Date ______ Subject ______

Hw(6) پارسین عسگریان

ج۱)

الف)

$$\text{Swish} = x \cdot \text{Sigmoid}(x)$$

تصویر نمودار با نام swish.png پیوست شد

$$\text{mish} = x \tanh(\text{softplus}(x))$$

تصویر نمودار با نام mish.png پیوست شد

ب)

$$\text{Swish} = x \cdot \text{Sigmoid}(x)$$

$$S'(x) = \text{Sigmoid}(x) + x \cdot \text{Sigmoid}(x)(1 - \text{Sigmoid}(x))$$

$$= \sigma(x) + x \cdot \sigma(x) - x \cdot \sigma(x)^2$$

$$= x \cdot \sigma(x) + \sigma(x)(1 - x \cdot \sigma(x))$$

$$= S(x) + \text{Sigmoid}(x)(1 - S(x))$$

CHITRA

* مشتق تابع 1st-derivative-Swish برابر است با

$$m(n) = n \cdot \tanh(\text{Softplus}(n))$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$\text{Softplus} = \ln(1 + e^{x})$$

$$m(n) = x \cdot \frac{e^{\ln(1+e^{n})} - e^{-\ln(1+e^{n})}}{e^{\ln(1+e^{n})} + e^{-\ln(1+e^{n})}}$$

~~$m'(n) = \dots$~~

$$m'(n) = n \cdot \left(\frac{(e^{n}+1)^{2} - 1}{(e^{n}+1)^{2} + 1} \right)$$

* مشتق تابع 1st-derivative-mish برابر است با

Date Subject

ت) تابع Relu نسبت به Sigmoid و tanh

هزینه‌ی محاسباتی کمتری دارد و در نتیجه سریع‌تر است

در تابع Sigmoid و tanh به این دلیل که در

$+\infty$ و $-\infty$ شیب صفر را دارند در نتیجه مدل ها دیرتر

یاد می‌گیرد. ~~[illegible]~~

تابع Relu در مقدارهای منفی مقدار صفر خلق را

برمی‌گرداند که در نتیجه شیب صفر دارد و عملاً اگر مقدار

منفی ای بدست آید مدل ~~[illegible]~~ در یادگیری دچار مشکل می‌شود

هرچند که روی یک batch لحاظ می‌شود و احتمال اینکه

همه مقادیر منفی بشوند بسیار کم است

دو مدل Swish و Mish در مقدارهای منفی

Date

Subject

دارای شیب صفر می باشد

ج ث)

$$Swish(x;\beta) = 2x \cdot \sigma(\beta x)$$

مثل PRelu میاند یک متغیر قابل آموزش β

اگر $\beta=0$ باشد $Swish(x,0)=x$ خلاصه شده است

و اما اگر به $\beta \to \infty$ میکرد مثل Sigmoid د/ر

ج) د/ر mish $\Delta(x)Swish(x)+\frac{mish(x)}{x}$

Swish مثل یک پیش شرط برای $\Delta(x)$

ی ماند که باعث میشود نقاط منفی محاسبه شده شیب بهتر

شود در نتیجه سریع تر اپتیمایز شود

Date Subject

(ج۲

ی

تست ۱) مقدار های گفته شده وابسته به مقداردهی Initial وزن ها هستند اگر وزن ها را صفر

بگیریم در epoch اول:

$$MSE = \frac{1}{n}\sum^{n} (y_i - \hat{y}_i)^2$$

$$y \begin{cases} 0 & (0-0.5)^2 = 0.25 \\ 1 & (1-0.5)^2 = 0.25 \end{cases}$$

binary cross entropy: $-(y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$

$$y \begin{cases} 0 & -(0 + \log(0.5)) \simeq 0.7 \\ 1 & -(0 + \log(0.5)) \simeq 0.7 \end{cases}$$

Date Subject

ج ـ ۱ چون مدل داده‌های b حفظ کرده و overfit شده است

ج ـ ۲ برای binary crossentropy در epoch 60 ...

قبل از آن کاهش

برای MSE در epoch 100 یا 80 ...

ج ۳)

با توجه به نمودار مقدار $\alpha = 0.5$ از همه بهتر بود

چون مقدار validation loss و loss خیلی نزدیک هم بود

و به نظر می‌رسید overfit رخ نداده

در مقدار $\alpha = 1$ بعد از epoch آخر، افزایش