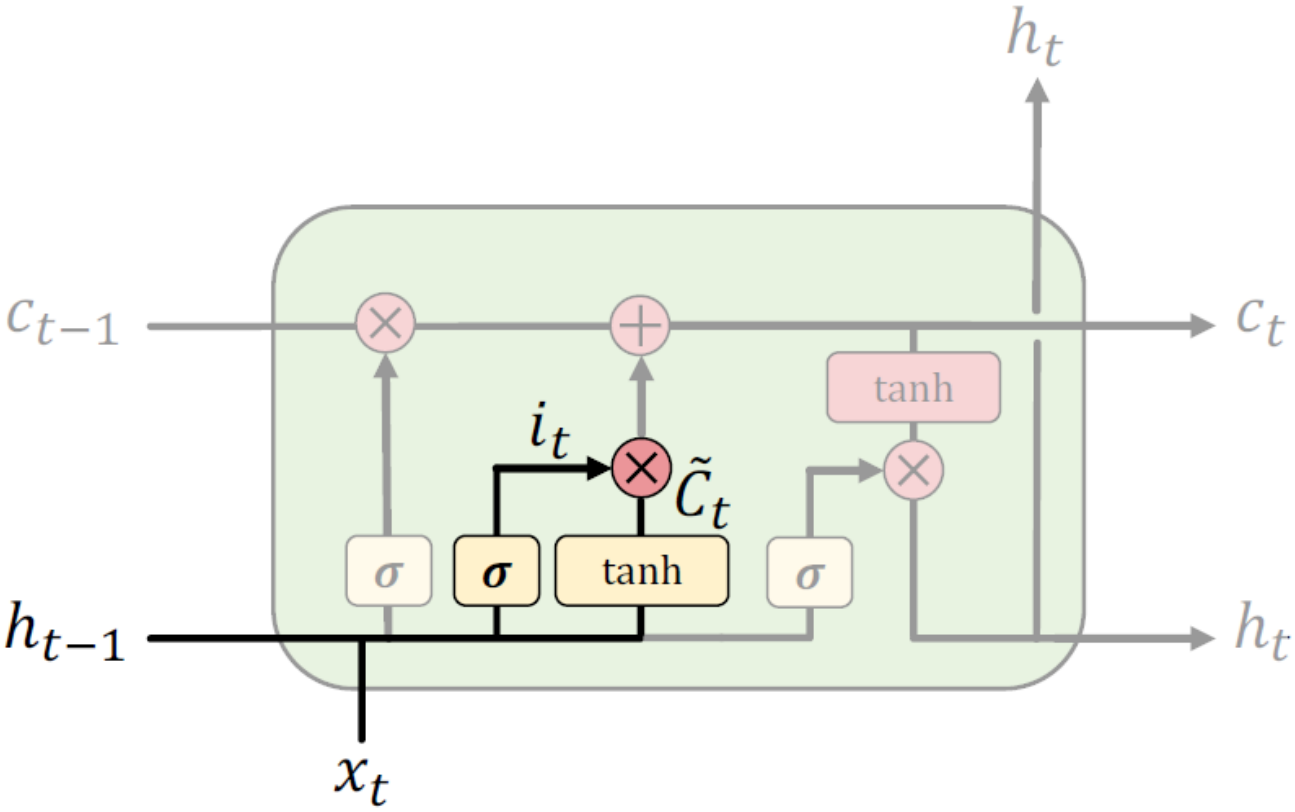
- گیت فراموشی بهنگام ورود به شبکه، تعیین می‌کند چه میزان از اطلاعات باید از بین برود، چگونه؟

$$f_t = \sigma\left(W_f[x_t; h_{t-1}] + b_f\right)$$

چه میزان از اطلاعات قبلی بایستی از سلول حافظه حذف گردند؟ $f_t C_{t-1}$!

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Input gate



- گیت ورودی

- گیت ورودی بهنگام ورود به شبکه، تعیین می‌کند چه میزان از اطلاعات جدید باید ذخیره گردد.

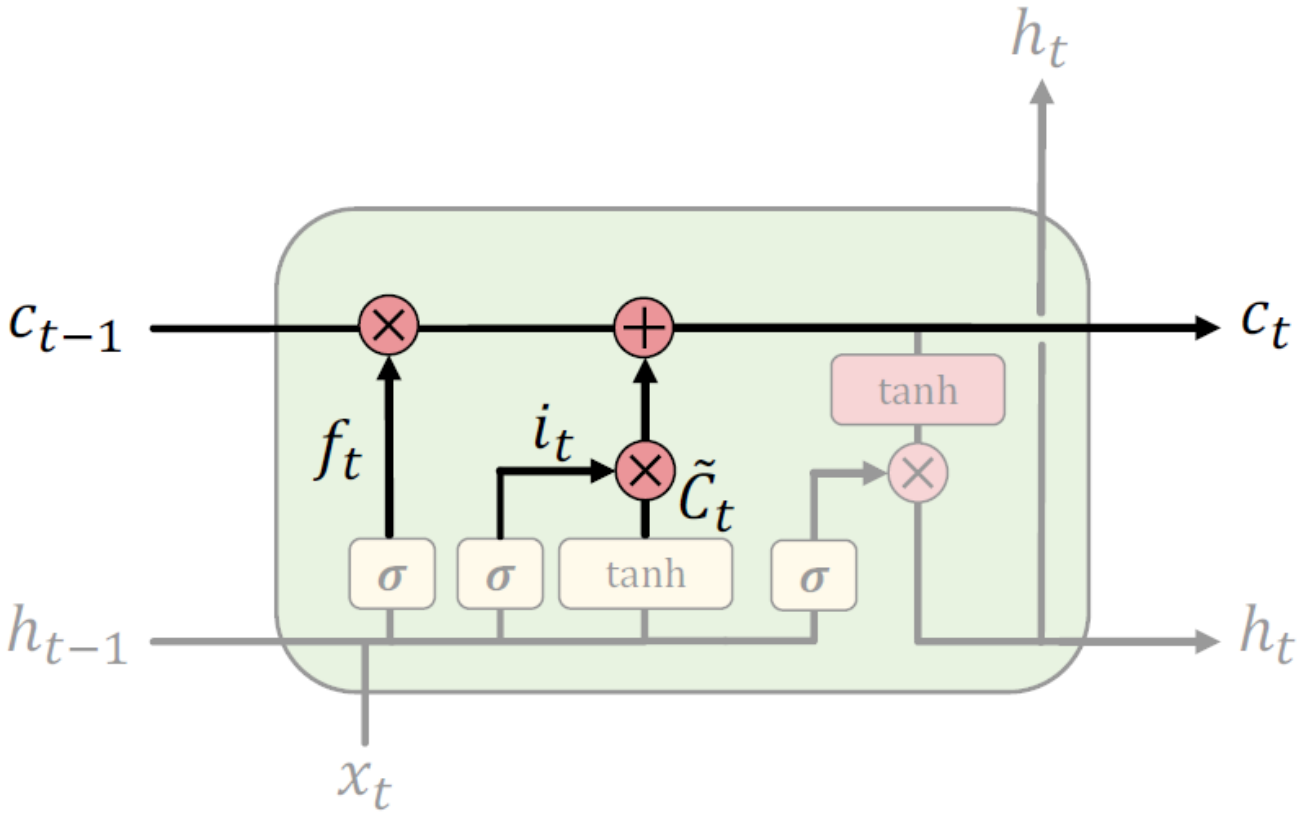
$$i_t = \sigma(W_i[x_t; h_{t-1}] + b_i)$$

چه میزان از اطلاعات فعلی باید ذخیر گردد؟ $i_t\tilde{C}_t$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Update cell state

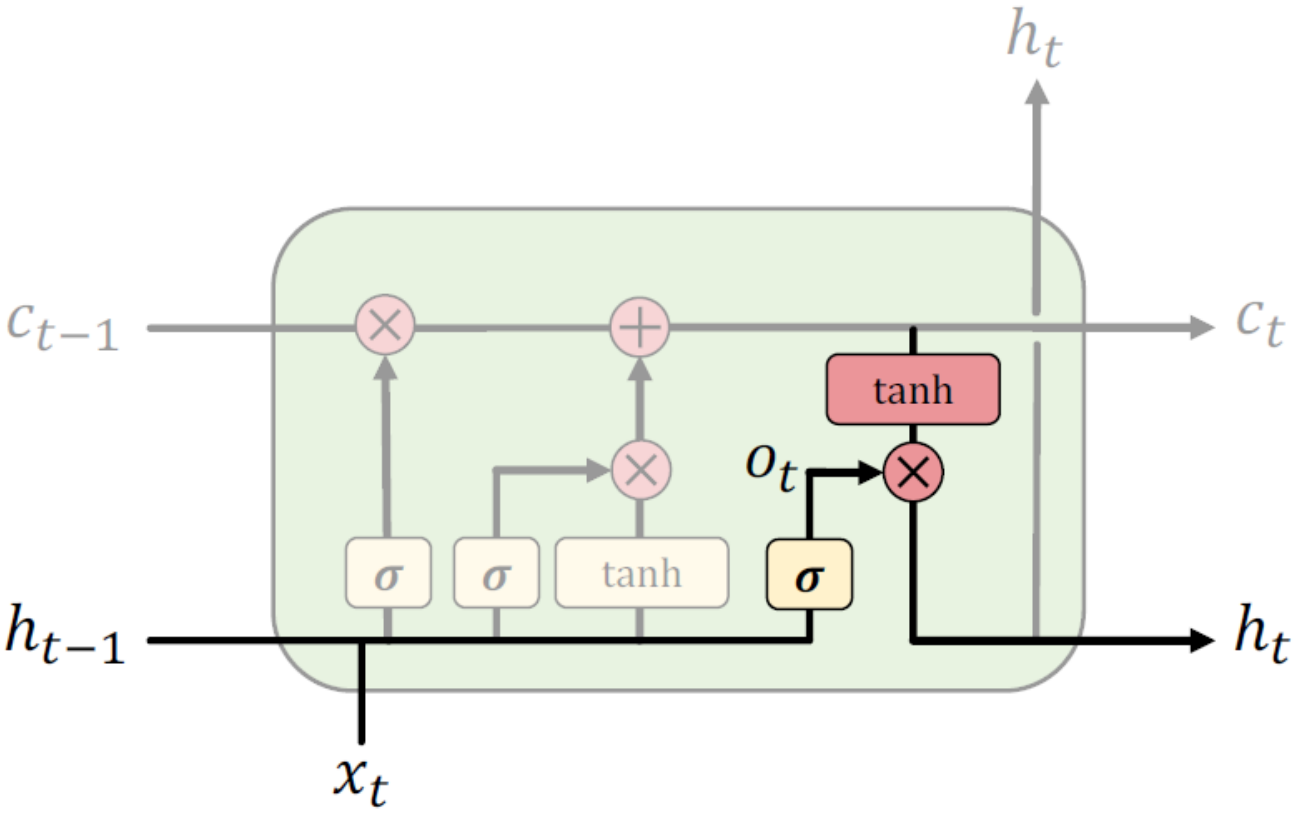
- سلول حافظه



$$C_t = i_t \odot \tilde{C}_t + f_t \odot C_{t-1}$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
گروه دایکه | dayche.com

Output gate



- گیت خروجی

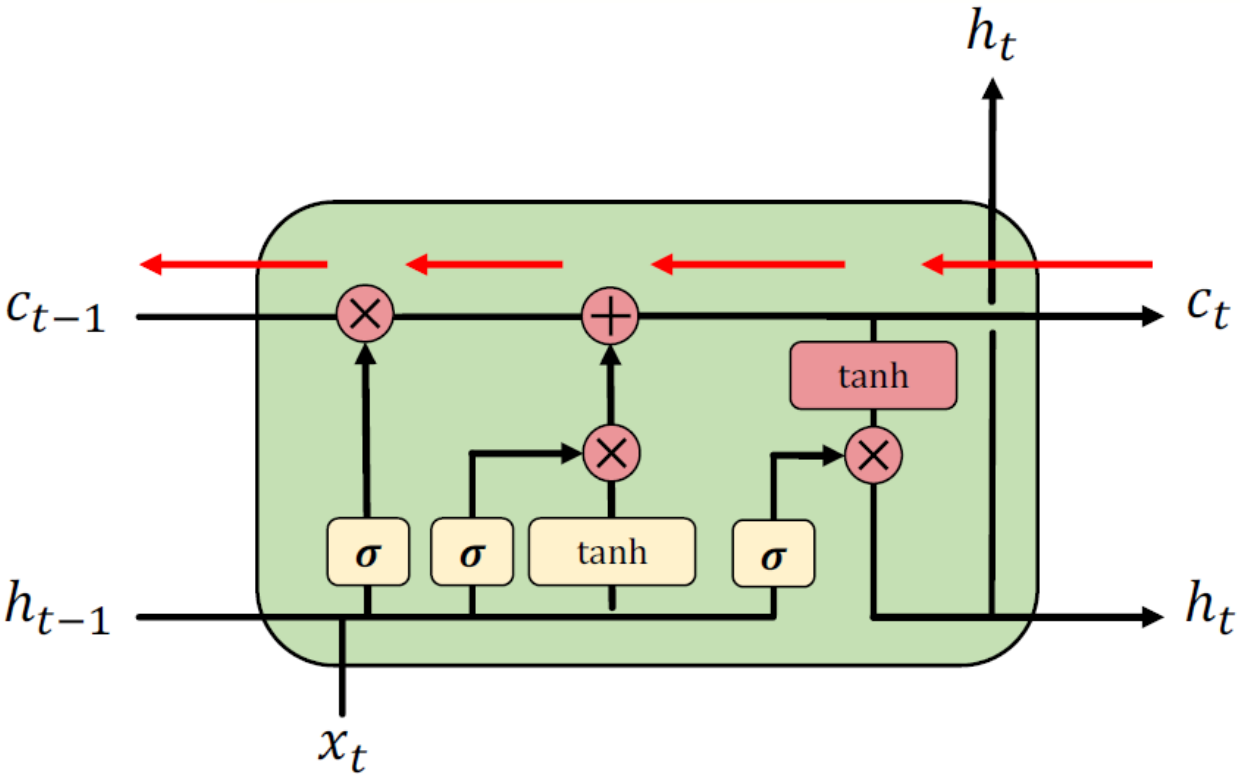
- گیت خروجی تعیین می‌کند که چه مقداری از حافظه قابل دسترسی باشد.

$$o_t = \sigma(W_o[x_t; h_{t-1}] + b_o)$$

$$\boldsymbol{h_t = o_t \odot \tanh(C_t)}$$

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

LSTM gradient flow



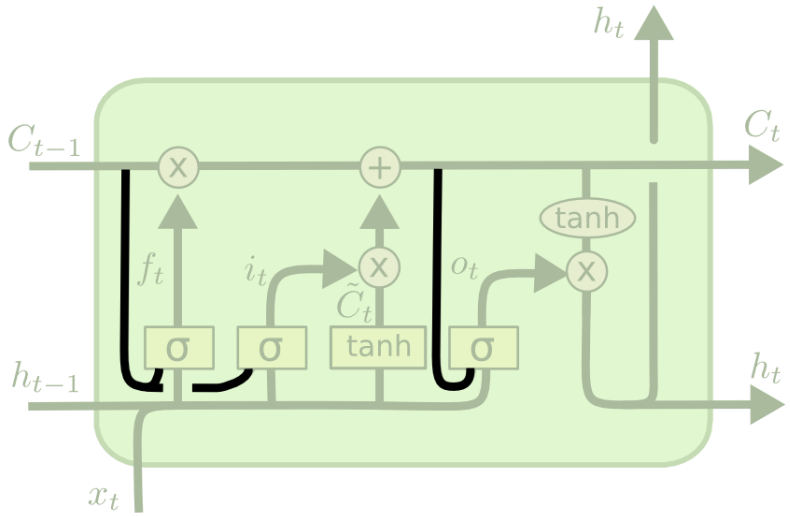
- الگوریتم از یک مقدار دهی اولیه (معمولا 0) برای سلول حافظه و متغیر حالت شروع به کار می‌کند.
- برای هر نرون لایه مخفی (هر بعد بردار حالت) این روابط تعریف می‌شود که تعریف یکپارچه آن به صورت رو به رو در آمده است.
- ساختار داده شده چگونه باعث جلوگیری از محوشدگی گرادیان می‌شود؟
  - دلیل اصلی محوشدگی در گرادیان در Vanilla recurrent

$$h_t = W_{hh}^T h_{t-1} \rightarrow h_t = \left(W_{hh}^T\right)^t h_0 = Q^T \lambda^t Q h_0$$

ضرب ماتریسی تبدیل به یک ضرب نقطه به نقطه شده است. بنابراین گرادیان برای هر واحد کوچک نخواهد شد.

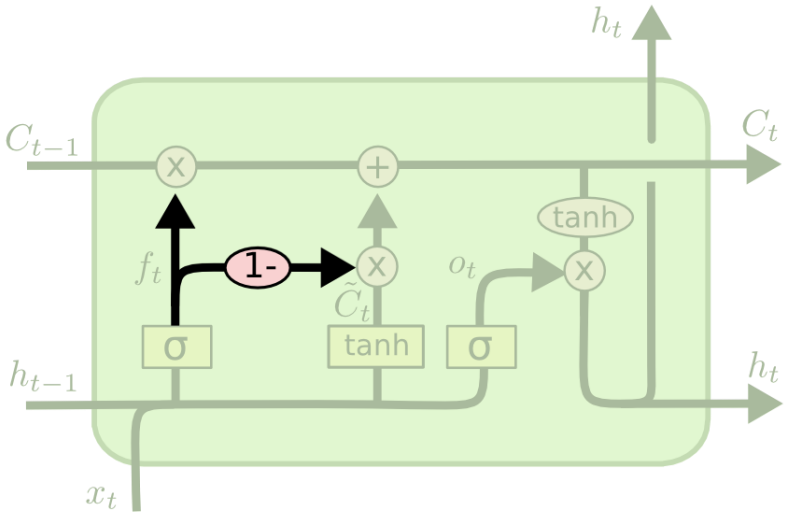
تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه

Variants of LSTM



Peephole connection

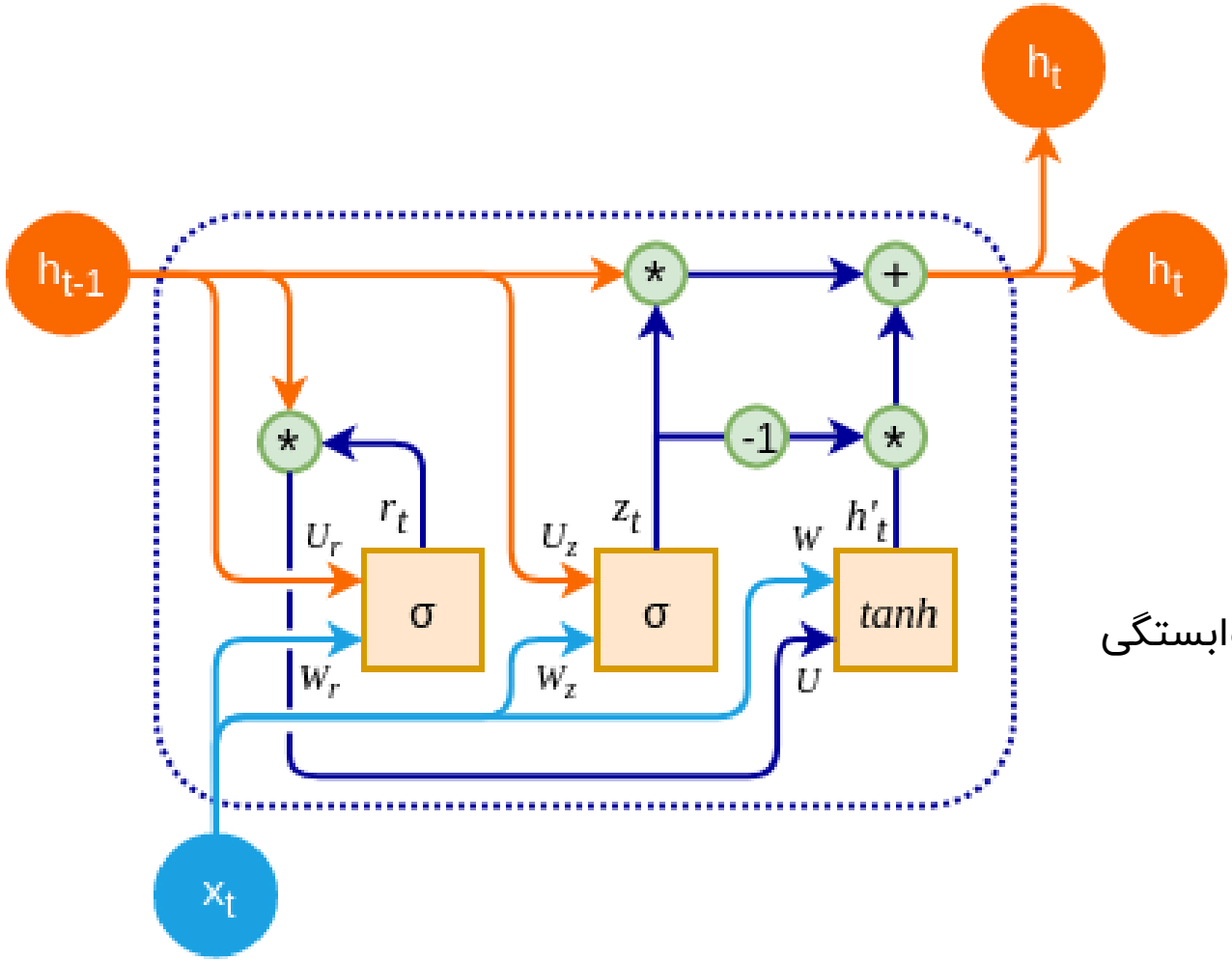
$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] \; + \; b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] \; + \; b_o\right)$$



Merged input and forget gate

$$C_t = f_t * C_{t-1} + (1 - \boldsymbol{f_t}) * \tilde{C}_t$$

Gated recurrent unit (GRU)



$$z_t = \sigma(W_z x_t + U_z h_{t-1}), \qquad r_t = \sigma(W_r x_t + U_r h_{t-1})$$

$$\tilde{h}_t = \tanh(W x_t + U h_{t-1} \odot r_t)$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

در این ساختار، گیت Reset وابستگی کوتاه مدت و گیت update وابستگی طولانی مدت را مدل می‌کند.

تولید محتوا: وحید محمدزاده ایوقی
daychegroup
daychegroup
dayche.com | گروه دایکه