

ملیکا محمدی فخار

۹۹۵۲۲۰۸۶

منابع مورد استفاده:

<https://stats.stackexchange.com/questions/386535/why-cant-we-use-back-propagation-in-hard-attention-but-we-can-use-it-in-relu>

[/https://chat.openai.com](https://chat.openai.com)

1.

بخش (a) گزینه‌های ب و ج صحیح هستند.

معماری many to one RNN برای تسک‌هایی مناسب هستند که ورودی آن‌ها فرم sequential دارد و خروجی واحد است. از آن‌جا که گزینه آ دارای یک خروجی واحد نیست انتخاب نشده است.

بخش (b) گزینه ج صحیح است.

زیرا اخلاق گربه به آب و هوای گذشته و امروز بستگی دارد اما به آب و هوای آینده بستگی ندارد و ما آن را نمی‌دانیم زیرا هنوز اتفاق نیفتاده است.

بخش (c) گزینه ج صحیح است.

در یک مدل RNN هدف پیش‌بینی یک مرحله با توجه به مراحل قبلی آن است.

①

(A)

③

(ب) با توجه به اینکه در RNN، هزینه محاسبه در هر گام به صورت زیر است:

۱) عرب سے سفر.

got

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial z_t} \times \frac{\partial z_t}{\partial h_{t-1}}$$

5-1-1

1-1-1

Subject:

Year:

Month:

Date:

$$h_t = \sigma(w_{hh} \cdot h_{t-1} + w_{hx} \cdot x_t) = \sigma(z_t) \quad (2)$$

$$\Rightarrow \frac{\partial h_t}{\partial w_{hh}} = \frac{\partial h_t}{\partial z_t} \times \frac{\partial z_t}{\partial w_{hh}} = \sigma'(z_t) \times h_{t-1}$$

$$\frac{\partial J_t}{\partial w_{hh}} = \sum_{i=1}^t \frac{\partial J_t}{\partial h_i} \times \frac{\partial h_i}{\partial w_{hh}} = \sum_{i=1}^t g_{h_{t,i}} \times \sigma'(z_t) \times h_{i-1}$$

$$\frac{\partial J}{\partial w_{hh}} = \sum_{t=1}^r \frac{\partial J_t}{\partial w_{hh}} = \sum_{t=1}^r g_{w_{hh},t} \quad (3)$$

3.

الف. در گام نخست ضرب داخلی پرسوجو و کلیدها را محاسبه می‌کنیم:

for key0 => $q.key0 = 3 - 2 - 3 = -2$

for key1 => $q.key1 = 6 - 2 - 1 = 3$

for key2 => $q.key2 = 0 - 1 + 1 = 0$

for key3 => $q.key3 = 0 + 2 + 4 = 6$

سپس اگر argmax را اعمال کنیم مشابه‌ترین کلید اندیس ۳ را دارد.

حال با استفاده از این اندیس، مقدار مربوطه را به دست می‌آوریم:

$\text{Values}[3] = [6, 1, 2]$

ب.

استفاده از argmax در مکانیسم‌های توجه، هم مزایا و هم معایبی دارد. توجه argmax شامل انتخاب عنصر با بالاترین مقدار از توزیع توجه است، که به طور مؤثر بر روی یک مکان یا ویژگی خاص تمرکز می‌کند. بیایید تأثیر آن بر آموزش مدل‌ها را بررسی کنیم:

مزایا:

کارایی: argmax attention به لحاظ محاسباتی بهینه‌تر از soft attention است. این گونه توجه هزینه محاسباتی کمتری حین $\text{forward/backward pass}$ ایجاد می‌کند.

تفسیرپذیری: وزن‌های حاصل از توجه argmax دودویی (0 یا 1) هستند، که تفسیر قسمت‌های مهم ورودی توسط مدل را آسان‌تر می‌کند.

معایب:

عدم مشتق‌پذیری: چالش اصلی با توجه argmax این است که قابل مشتق‌گیری نیست. این یک مشکل حین فرآیند backpropagation ایجاد می‌کند که در آن گرادیان‌ها محاسبه و برای به‌روزرسانی پارامترهای مدل استفاده می‌شوند.

از بین رفتن اطلاعات: عدم مشتق‌پذیری توجه argmax منجر به از دست رفتن اطلاعات گرادیان می‌شود، که باعث سختی یادگیری و سازگاری مدل در طول آموزش می‌شود. این می‌تواند به همگرایی غیر بهینه و کارایی کمتر منجر شود.

4.

فایل مربوطه ضمیمه شده است.