به نام خدا

روش مقاله:

روش nag به صورت زیر میباشد:

Algorithm 3 Nesterov's accelerated gradient

$$\mathbf{g}_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1} - \eta\mu\mathbf{m}_{t-1})$$
$$\mathbf{m}_t \leftarrow \mu\mathbf{m}_{t-1} + \mathbf{g}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \eta\mathbf{m}_t$$

مشاهده میکنیم که مومنتوم دوبار روی پارامتر ها اعمال شده است. یک بار برای پارامتر موقت:

$$\theta' = \theta_{t-1} - \eta\mu\, m_{t-1}$$

و یک بار برای قانون آپدیت:

$$\theta_t = \theta_{t-1} - \mu\, m_t$$

فرمول بالا را باز نویسی میکنیم تا خوانایی آن برای ادغام با adam بیشتر شود:

Algorithm 7 NAG rewritten

$$\mathbf{g}_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$$
$$\mathbf{m}_t \leftarrow \mu_t\mathbf{m}_{t-1} + \mathbf{g}_t$$
$$\bar{\mathbf{m}}_t \leftarrow \mathbf{g}_t + \mu_{t+1}\mathbf{m}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \eta\bar{\mathbf{m}}_t$$

با این فرمول ما مومنتوم را فقط در قانون آپدیت اعمال میکنیم. $\bar{m}_t$ حاوی گرادیان در تایم استپ فعلی به علاوه آپدیت بردار مومنتوم برای تایم استپ بعدی میباشد $\mu_{t+1} m_t$.

در ادامه از منبعی که در آخر ذکر کرده ام استفاده میکنم تا روش ادغام را توضیح دهم. در این منبع علامت ها و فرمول ها کمی متفاوت هستند اما استدلالی که توضیح داده میشود ما را به یک نتیجه میرساند برای همین دوباره آن ها ذکر میکنم.

Nag:

$$\tilde{\theta} = \theta_t + \alpha v_t$$
$$g_{NAG} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathcal{L}(x^{(i)}, y^{(i)}, \tilde{\theta}) \tag{1}$$

$$v_{t+1} = \alpha v_t - \eta g_{NAG}$$
$$\theta_{t+1} = \theta_t + v_{t+1}. \tag{2}$$

Rewrite nag:

$$g_t = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathcal{L}\left(x^{(i)}, y^{(i)}, \theta_t\right)$$
$$m_t = \rho_1 m_{t-1} - \eta g_t \tag{3}$$
$$\bar{m}_t = \rho_1 m_t - \eta g_t$$
$$\theta_{t+1} = \theta_t + \bar{m}_t$$

چون آدام از مومنتوم استفاده میکند پس از تغییر مومنتوم کلاسیک شروع میکنیم. فرمول مومنتوم کلاسیک:

Classical mom:

$$\theta_{t+1} = \theta_t + m_t$$
$$\theta_{t+1} = \theta_t + \rho_1 m_{t-1} - \eta g_t \tag{4}$$

در عبارت دوم m_t با مقدار سمت راست معادله اش در (3) جایگزین شده است.

از آنجایی که m_{t-1} به گرادیان فعلی وابستگی ندارد این ترم را تغییر میدهیم و از ترفند نستروو استفاده میکنیم تا به rewritten nag برسیم:

$$\begin{aligned}\theta_{t+1} &= \theta_t + \rho_1(\rho_1 m_{t-1} - \eta g_t) - \eta g_t \\ \theta_{t+1} &= \theta_t + \rho_1 m_t - \eta g_t \\ \theta_{t+1} &= \theta_t + \bar{m}_t\end{aligned} \tag{5}$$

حالا به سراغ فرمول adam میرویم:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t + \varepsilon}} \tag{6}$$

فرمول m_t را مانند کاری که برای مومنتوم کلاسیک انجام دادیم باز میکنیم:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \cdot \frac{\rho_1 m_{t-1} + (1-\rho_1) g_t}{\sqrt{v_t + \varepsilon}} \\ \theta_{t+1} &= \theta_t - \eta \cdot \frac{\rho_1 m_{t-1}}{\sqrt{v_t + \varepsilon}} - \eta \cdot \frac{(1-\rho_1) g_t}{\sqrt{v_t + \varepsilon}}\end{aligned} \tag{7}$$

ما به طور مستقیم نمیتوانیم nesterov را روی دومین ترم عبارت سمت راست فرمول بالا اعمال کنیم چون v_t به g_t که گرادیان فعلی میباشد وابسته است:

$$v_t = \rho_2 v_{t-1} + (1-\rho_2) g_t^2 \tag{8}$$

از آنجایی که ρ_2 خیلی بزرگ مقدار دهی میشود (بین 0.9 تا 0.999) پس تفاوت بین v_t و v_{t-1} کوچک میباشد و میتوانیم جای v_t ، v_{t-1} را بگذاریم که فرمول آدام به صورت زیر میشود:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\rho_1 m_{t-1}}{\sqrt{v_{t-1} + \varepsilon}} - \eta \cdot \frac{(1-\rho_1) g_t}{\sqrt{v_t + \varepsilon}} \tag{9}$$

حالا میتوانیم ترفند nesterov را اعمال کنیم و به عبارت زیر برسیم:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\rho_1(\rho_1 m_{t-1} + (1-\rho_1)g_t)}{\sqrt{(\rho_2 v_{t-1} + (1-\rho_2)g_t^2) + \varepsilon}} - \eta \cdot \frac{(1-\rho_1)g_t}{\sqrt{v_t + \varepsilon}}$$
$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\rho_1 m_t}{\sqrt{v_t + \varepsilon}} - \eta \cdot \frac{(1-\rho_1)g_t}{\sqrt{v_t + \varepsilon}} \qquad (10)$$
$$\theta_{t+1} = \theta_t - \eta \cdot \frac{(\rho_1 m_t + (1-\rho_1)g_t)}{\sqrt{v_t + \varepsilon}}$$
$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\bar{m}_t}{\sqrt{v_t + \varepsilon}}$$

در ادامه تصحیح بایاس هم انجام می شود که به فرمول نهایی مقاله میرسیم:

Algorithm 8 Nesterov-accelerated adaptive moment estimation

$$\mathbf{g}_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$$
$$\hat{\mathbf{g}} \leftarrow \frac{\mathbf{g}_t}{1-\prod_{i=1}^{t}\mu_i}$$
$$\mathbf{m}_t \leftarrow \mu \mathbf{m}_{t-1} + (1-\mu)\mathbf{g}_t$$
$$\hat{\mathbf{m}}_t \leftarrow \frac{\mathbf{m}_t}{1-\prod_{i=1}^{t+1}\mu_i}$$
$$\mathbf{n}_t \leftarrow \nu \mathbf{n}_{t-1} + (1-\nu)\mathbf{g}_t^2$$
$$\hat{\mathbf{n}}_t \leftarrow \frac{\mathbf{n}_t}{1-\nu^t}$$
$$\bar{\mathbf{m}}_t \leftarrow (1-\mu_t)\hat{\mathbf{g}}_t + \mu_{t+1}\hat{\mathbf{m}}_t$$
$$\theta_t \leftarrow \theta_{t-1} - \eta \frac{\bar{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{n}}_t + \varepsilon}}$$

مزایای nadam نسبت به adam؟ :

در 2 بنچمارک از 3 بنچمارک متوجه شدیم که عملکرد nadam به طور قابل ملاحظه ای بهتر از adam میباشد. چرا؟ مومنتوم کلاسیک به عنوان یک ورژن از مومنتوم نستروو است که بردار مومنتوم قدیمی و منسوخ را اعمال میکند. این بردار فقط از گرادیان های گذشته استفاده میکند در حالیکه در نستروو بردار مومنتوم اخیر و به روزی که گرادیان فعلی را هم استفاده میکند اعمال میشود. پس به طور منطقی نستروو آپدیت گرادیان بهتری را نسبت به مومنتوم کلاسیک تولید میکند. از طرفی RMSProp به کیفیت step الگوریتم در اپدیت ها توجه میکند و learning rate برای هر پارامتر متناسب با خودش است. پس هنگامی که به طور همزمان از RMSProp و Nestrov استفاده کنیم میتوانیم مزیت کامل را داشته باشیم(NAdam).

به نظر شما چه عاملی در بهینه سازها سبب میشود تا در مسئله های متفاوت عملکردهای متفاوت داشته باشند:

با توجه به مطالب گفته شده در discussion مقاله:

- سایز مدلی که انتخاب میکنیم (مثلا در مثال mnist گفته شده:
  However, [5] use a larger CNN and find that Adam achieves the lowest performance
- سادگی یا پیچیدگی مسئله.
- مقدار هایپرپارامتر هایی که تنظیم میکنیم.
- نوسانات ناشی از dropout
- Overfitting

این ها عواملی هستند که با توجه به هر مسئله تغییر میکنند و باعث عملکرد متفاوت بهینه ساز ها میشوند.

منبع استفاده شده برای شرح روش nadam:

https://medium.com/konvergen/modifying-adam-to-use-nesterov-accelerated-gradients-nesterov-accelerated-adaptive-moment-67154177e1fd