Adafactor: Adaptive Learning Rates with Sublinear Memory Cost

Noam Shazeer
Google Brain

Mitchell Stern
UC Berkeley
Gradient-Based Optimization
Gradient-Based Optimization

Method: Gradient
       SGD

Auxiliary Memory: 0
Gradient-Based Optimization

Method
- Gradient
- SGD

Momentum
- Heavy Ball
- Nesterov

Auxiliary Memory
- 0
- N
Gradient-Based Optimization

Method
- Gradient
- SGD

Momentum
- Heavy Ball
- Nesterov

Adaptivity
- Adagrad
- Adadelta
- RMSProp

Auxiliary Memory
- 0

N
Gradient-Based Optimization

Method

Gradient

SGD

Momentum

Heavy Ball

Nesterov

Adaptivity

Adagrad

Adadelta

RMSProp

Combined

Adam

Nadam

AMSGrad

Auxiliary Memory

0

N

2N
Gradient-Based Optimization

<table>
<thead>
<tr>
<th>Method</th>
<th>Gradient</th>
<th>Adaptness</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SGD</td>
<td>Adafactor</td>
<td></td>
</tr>
<tr>
<td>Momentum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heavy Ball</td>
<td>Nesterov</td>
<td></td>
</tr>
<tr>
<td>Adaptness</td>
<td>Adagrad</td>
<td>Adadelta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSProp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Auxiliary Memory | 0    | o(N)  | N    | 2N   |

0, 0(N), N, 2N
Adam vs. Adafactor

Algorithm 1 Adam (Kingma & Ba, 2015)
1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g^2_t$
7: $\hat{m}_t = m_t/(1 - \beta_1^t)$
8: $\hat{v}_t = v_t/(1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: **end for**

Algorithm 4 Adafactor for weight matrices.
1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2\}_{t=1}^T$ such that $\beta_2^t = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $\delta$
2: **for** $t = 1$ **to** $T$ **do**
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2) (G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2) (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_n^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / \delta)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: **end for**
Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$

2: Initialize $m_0 = 0$ and $v_0 = 0$

3: **for** $t = 1$ **to** $T$ **do**

4: $g_t = \nabla f_t(x_{t-1})$

5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$

6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

7: $\hat{m}_t = m_t / (1 - \beta_1^t)$

8: $\hat{v}_t = v_t / (1 - \beta_2^t)$

9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$

10: **end** **for**

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\hat{\beta}_2^t\}_{t=1}^T$ such that $\hat{\beta}_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$

2: **for** $t = 1$ **to** $T$ **do**

3: $\alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$

4: $G_t = \nabla f_t(X_{t-1})$

5: $R_t = \hat{\beta}_2^t R_{t-1} + (1 - \hat{\beta}_2^t) (G_t^2 + \epsilon_1 1_n 1_m^T 1_m)$

6: $C_t = \hat{\beta}_2^t C_{t-1} + (1 - \hat{\beta}_2^t) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$

7: $\hat{V}_t = R_t C_t / 1_n R_t$

8: $U_t = G_t / \sqrt{\hat{V}_t}$

9: $\hat{U}_t = U_t / \max (1, \text{RMS}(U_t) / d)$

10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$

11: **end** **for**
Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: **end for**

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\tilde{\beta}_2\}_{t=1}^T$ such that $\tilde{\beta}_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: **for** $t = 1$ **to** $T$ **do**
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $g_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_m 1_n^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: **end for**
**Algorithm 1** Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: **end for**

**Algorithm 4** Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: **for** $t = 1$ **to** $T$ **do**
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2 + \epsilon_1 1_n 1_m^T 1_m)$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2)(G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / (1_n^T R_t)$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: **end for**

Update first moment estimate
Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ to $T$ **do**
4: $\quad g_t = \nabla f_t(x_{t-1})$
5: $\quad m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
6: $\quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
7: $\quad \hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\quad \hat{v}_t = v_t / (1 - \beta_2^t)$
9: $\quad x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: **end for**

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2_t\}_{t=1}^T$ such that $\beta_2^t = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: **for** $t = 1$ to $T$ **do**
3: $\quad \alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $\quad G_t = \nabla f_t(X_{t-1})$
5: $\quad R_t = \beta_2_t R_{t-1} + (1 - \beta_2_t) (G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $\quad C_t = \beta_2_t C_{t-1} + (1 - \beta_2_t) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\quad \hat{V}_t = R_t C_t / 1_n R_t$
8: $\quad U_t = G_t / \sqrt{\hat{V}_t}$
9: $\quad \hat{U}_t = U_t / \max (1, \text{RMS}(U_t) / d)$
10: $\quad X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: **end for**

Update second moment estimate
Adam

Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs**: initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

Algorithm 4 Adafactor for weight matrices.

1: **Inputs**: initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2\}_{t=1}^T$ such that $\beta_2^t = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2) (G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T) 1_n$
7: $\hat{V}_t = R_t C_t / 1_n^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for

Perform bias correction
Adam

Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

---

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2^t\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $\delta$
2: for $t = 1$ to $T$
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_{2t} R_{t-1} + (1 - \beta_{2t})(G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \beta_{2t} C_{t-1} + (1 - \beta_{2t}) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_n^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for

Compute update vector
Adam

**Algorithm 1** Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$
   4: $g_t = \nabla f_t(x_{t-1})$
   5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
   6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
   7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
   8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
   9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
4: end for

**Algorithm 4** Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\hat{\beta}_t\}_{t=1}^T$ such that $\hat{\beta}_1 = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$
3: $\alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \hat{\beta}_t R_{t-1} + (1 - \hat{\beta}_t)(G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \hat{\beta}_t C_{t-1} + (1 - \hat{\beta}_t) 1_n (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_n R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max (1, \text{RMS}(U_t)/d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for

Update parameters
Adam vs. Adafactor

**Algorithm 1** Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4:   $g_t = \nabla f_t(x_{t-1})$
5:   $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6:   $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$
7:   $\hat{m}_t = m_t/(1 - \beta_1^t)$
8:   $\hat{v}_t = v_t/(1 - \beta_2^t)$
9:   $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

**Algorithm 4** Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3:   $\alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4:   $G_t = \nabla f_t(X_{t-1})$
5:   $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2 + \epsilon_1 1_n 1_m^\top) 1_m$
6:   $C_t = \beta_2 C_{t-1} + (1 - \beta_2)1_n^\top (G_t^2 + \epsilon_1 1_n 1_m^\top)$
7:   $\hat{V}_t = R_t C_t / 1_n R_t$
8:   $U_t = G_t / \sqrt{\hat{V}_t}$
9:   $\hat{U}_t = U_t / \max (1, \text{RMS}(U_t)/d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for
Algorithm 1 Adam (Kingma & Ba, 2015)

1: Inputs: initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4:    $g_t = \nabla f_t(x_{t-1})$
5:    $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
6:    $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
7:    $\hat{m}_t = m_t / (1 - \beta_1^t)$
8:    $\hat{v}_t = v_t / (1 - \beta_2^t)$
9:    $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

Algorithm 4 Adafactor for weight matrices.

1: Inputs: initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2^t\}_{t=1}^T$ such that $\beta_2^t = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3:    $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4:    $G_t = \nabla f_t(X_{t-1})$
5:    $R_t = \beta_2^t R_{t-1} + (1 - \beta_2^t)(G_t^2 + \epsilon_1 1_n 1_m^\top 1_m)$
6:    $C_t = \beta_2^t C_{t-1} + (1 - \beta_2^t) 1_n^\top (G_t^2 + \epsilon_1 1_n 1_m^\top)$
7:    $\hat{V}_t = R_t C_t / 1_n^\top R_t$
8:    $U_t = G_t / \sqrt{\hat{V}_t}$
9:    $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for
Adam vs. Adafactor

Algorithm 1 Adam (Kingma & Ba, 2015)
1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4:  \[ g_t = \nabla f_t(x_{t-1}) \]
5:  \[ m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \]
6:  \[ v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \]
7:  \[ \hat{m}_t = m_t/(1 - \beta_1^t) \]
8:  \[ \hat{v}_t = v_t/(1 - \beta_2^t) \]
9:  \[ x_t = x_{t-1} - \alpha_t \hat{m}_t/((\sqrt{\hat{v}_t} + \epsilon) \]
10: end for

Algorithm 4 Adafactor for weight matrices.
1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2 t\}_{t=1}^T$ such that $\beta_{2t} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3:  \[ \alpha_t = \max\left(\epsilon_2, \text{RMS}(X_{t-1})\right) \rho_t \]
4:  \[ G_t = \nabla f_t(X_{t-1}) \]
5:  \[ R_t = \beta_{2t} R_{t-1} + (1 - \beta_{2t})(G_t^2 + \epsilon_1 l_n 1_m^T) 1_m \]
6:  \[ C_t = \beta_{2t} C_{t-1} + (1 - \beta_{2t}) 1_n^T (G_t^2 + \epsilon_1 l_n 1_m^T) \]
7:  \[ \hat{V}_t = R_t C_t / 1_n^T R_t \]
8:  \[ U_t = G_t / \sqrt{\hat{V}_t} \]
9:  \[ \hat{U}_t = U_t / \max\left(1, \text{RMS}(U_t)/d\right) \]
10: \[ X_t = X_{t-1} - \alpha_t \hat{U}_t \]
11: end for

Update first moment estimate

No momentum
Adam vs. Adafactor

Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_{2t}\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $\delta$
2: for $t = 1$ to $T$
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_{2t} R_{t-1} + (1 - \beta_{2t}) (G_t^2 + \epsilon_1 1_n 1_m^T) 1_m$
6: $C_t = \beta_{2t} C_{t-1} + (1 - \beta_{2t}) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_n^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / \delta)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for

Update second moment estimate
Update low-rank second moment estimate
Algorithm 1 Adam (Kingma & Ba, 2015)

1: Inputs: initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4: \[ g_t = \nabla f_t(x_{t-1}) \]
5: \[ m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t \]
6: \[ v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2 \]
7: \[ \hat{m}_t = m_t / (1 - \beta_1^t) \]
8: \[ \hat{v}_t = v_t / (1 - \beta_2^t) \]
9: \[ x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) \]
10: end for

Perform bias correction

Algorithm 4 AdaFactor for weight matrices.

1: Inputs: initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3: \[ \alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t \]
4: \[ G_t = \nabla f_t(X_{t-1}) \]
5: \[ R_t = \beta_{2t} R_{t-1} + (1 - \beta_{2t})(G_t^2 + \epsilon_1 1_n 1_m^T) 1_m \]
6: \[ C_t = \beta_{2t} C_{t-1} + (1 - \beta_{2t}) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T) \]
7: \[ \hat{V}_t = R_t C_t / 1_n^T R_t \]
8: \[ U_t = G_t / \sqrt{\hat{V}_t} \]
9: \[ \hat{U}_t = U_t / \max (1, \text{RMS}(U_t) / d) \]
10: \[ X_t = X_{t-1} - \alpha_t \hat{U}_t \]
11: end for

No bias correction needed
Adam vs. Adafactor

**Algorithm 1** Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^{T}$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$
7: $\hat{m}_t = m_t/(1 - \beta_1^t)$
8: $\hat{v}_t = v_t/(1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t/\sqrt{\hat{v}_t + \epsilon}$
10: **end for**

**Algorithm 4** Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^{T}$, second moment decay $\{\beta_2\}_{t=1}^{T}$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: **for** $t = 1$ **to** $T$ **do**
3: $\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2 + \epsilon_1 1_n 1_m^\top)1_m$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2)1_n^\top(G_t^2 + \epsilon_1 1_n 1_m^\top)$
7: $\hat{V}_t = R_tC_t/1_n^\top R_t$
8: $U_t = G_t/\sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t/\max(1, \text{RMS}(U_t)/d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: **end for**

Compute update vector	Compute update vector
Adam vs. Adafactor

Algorithm 1 Adam (Kingma & Ba, 2015)

1: **Inputs:** initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: **for** $t = 1$ **to** $T$ **do**
4: \[ g_t = \nabla f_t(x_{t-1}) \]
5: \[ m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \]
6: \[ v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \]
7: \[ \hat{m}_t = m_t / (1 - \beta_1^t) \]
8: \[ \hat{v}_t = v_t / (1 - \beta_2^t) \]
9: \[ x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) \]
10: **end for**

Algorithm 4 Adafactor for weight matrices.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_{2t}\}_{t=1}^T$ such that $\beta_{21} = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: **for** $t = 1$ **to** $T$ **do**
3: \[ \alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t \]
4: \[ G_t = \nabla f_t(X_{t-1}) \]
5: \[ R_t = \beta_{2t} R_{t-1} + (1 - \beta_{2t})(G_t^2 + \epsilon_1 n_1 m_1) \]
6: \[ C_t = \beta_{2t} C_{t-1} + (1 - \beta_{2t}) n_1 m_1 (G_t^2 + \epsilon_1 n_1 m_1) \]
7: \[ \hat{V}_t = R_t C_t / n_1 m_1 \]
8: \[ U_t = G_t / \sqrt{\hat{V}_t} \]
9: \[ \hat{U}_t = U_t / \max(1, \text{RMS}(U_t) / d) \]
10: \[ X_t = X_{t-1} - \alpha_t \hat{U}_t \]
11: **end for**

Perform update clipping
Adam vs. Adafactor

Algorithm 1 Adam (Kingma & Ba, 2015)
1: Inputs: initial point $x_0$, step sizes $\{\alpha_t\}_{t=1}^T$, first moment decay $\beta_1$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $m_0 = 0$ and $v_0 = 0$
3: for $t = 1$ to $T$ do
4: $g_t = \nabla f_t(x_{t-1})$
5: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
6: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
7: $\hat{m}_t = m_t / (1 - \beta_1^t)$
8: $\hat{v}_t = v_t / (1 - \beta_2^t)$
9: $x_t = x_{t-1} - \alpha_t \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
10: end for

Algorithm 4 Adafactor for weight matrices.
1: Inputs: initial point $X_0 \in \mathbb{R}^{n \times m}$, relative step sizes $\{\rho_t\}_{t=1}^T$, second moment decay $\{\beta_2^t\}_{t=1}^T$ such that $\beta_2^1 = 0$, regularization constants $\epsilon_1$ and $\epsilon_2$, clipping threshold $d$
2: for $t = 1$ to $T$ do
3: $\alpha_t = \max (\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \hat{\beta_2} R_{t-1} + (1 - \hat{\beta_2}) (G_t^2 + \epsilon_1 1_n 1_m^T 1_m)$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2) 1_n^T (G_t^2 + \epsilon_1 1_n 1_m^T)$
7: $\hat{V}_t = R_t C_t / 1_n 1_m^T R_t$
8: $U_t = G_t / \sqrt{\hat{V}_t}$
9: $\hat{U}_t = U_t / \max (1, \text{RMS}(U_t) / d)$
10: $X_t = X_{t-1} - \alpha_t \hat{U}_t$
11: end for

Update parameters
Update parameters using relative step size
Key Changes in Adafactor

• Factored second moment estimation
• $\beta_2$ varies with time
• Update clipping
• Relative step sizes
• No momentum
Factored Second Moment Estimation
Factored Second Moment Estimation

- Consider a matrix-shaped parameter subset $X$ with second moment estimate $V$
Factored Second Moment Estimation

• Consider a matrix-shaped parameter subset $X$ with second moment estimate $V$

• Idea: want a low-rank representation $V \approx RS$ compatible with exponential moving averaging
Factored Second Moment Estimation

• Consider a matrix-shaped parameter subset $X$ with second moment estimate $V$

• Idea: want a low-rank representation $V \approx RS$ compatible with exponential moving averaging

• Formally: if factorization $F : V \mapsto (R, S)$, want $F(\beta V_{t-1} + (1 - \beta)G_t^2) = \beta F(V_{t-1}) + (1 - \beta)F(G_t^2)$
Factored Second Moment Estimation

Non-negative factorization using I-divergence:

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{n} \sum_{j=1}^{m} d(V_{ij}, [RS]_{ij}) \\
\text{subject to} & \quad R_{ij} \geq 0, S_{ij} \geq 0.
\end{align*}
\]

\[
d(p, q) = p \log(p/q) - p + q
\]
Factored Second Moment Estimation

Simple analytic solution set for rank-1 case:

\[ \{(R, S) : RS = V1_m1_n^T V / 1_n^T V 1_m\} \]
Factored Second Moment Estimation

Simple analytic solution set for rank-1 case:

\[ \{(R, S) : RS = V_1^\top m 1_n^\top V / 1_n^\top V 1_m\} \]

Row sums
Factored Second Moment Estimation

Simple analytic solution set for rank-1 case:

\[ \{(R, S) : RS = V1_m 1_n^\top V/1_n^\top V1_m \} \]

Row sums  Column sums
Factored Second Moment Estimation

Simple analytic solution set for rank-1 case:

\[
\{(R, S) : RS = V 1_m 1_n^\top V / 1_n^\top V 1_m \}
\]

- Row sums
- Column sums
- Sum of all entries
Factored Second Moment Estimation

Simple analytic solution set for rank-1 case:

\[ \{(R, S) : RS = V 1_m 1_n^\top V/1_n^\top V 1_m \} \]

Row sums  Column sums  Sum of all entries

Each component commutes with exponential moving averaging!
Factored Second Moment Estimation

Algorithm 2 Adam for a matrix parameter $X$ with factored second moments and first moment decay parameter $\beta_1 = 0$.

1: Inputs: initial point $X_0 \in \mathbb{R}^{n \times m}$, step sizes $\{\alpha_t\}_{t=1}^T$, second moment decay $\beta_2$, regularization constant $\epsilon$
2: Initialize $R_0 = 0$ and $C_0 = 0$
3: for $t = 1$ to $T$ do
4: $G_t = \nabla f_t(X_{t-1})$
5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2)1_m$
6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2)1_n (G_t^2)$
7: $\hat{V}_t = (R_t C_t/1_n R_t)/(1 - \beta_2^t)$
8: $X_t = X_{t-1} - \alpha_t G_t/(\sqrt{\hat{V}_t} + \epsilon)$
9: end for
Only need to keep track of moving averages of row and column sums
Factored Second Moment Estimation

**Algorithm 2** Adam for a matrix parameter $X$ with factored second moments and first moment decay parameter $\beta_1 = 0$.

1: **Inputs:** initial point $X_0 \in \mathbb{R}^{n \times m}$, step sizes $\{\alpha_t\}_{t=1}^T$, second moment decay $\beta_2$, regularization constant $\epsilon$

2: Initialize $R_0 = 0$ and $C_0 = 0$

3: **for** $t = 1$ **to** $T$ **do**

4: $G_t = \nabla f_t(X_{t-1})$

5: $R_t = \beta_2 R_{t-1} + (1 - \beta_2)(G_t^2)1_m$

6: $C_t = \beta_2 C_{t-1} + (1 - \beta_2)1_n^T (G_t^2)$

7: $\hat{V}_t = (R_tC_t/1_n^T R_t)/(1 - \beta_2^t)$

8: $X_t = X_{t-1} - \alpha_t G_t/\left(\sqrt{\hat{V}_t} + \epsilon\right)$

9: **end for**

Only need to keep track of moving averages of row and column sums

Can compute second moment estimate on the fly
A Problem with Adam
A Problem with Adam

• When $\beta_2$ is low, results are worse
  – 7-point gap in BLEU score for machine translation
A Problem with Adam

• When $\beta_2$ is low, results are worse
  – 7-point gap in BLEU score for machine translation

• When $\beta_2$ is high, training is unstable
  – Requires non-monotonic learning rate schedule with warm-up period followed by gradual decay
A Problem with Adam

Components of update $U_t$ should be close to 1, but diverge for large $\beta_2$.

$$U_t = \frac{G_t}{\sqrt{\hat{V}_t}}$$
Solution 1: Update Clipping

Scale down update whenever $\text{RMS}(U_t)$ exceeds a threshold value $d$:

$$\hat{U}_t = \frac{U_t}{\max(1, \text{RMS}(U_t)/d)}$$
Solution 2: Increasing $\beta_2$

Start from 0 and increase toward 1 to achieve best of both worlds:

$$\hat{\beta}_{2t} = 1 - \frac{1}{t^c}, \quad 0 < c \leq 1$$

Also avoids need for bias correction!
Relative Step Sizes

Training works well if the magnitudes of the parameter updates are about $10^{-2}$ to $10^{-3}$ times the magnitudes of the parameters.

―Geoff Hinton

$$\alpha_t = \max(\epsilon_2, \text{RMS}(X_{t-1})) \rho_t$$
Experiments

• English-German machine translation task
• State-of-the-art Transformer model
• WMT 2014 dataset
• Trained all models on TPU for 100,000 steps
<table>
<thead>
<tr>
<th>Factored Second-Moment Estimation</th>
<th>$\hat{\beta}_{1t}$</th>
<th>$\hat{\beta}_{2t}$</th>
<th>Update Clipping $d$</th>
<th>(Relative) Step Size</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C) yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D) use row-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E) use col-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>0</td>
<td>$\beta_2 = 0.99$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>0</td>
<td>$\beta_2 = 0.9$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H) yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td></td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>2.0</td>
<td></td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J) yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td></td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K)</td>
<td>0</td>
<td>$1 - t^{-0.5}$</td>
<td></td>
<td></td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td></td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>0</td>
<td>$1 - t^{-1.0}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td></td>
<td>25.9</td>
<td>22.4</td>
</tr>
<tr>
<td>(O) yes</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>$\rho_t = s_t$</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P) yes</td>
<td>0.9</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>$lr = 1 \cdot s_t$</td>
<td>24.9</td>
<td>25.3</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 10 \cdot s_t$</td>
<td></td>
<td></td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 100 \cdot s_t$</td>
<td></td>
<td></td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 150 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 200 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 300 \cdot s_t$</td>
<td></td>
<td></td>
<td>diverged</td>
<td>diverged</td>
</tr>
<tr>
<td>Factored Second-Moment Estimation</td>
<td>Update Clipping $d$</td>
<td>(Relative Step Size) $\alpha_t$</td>
<td>BLEU with warmup</td>
<td>BLEU no warmup</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------</td>
<td>---------------------------------</td>
<td>-----------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A)</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>(B)</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>23.1</td>
<td></td>
</tr>
<tr>
<td>(C) yes</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>(D) use row-mean</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.2</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>(E) use col-mean</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>0.3</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>(F)</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>(G)</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
<td>15.6</td>
<td></td>
</tr>
<tr>
<td>(H) yes</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 1$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>21.5</td>
<td></td>
</tr>
<tr>
<td>(I)</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 2$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.7</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>(J) yes</td>
<td>$\beta_1 = 0.999$</td>
<td>$\beta_2 = 1$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>(K)</td>
<td>$\beta_1 = 1 - t^{-0.5}$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>21.1</td>
<td></td>
</tr>
<tr>
<td>(L)</td>
<td>$\beta_1 = 1 - t^{-0.8}$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>(M)</td>
<td>$\beta_1 = 1 - t^{-1.0}$</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>(N)</td>
<td>$\beta_1 = 1 - t^{-0.8}$</td>
<td>$\beta_2 = 1$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td><strong>25.9</strong></td>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>(O) yes</td>
<td>$\beta_1 = 1 - t^{-0.8}$</td>
<td>$\beta_2 = 1$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>25.5</td>
<td></td>
</tr>
<tr>
<td>(P) yes</td>
<td>$\beta_1 = 0.9$</td>
<td>$\beta_2 = 1 - t^{-0.8}$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>24.9</td>
<td>25.3</td>
<td></td>
</tr>
</tbody>
</table>

<p>| (Q)                              | SGD                 | $lr = 1 \cdot s_t$            | 0.6  | 0.8 |
| (Q)                              | SGD                 | $lr = 10 \cdot s_t$           | 8.2  | 9.1 |
| (Q)                              | SGD                 | $lr = 100 \cdot s_t$          | 22.9 | diverged |
| (Q)                              | SGD                 | $lr = 150 \cdot s_t$          | 24.0 | diverged |
| (Q)                              | SGD                 | $lr = 200 \cdot s_t$          | 24.3 | diverged |
| (Q)                              | SGD                 | $lr = 300 \cdot s_t$          | diverged | diverged |</p>
<table>
<thead>
<tr>
<th>Factored</th>
<th>Second-Moment Estimation</th>
<th>Update Clipping ( d )</th>
<th>(Relative) Step Size</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C) yes</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D) use row-mean</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E) use col-mean</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.9 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.9 )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H) yes</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>1.0</td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>2.0</td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J) yes</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( \hat{\beta}_2 = 0.999 )</td>
<td>1.0</td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K) yes</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-0.5} )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-0.8} )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-1.0} )</td>
<td>( \alpha_t = 0.1 \cdot s_t )</td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-0.8} )</td>
<td>1.0</td>
<td>25.9</td>
<td>22.4</td>
</tr>
<tr>
<td>(O) yes</td>
<td>0 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-0.8} )</td>
<td>1.0</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P) yes</td>
<td>0.9 ( \hat{\beta}_1 = 0 )</td>
<td>( 1 - t^{-0.8} )</td>
<td>1.0</td>
<td>24.9</td>
<td>25.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>( lr ) Setting</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>( lr = 1 \cdot s_t )</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>SGD</td>
<td>( lr = 10 \cdot s_t )</td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td>SGD</td>
<td>( lr = 100 \cdot s_t )</td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td>SGD</td>
<td>( lr = 150 \cdot s_t )</td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td>SGD</td>
<td>( lr = 200 \cdot s_t )</td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td>SGD</td>
<td>( lr = 300 \cdot s_t )</td>
<td>diverged</td>
<td>diverged</td>
</tr>
<tr>
<td>Factored Second-Moment Estimation</td>
<td>Update Clipping Step Size</td>
<td>(Relative)</td>
<td>BLEU with warmup</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------------------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td>(A)</td>
<td>$\hat{\beta}_1$ = 0, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
</tr>
<tr>
<td>(C)</td>
<td>yes, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
</tr>
<tr>
<td>(D)</td>
<td>use row-mean, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.2</td>
</tr>
<tr>
<td>(E)</td>
<td>use col-mean, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>0.3</td>
</tr>
<tr>
<td>(F)</td>
<td>0, $\hat{\beta}_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
</tr>
<tr>
<td>(G)</td>
<td>0, $\hat{\beta}_2 = 0.9$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
</tr>
<tr>
<td>(H)</td>
<td>yes, $\hat{\beta}_2 = 0.999$</td>
<td>1.0</td>
<td>25.4</td>
</tr>
<tr>
<td>(I)</td>
<td>0, $\hat{\beta}_2 = 0.999$</td>
<td>2.0</td>
<td>25.7</td>
</tr>
<tr>
<td>(J)</td>
<td>yes, $\hat{\beta}_2 = 0.999$</td>
<td>1.0</td>
<td>25.6</td>
</tr>
<tr>
<td>(K)</td>
<td>0, $1 - t^{-0.5}$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
</tr>
<tr>
<td>(L)</td>
<td>0, $1 - t^{-0.8}$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
</tr>
<tr>
<td>(M)</td>
<td>0, $1 - t^{-1.0}$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
</tr>
<tr>
<td>(N)</td>
<td>0, $1 - t^{-0.8}$</td>
<td>1.0</td>
<td>25.9</td>
</tr>
<tr>
<td>(O)</td>
<td>yes, $1 - t^{-0.8}$</td>
<td>1.0</td>
<td>25.0</td>
</tr>
<tr>
<td>(P)</td>
<td>yes, 0.9, $1 - t^{-0.8}$</td>
<td>1.0</td>
<td>24.9</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 1 \cdot s_t$</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 10 \cdot s_t$</td>
<td></td>
<td>8.2</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 100 \cdot s_t$</td>
<td></td>
<td>22.9</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 150 \cdot s_t$</td>
<td></td>
<td>24.0</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 200 \cdot s_t$</td>
<td></td>
<td>24.3</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD, $lr = 300 \cdot s_t$</td>
<td></td>
<td>diverged</td>
</tr>
</tbody>
</table>

Vanilla SGD
<table>
<thead>
<tr>
<th>Factored Second-Moment Estimation</th>
<th>$\hat{\beta}_{1t}$</th>
<th>$\hat{\beta}_{2t}$</th>
<th>Update Clipping $d$</th>
<th>(Relative) Step Size</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td></td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>2.0</td>
<td></td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td></td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-0.5}$</td>
<td></td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-1.0}$</td>
<td></td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td>25.9</td>
<td>22.4</td>
</tr>
<tr>
<td>(O)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-0.8}$</td>
<td>$\rho_t = s_t$</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P)</td>
<td>yes</td>
<td>$\beta_2 = 0.999$</td>
<td>$1 - t^{-0.8}$</td>
<td>$lr = 1 \cdot s_t$</td>
<td>24.9</td>
<td>25.3</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD</td>
<td>$lr = 1 \cdot s_t$</td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 10 \cdot s_t$</td>
<td></td>
<td></td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 100 \cdot s_t$</td>
<td></td>
<td></td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 150 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 200 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 300 \cdot s_t$</td>
<td></td>
<td></td>
<td>diverged</td>
<td>diverged</td>
</tr>
</tbody>
</table>

Adam with factored second moment estimation
<table>
<thead>
<tr>
<th>Factored Second-Moment Estimation</th>
<th>$\hat{\beta}_{1t}$</th>
<th>$\hat{\beta}_{2t}$</th>
<th>Update Clipping $d$</th>
<th>(Relative Step Size)</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D)</td>
<td>use row-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E)</td>
<td>use col-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>0</td>
<td>$\beta_2 = 0.9$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>0</td>
<td>$\beta_2 = 0.9$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>2.0</td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K)</td>
<td>0</td>
<td>$1 - t^{-0.5}$</td>
<td></td>
<td></td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td></td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>0</td>
<td>$1 - t^{-1.0}$</td>
<td></td>
<td></td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td></td>
<td>25.9</td>
<td>22.4</td>
</tr>
<tr>
<td>(O)</td>
<td>yes</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P)</td>
<td>yes</td>
<td>0.9</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>24.9</td>
<td>25.3</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD</td>
<td>$lr = 1 \cdot s_t$</td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 10 \cdot s_t$</td>
<td></td>
<td></td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 100 \cdot s_t$</td>
<td></td>
<td></td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 150 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 200 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 300 \cdot s_t$</td>
<td></td>
<td></td>
<td>diverged</td>
<td>diverged</td>
</tr>
</tbody>
</table>

Update clipping
<table>
<thead>
<tr>
<th>Factored Second-Moment Estimation</th>
<th>$\hat{\beta}_{1t}$</th>
<th>$\hat{\beta}_{2t}$</th>
<th>Update Clipping $d$</th>
<th>(Relative) Step Size</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D)</td>
<td>use row-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E)</td>
<td>use col-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>0</td>
<td>$\beta_2 = 0.99$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>0</td>
<td>$\beta_2 = 0.9$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>2.0</td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K)</td>
<td>0</td>
<td>$1 - t^{-0.5}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td></td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>0</td>
<td>$1 - t^{-1.0}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.9</td>
<td>22.4</td>
</tr>
<tr>
<td>(O)</td>
<td>yes</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P)</td>
<td>yes</td>
<td>0.9</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>24.9</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 1 \cdot s_t$</td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 10 \cdot s_t$</td>
<td></td>
<td></td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 100 \cdot s_t$</td>
<td></td>
<td></td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 150 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 200 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 300 \cdot s_t$</td>
<td></td>
<td></td>
<td>diverged</td>
<td>diverged</td>
</tr>
</tbody>
</table>

Increasing $\beta_2$
<table>
<thead>
<tr>
<th>Factored Second-Moment Estimation</th>
<th>$\hat{\beta}_{1t}$</th>
<th>$\hat{\beta}_{2t}$</th>
<th>Update Clipping $d$</th>
<th>(Relative) Step Size</th>
<th>BLEU with warmup</th>
<th>BLEU no warmup</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(B)</td>
<td>0.9</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>23.1</td>
</tr>
<tr>
<td>(C)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.2</td>
</tr>
<tr>
<td>(D)</td>
<td>use row-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(E)</td>
<td>use col-mean</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>(F)</td>
<td>0</td>
<td>$\beta_2 = 0.99$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.0</td>
<td>0.4</td>
</tr>
<tr>
<td>(G)</td>
<td>0</td>
<td>$\beta_2 = 0.9$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>(H)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.4</td>
<td>21.5</td>
</tr>
<tr>
<td>(I)</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td></td>
<td>2.0</td>
<td>25.7</td>
<td>0.2</td>
</tr>
<tr>
<td>(J)</td>
<td>yes</td>
<td>0</td>
<td>$\beta_2 = 0.999$</td>
<td>1.0</td>
<td>25.6</td>
<td>22.4</td>
</tr>
<tr>
<td>(K)</td>
<td>0</td>
<td>$1 - t^{-0.5}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>21.1</td>
</tr>
<tr>
<td>(L)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.6</td>
<td>0.1</td>
</tr>
<tr>
<td>(M)</td>
<td>0</td>
<td>$1 - t^{-1.0}$</td>
<td></td>
<td>$\alpha_t = 0.1 \cdot s_t$</td>
<td>25.4</td>
<td>0.1</td>
</tr>
<tr>
<td>(N)</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td></td>
<td>1.0</td>
<td><strong>25.9</strong></td>
<td>22.4</td>
</tr>
<tr>
<td>(O)</td>
<td>yes</td>
<td>0</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>25.0</td>
<td>25.5</td>
</tr>
<tr>
<td>(P)</td>
<td>yes</td>
<td>0.9</td>
<td>$1 - t^{-0.8}$</td>
<td>1.0</td>
<td>24.9</td>
<td>25.3</td>
</tr>
<tr>
<td>(Q)</td>
<td>SGD</td>
<td>$lr = 1 \cdot s_t$</td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 10 \cdot s_t$</td>
<td></td>
<td></td>
<td>8.2</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 100 \cdot s_t$</td>
<td></td>
<td></td>
<td>22.9</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 150 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.0</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 200 \cdot s_t$</td>
<td></td>
<td></td>
<td>24.3</td>
<td>diverged</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>$lr = 300 \cdot s_t$</td>
<td></td>
<td></td>
<td>diverged</td>
<td>diverged</td>
</tr>
</tbody>
</table>
Algorithm 6 Proposed hyperparameters for Adafactor

1: $\epsilon_1 = 10^{-30}$
2: $\epsilon_2 = 10^{-3}$
3: $d = 1$
4: $\rho_t = \min \left( 10^{-2}, \frac{1}{\sqrt{t}} \right)$
5: $\hat{\rho}_{2t} = 1 - t^{-0.8}$
**Algorithm 6** Proposed hyperparameters for Adafactor

1. \( \epsilon_1 = 10^{-30} \)
2. \( \epsilon_2 = 10^{-3} \)
3. \( d = 1 \)
4. \( \rho_t = \min \left( 10^{-2}, \frac{1}{\sqrt{t}} \right) \)
5. \( \hat{\beta}_{2t} = 1 - t^{-0.8} \)

<table>
<thead>
<tr>
<th>Emb init. ( \sigma )</th>
<th>Multiplier ( d_{model} )</th>
<th>BLEU (Adam)</th>
<th>BLEU (Adafactor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{\sqrt{d_{model}}} )</td>
<td>( \sqrt{d_{model}} )</td>
<td>26.4</td>
<td><strong>26.6</strong></td>
</tr>
<tr>
<td>( \frac{1}{\sqrt{d_{model}}} )</td>
<td>1</td>
<td>25.8</td>
<td>26.4</td>
</tr>
<tr>
<td>( \frac{1}{\sqrt{d_{model}}} )</td>
<td>1</td>
<td>24.2</td>
<td>25.4</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion

- Adafactor is a memory-efficient adaptive learning method
Conclusion

• Adafactor is a memory-efficient adaptive learning method
• Adafactor matches the performance of Adam on large-scale deep learning problems
Conclusion

- Adafactor is a memory-efficient adaptive learning method
- Adafactor matches the performance of Adam on large-scale deep learning problems
- Code available in the Tensor2Tensor library: https://github.com/tensorflow/tensor2tensor
Thanks!