به نام خدا

یادگیری ژرف

نیم‌سال دوم ۰۳ – ۰۲

مدرس: دکتر مهدیه سلیمانی

سید امیر کسائی- ۴۰۲۲۱۲۲۱- همفکری با: امیر محمد عزتی

سوال اول: مشتق جزئی (۱۲ نمره)

فرض کنید یک ماتریس دلخواه $A_{m\times n}$ داریم. همچنین بردارهای y و x که به ترتیب m و n بعدی هستند و به صورت $y = Ax$ به هم مرتبط می شوند. مشتق جزئی y نسبت به x به صورت زیر تعریف می شود.

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

از روی فرم باز شده ی یک ضرب ماتریسی یعنی $y_i = \sum_{k=1}^{n} a_{ik}x_k$ عبارات زیر را بدست آورید.

- $\frac{\partial y}{\partial x} = A$

$$y_i = \sum_{k=1}^{n} a_{ik}x_k$$

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j}(\sum_{k=1}^{n} a_{ik}x_k)$$

$$\rightarrow \frac{\partial}{\partial x_j}(a_{ij}x_j) = a_{ij} \Rightarrow \frac{\partial y_i}{\partial x_j} = a_{ij} \Rightarrow \frac{\partial y}{\partial x} = A$$

- اگر x یک تابع از z و A مستقل از z باشد، ثابت کنید $\frac{\partial y}{\partial z} = A\frac{\partial x}{\partial z}$

$$\frac{\partial y}{\partial z} = A\frac{\partial x}{\partial z}$$

$$\frac{\partial y}{\partial z} = \frac{\partial (Ax)}{\partial z} = \frac{\partial A}{\partial z}x + A\frac{\partial x}{\partial z} = 0 + A\frac{\partial x}{\partial z} = A\frac{\partial x}{\partial z}$$

- اگر تعریف کنیم $\alpha = y^T Ax$ ثابت کنید $\frac{\partial \alpha}{\partial x} = y^T A$ و $\frac{\partial \alpha}{\partial y} = x^T A^T$

$$\alpha = y^T Ax = B$$

$$\rightarrow b_i = \sum_{k=1}^{n} a_{ik}\, x_k$$

$$y^T B = \sum_{k=1}^{m} y_k b_k$$

$$\alpha = y^T A x = \sum_{i=1}^{m} y_i \sum_{j=1}^{n} a_{ij}\, x_j$$

$$\frac{\partial \alpha}{\partial y_k} = \frac{\partial}{\partial y_k}\left(y_k \sum_{j=1}^{n} a_{kj} x_j \right) = \sum_{j=1}^{n} a_{kj}\, x_j$$

$$\frac{\partial \alpha}{\partial y} = \left(\frac{\partial \alpha}{\partial y_1}, \dots, \frac{\partial \alpha}{\partial y_m} \right) = x^T A^T$$

$$\frac{\partial \alpha}{\partial x} = \frac{\partial}{\partial x}(y^T A x)$$

$$\frac{\partial \alpha}{\partial x_k} = \frac{\partial}{\partial y_k}(x_k \sum_{i=1}^{m} y_i a_{ik}) = \sum_{i=1}^{m} y_i a_{ik}$$

$$\Longrightarrow \frac{\partial \alpha}{\partial x} = y^T A$$

- اگر x و y دو بردارد n بعدی که تابعی از متغیر z اند و $\alpha = y^T x$ ثابت کنید: $\frac{\partial \alpha}{\partial z} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z}$

$$\alpha = y^T x = \sum_{i=1}^{n} y_i x_i$$

$$\frac{\partial \alpha}{\partial z} = \sum_{i=1}^{n} (\frac{\partial y_i}{\partial z}\, x_i + y_i \frac{\partial x_i}{\partial z})$$

$$\frac{\partial \alpha}{\partial z} = \left(\frac{\partial y_1}{\partial z}, \frac{\partial y_2}{\partial z}, \dots, \frac{\partial y_n}{\partial z} \right) \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{pmatrix} + \left(\frac{\partial x_1}{\partial z}, \frac{\partial x_2}{\partial z}, \dots, \frac{\partial x_n}{\partial z} \right) \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z}$$

$$\frac{\partial \alpha}{\partial z} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z}$$

- اگر $A_{m\times m}$ یک ماتریس non singular باشد که درایه های آن تابع هایی از مقدار اسکالر α باشد، ثابت کنید: $\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$

$$AA^{-1} = I$$

$$\frac{\partial (AA^{-1})}{\partial \alpha} = \frac{\partial A}{\partial \alpha} A^{-1} + \frac{\partial A^{-1}}{\partial \alpha} A = \frac{\partial (I)}{\partial \alpha} = 0 \Rightarrow \frac{\partial A^{-1}}{\partial \alpha} A = -\frac{\partial A}{\partial \alpha} A^{-1} \xRightarrow{\times A^{-1}}$$

$$I \frac{\partial A^{-1}}{\partial \alpha} = -\frac{\partial A}{\partial \alpha} A^{-1} A^{-1} \Rightarrow \frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

سوال دوم: ماتریس Hessian (۸ نمره)

نشان دهید ماتریس Hessian یک تبدیل مثل $y = \psi(u,v,z)$ را می توان به صورت ماتریس ژاکوبی گرادیان این تبدیل نوشت. توجه کنید که متغیرهای u, v, z تک بعدی و y نیز تابعی بر حسب آنها است.

$$y = \psi(u,v,z)$$

$$\nabla\psi = \left(\frac{\partial\psi}{\partial u}, \frac{\partial\psi}{\partial v}, \frac{\partial\psi}{\partial z}\right) \xRightarrow{J} J(\nabla\psi) = \begin{bmatrix} \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial u}\left(\frac{\partial\psi}{\partial z}\right) \\ \frac{\partial}{\partial v}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial v}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial v}\left(\frac{\partial\psi}{\partial z}\right) \\ \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial u}\right) & \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial v}\right) & \frac{\partial}{\partial z}\left(\frac{\partial\psi}{\partial z}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial^2\psi}{\partial u^2} & \frac{\partial^2\psi}{\partial u\partial v} & \frac{\partial^2\psi}{\partial u\partial z} \\ \frac{\partial^2\psi}{\partial v\partial u} & \frac{\partial^2\psi}{\partial v^2} & \frac{\partial^2\psi}{\partial v\partial z} \\ \frac{\partial^2\psi}{\partial z\partial u} & \frac{\partial^2\psi}{\partial z\partial v} & \frac{\partial^2\psi}{\partial z^2} \end{bmatrix}$$

$$Hessian = \begin{bmatrix} \frac{\partial^2\psi}{\partial u^2} & \frac{\partial^2\psi}{\partial u\partial v} & \frac{\partial^2\psi}{\partial u\partial z} \\ \frac{\partial^2\psi}{\partial v\partial u} & \frac{\partial^2\psi}{\partial v^2} & \frac{\partial^2\psi}{\partial v\partial z} \\ \frac{\partial^2\psi}{\partial z\partial u} & \frac{\partial^2\psi}{\partial z\partial v} & \frac{\partial^2\psi}{\partial z^2} \end{bmatrix}$$

$$\Longrightarrow H(\psi) = J(\nabla\psi)$$

سوال سوم: جلوگیری از نامتقارن شدن(۱۰ نمره)

برخی از انواع مجموعه دادگان، مانند برخی از انواع سری زمانی یا تصاویر صورت، دارای یک شبه تقارن ذاتی هستند. ابتدا تحقیق کنید که منظور از این تقارن چیست و چگونه به آموزش بهتر یک مدل دسته بندی کمک می کنند. حال فرض کنید که یک عکس کوچک با ابعاد 1×2 در اختیار داریم. ترم رگولاریزیشن L_2 به صورت $R(\omega) = \omega^T\omega = \omega^T I\omega$ تعریف می شود. ماتریس S را بیابید به گونه ای که $R(\omega) = \omega^T S\omega$ از نامتقارن شدن وزن ها جلوگیری کند.

شبه تقارن ذاتی داده ها به معنای وجود الگو ها یا تقارن های ساختاری در خود داده هاست که می‌تواند برای درک و مدل سازی بهتر مورد استفاده قرار گیرد. از مثال های آن در حوزه تصویر می‌توان به تقارن چرخشی،تقارن انعکاسی و تقارن translational اشاره کرد. در تقارن چرخشی جسم پس از درجه خاصی از چرخش مثل حالت اولیه به نظر می‌رسد. در تقارن translational الگو هایی مانند آنچه در کاشی ها است وجود دارد و در واقع شکل دارای واحد های تکرار شونده است. تقارن انعکاسی مانند تصاویر پروانه ها است که در امتداد محور خاصی تقارن وجود دارد. در حوزه متن هم می‌توان به وجود ساختار متقارن از لحاظ معنا یا وجود جفت های واژگان مانند مترادف و متضاد اشاره کرد. داده های سری زمانی هم می‌توانند دارای تقارن باشند. مانند داده های فصلی. شناسایی این تقارن ها کاربرد های مختلفی در زمینه مدل سازی دارد.

Regularization: درک شبه تقارن ها می‌توان منجر به استفاده بهتر از تکنیک های منظم سازی شود که وزن های شبکه را تشویق به یادگیری تقارن ها می‌کند.

Feature representation: شبه تقارن ذاتی می‌تواند در انتخاب یا حتی ایجاد ویژگی ها کمک کننده باشد. با شناسایی تقارن ها،محققان می‌توانند ویژگی هایی طراحی کنند که به یادگیری الگو های متمایز توسط مدل کمک کنند.

Data Augmentation: با اعمال چرخش یا بازتاب یا پیدا کردن الگو های تکرار شونده می‌توان داده های جدیدی تولید کرد.

معماری مدل: برای مثالی از این کاربرد می‌توان به شبکه های عصبی CNN اشاره کرد که از تقارن در تصاویر استفاده می‌کنند.

تفسیر پذیری: درک تقارن ذاتی در داده ها می‌تواند منجر به داشتن مدل های قابل تفسیر بیشتری شود. زیرا با شناسایی و مدل سازی این تقارن ها می‌توان بینش خوبی درباره ساختار داده ها و عوامل موثر در طبقه بندی بدست آورد.

$$R(\omega) = \omega^T\omega = \omega^T I\omega$$

$$\omega = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$R(\omega) = \omega^T S\omega \Longrightarrow R(\omega) = [w_1 \quad w_2]\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = [w_1 \quad w_2]\begin{bmatrix} aw_1 + bw_2 \\ cw_1 + dw_2 \end{bmatrix} = (aw_1{}^2 + bw_1w_2 + cw_1w_2 + dw_2{}^2)$$

$$To\ prevent\ being\ asymmetric \Longrightarrow a = d, b = c$$

$$R(\omega) = a(w_1{}^2 + w_2{}^2) + 2bw_1w_2$$

$$R(\omega) = (w_1 - w_2)^2 = w_1{}^2 + w_2{}^2 - 2w_1w_2$$

$$\Longrightarrow a = 1, b = -1 \Longrightarrow S = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

سوال چهارم: Backpropagation Basics (۱۰ نمره)

فرض کنید شبکه عصبی دو لایه مانند زیر داریم

$$z_1 = W_1x^{(i)} + b_1$$
$$a_1 = ReLU(z_1)$$
$$z_2 = W_2a_1 + b_2$$
$$\hat{y}^{(i)} = \sigma(z_2)$$
$$L^{(i)} = y^{(i)} * \log(\hat{y}^{(i)}) + (1 - y^{(i)}) * \log(1 - \hat{y}^{(i)})$$
$$J = \frac{-1}{m}\sum_{i=1}^{m} L^{(i)}$$

توجه کنید که $x^{(i)}$ نشان دهنده یک نمونه ورودی با ابعاد $D_x \times 1$ است. همچنین $y^{(i)}$ برچسب یک نمونه است و به صورت اسکالر می‌باشد. دیتاست شامل m نمونه است. همچنین z_1 ابعاد $D_{a_1} \times 1$ دارد.

- ابعاد W_1, b_1, W_2, b_2 را بنویسید.

$$Z_1 = W_1x^{(i)} + b_1$$

$$x^{(i)}: [D_x, 1]$$

$$Z_1: [D_{a1}, 1]$$

$$[D_{a1}, 1] = W_1[D_x, 1] + b_1$$

$$W_1: [D_a, D_x], b_1: [D_a, 1]$$

- نتیجه $\partial J/\partial \hat{y}^{(i)}$ را بدست آوردید و آن را با δ_1 نشان دهید.

$$\delta_1 = \frac{\partial J}{\partial \hat{y}^{(i)}} = \frac{-1}{m}\frac{\partial L^{(i)}}{\partial \hat{y}^{(i)}} = \frac{-1}{m}(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1 - y^{(i)}}{1 - \hat{y}^{(i)}})$$

- نتیجه $\partial \hat{y}^{(i)}/\partial z_2$ را بدست آوردید و آن را با δ_2 نشان دهید.

$$\delta_2 = \frac{\partial \hat{y}^{(i)}}{\partial z_2} = \sigma(z_2)(1 - \sigma(z_2))$$

- نتیجه $\partial z_2/\partial a_1$ را بدست آوردید و آن را با δ_3 نشان دهید.

$$\delta_3 = \frac{\partial z_2}{\partial a_1} = W_2$$

- نتیجه $\partial a_1/\partial z_1$ را بدست آوردید و آن را با δ_4 نشان دهید.

$$\delta_4 = \frac{\partial a_1}{\partial z_1} = \begin{cases} 0, & z_1 < 0 \\ 1, & z_1 \geq 0 \end{cases}$$

- نتیجه $\partial z_1/\partial W_1$ را بدست آوردید و آن را با δ_5 نشان دهید.

$$\delta_5 = \frac{\partial z_1}{\partial W_1} = x^T$$

- نتیجه $\partial J/\partial W_1$ را با استفاده از نتایج قبلی بدست آورید.

$$\frac{\partial J}{\partial W_1} = \delta_1 \times \delta_2 \times \delta_3 \times \delta_4 \times \delta_5 = \frac{-1}{m}\left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}}\right)\sigma(z_2)(1-\sigma(z_2))W_2\delta_4 x^T$$

$$\frac{\partial J}{\partial W_1} = \begin{cases} 0, & z_1 < 0 \\ \frac{-1}{m}\left(\frac{y^{(i)}}{\hat{y}^{(i)}} - \frac{1-y^{(i)}}{1-\hat{y}^{(i)}}\right)\sigma(z_2)(1-\sigma(z_2))W_2 x^T, & z_1 \geq 0 \end{cases}$$

سوال پنجم: بهینه سازی (۲۰ نمره)

۱. تابع زیر که Beale نام دارد را درنظر بگیرید:

$$f(\underline{x}) = (1.5 - x_1 + x_1x_2)^2 + (2.25 - x_1 + x_1x_2^2)^2 + (2.625 - x_1 + x_1x_2^3)^2 \qquad (۱)$$

- با استدلال مناسب بیان کنید که این تابع محدب است یا نامحدب؟ این مسئله در بهینه‌سازی توابع چه اهمیتی دارد؟
- حال با درنظر گرفتن نقطه شروع $(0,1)$، جهت گرادیان را پیدا کرده و سپس با کمک الگوریتم گرادیان کاهشی، مقدار جدید نقطه شروع پس از یک بار به‌روزرسانی را به‌دست آورید.

۲. تابع هدف محدب درجه‌دو زیر که در آن Q یک ماتریس مثبت معین (Positive definite) است را درنظر بگیرید:

$$h(x) = \frac{1}{2}x^TQx + x^Tc + b \qquad (۲)$$

ثابت کنید که الگوریتم بهینه‌سازی گرادیان کاهشی برای چنین توابعی معادل بهینه‌سازی درجهت بردارهای ویژه ماتریس درجه‌دو Q است.

۳. با الگوریتم بهینه‌سازی Adam در درس آشنا شدید. در این الگوریتم قانون به‌روزرسانی هر وزن این است که گرادیان آن را به‌صورت متناسب با معکوس نُرم‌دو گرادیان‌های فعلی و قبلی آن مقیاس کنیم.

- با جایگزینی نُرم‌دو مذکور با نُرم‌بی‌نهایت به الگوریتم بهینه‌سازی جدیدی برسید. (توجه کنید که توان β_2 را برابر p که $p \to \infty$ درنظر بگیرید.)
- الگوریتم حاصل را با روش Adam مقایسه و بیان کنید در چه شرایطی استفاده از الگوریتم حاصل‌شده بهتر است؟

۱)

a) در مورد تابع Beale، یک saddle point در (0,1) دارد، که نقطه ای است که نه بیشینه محلی است و نه کمینه محلی.
saddle point مکانی است که در آن تابع در یک جهت منحنی رو به بالا می شود و در جهت دیگر منحنی رو به پایین می شود، مانند زین و این ویژگی توابع غیر محدب است.

b)

$$f(x,y) = (1.5 - x + xy)^2 + (2.25 - x + xy^2)^2 + (2.625 - x + xy^3)^2$$

$$\frac{\partial f}{\partial x} = 2(1.5 - x + xy)(-1 + y) + 2(2.25 - x + xy^2)(-1 + y^2) + 2(2.625 - x + xy^3)(-1 + y^3)$$

$$\rightarrow \frac{\partial f}{\partial x}(0,1) = 0$$

$$\frac{\partial f}{\partial y} = 2(1.5 - x + xy)x + 2(2.25 - x + xy^2)xy + 2(2.625 - x + xy^3)x^2$$

$$\rightarrow \frac{\partial f}{\partial y}(0,1) = 0$$

$$A = (0,1) \rightarrow A_{new} = A - \alpha\left(\frac{\partial f}{\partial x}(A), \frac{\partial f}{\partial y}(A)\right) = (0,1) - \alpha\left(\frac{\partial f}{\partial x}(0,1), \frac{\partial f}{\partial y}(0,1)\right) = (0,1)$$

۲) برای سادگی فرض میکنیم :

$$h(x) = \frac{1}{2}x^TQx \qquad , \qquad x_{i+1} = x_i - \alpha\nabla h(x_i)$$

$$\xrightarrow{\text{Q is PSD}} \nabla h(x) = Qx \rightarrow\ x_{i+1} = x_i - \alpha(Qx_i) = (I - \alpha Q)x_i \xrightarrow{\textit{due to reccurence}} x_k = (I - \alpha Q)^k x_0$$

برای بررسی توان در ماتریس بدست آمده، فرض میکنیم $Q = S\Lambda S^T$ ($S^{-1} = S^T$) یک eigenvalue decomposition برای Q است که در آن S ماتریسی متعامد ($SS^{-1} = I$) از بردارهای ویژه و Λ یک ماتریس قطری از مقدار های ویژه است. بنابراین داریم:

$$x_k = (I - \alpha Q)^k x_0 = (I - \alpha S\Lambda S^T)^k x_0 = [S(I - \alpha\Lambda)S^T]^k x_0 = S(I - \alpha\Lambda)^k S^T$$

از آنجایی که گرادیان کاهشی نسبت به جابجایی تغییر ناپذیر است پس این رابطه برای $h(x) = \frac{1}{2}x^TQx + x^Tc + b$ نیز صادق است.
بنابراین گرادیان کاهشی معدل بهینه سازی در جهت پردار های ویژه ماتریس Q (S) است.

[CS Toronto]

۳)

a) الگوریتم Adam به صورت زیر است:

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)

$\widehat{m}_t \leftarrow m_t/(1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)

$\hat{v}_t \leftarrow v_t/(1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t/\left(\sqrt{\hat{v}_t} + \epsilon\right)$ (Update parameters)

$$if\ v_0 = 0 \xrightarrow{recurrence} v_t = (1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}.g_i^2$$

در الگوریتم جدید (Adamax) داریم:

$$v_t = \beta_2^p v_{t-1} + \left(1-\beta_2^p\right)|g_t|^p = (1-\beta_2)\sum_{i=1}^{t}\beta_2^{p(t-i)}.|g_i|^p \quad (p\to\infty)$$

$$u_t = \lim_{p\to\infty}(v_t)^{\frac{1}{p}} = \lim_{p\to\infty}\left(\left(1-\beta_2^p\right)\sum_{i=1}^{t}\beta_2^{p(t-i)}\cdot|g_i|^p\right)^{\frac{1}{p}}$$
$$= \lim_{p\to\infty}\left(1-\beta_2^p\right)^{\frac{1}{p}}\left(\sum_{i=1}^{t}\beta_2^{p(t-i)}\cdot|g_i|^p\right)^{\frac{1}{p}}$$
$$= \lim_{p\to\infty}\left(\sum_{i=1}^{t}\left(\beta_2^{(t-i)}\cdot|g_i|\right)^p\right)^{\frac{1}{p}}$$
$$= max(\beta_2^{t-1}|g_1|, \beta_2^{t-2}|g_2|, \dots, \beta_2|g_{t-1}|, |g_t|)$$
$$\to u_t = \max(\beta_2.u_{t-1}, |g_t|)$$

در نتیجه داریم:

$$m_t \leftarrow \beta_1\cdot m_{t-1} + (1-\beta_1)\cdot g_t \text{ (Update biased first moment estimate)}$$

$$u_t \leftarrow \max(\beta_2\cdot u_{t-1}, |g_t|) \text{ (Update the exponentially weighted infinity norm)}$$

$$\theta_t \leftarrow \theta_{t-1} - \left(\alpha/(1-\beta_1^t)\right)\cdot m_t/u_t \text{ (Update parameters)}$$

b) Adam و Adamax الگوریتم های بهینه سازی هستند که در آموزش مدل های یادگیری عمیق استفاده می شوند. Adam مزایای AdaGrad و RMSProp را ترکیب می کند و نرخ یادگیری را بر اساس لحظات اول و دوم گرادیان تنظیم می کند. Adamax، گونه‌ای از Adam بر اساس نورم بی‌نهایت، از حداکثر مقدار مطلق برای مقیاس‌بندی گرادیان‌ها استفاده می‌کند، و آن را به طور بالقوه پایدارتر و قوی‌تر می‌کند، به خصوص در موقعیت‌هایی با sparse gradients، noisy optimization، یا زمانی که گرفتن وابستگی‌های long-term بسیار مهم است. در حالی که Adam به طور گسترده برای کارایی خود در طیف گسترده ای از وظایف استفاده می شود، Adamax ممکن است به دلیل رویکرد متفاوت خود در مقیاس گرادیان، مزایایی را در سناریوهای خاص ارائه دهد.

[Adamax]