lèa R›c©x ÿ Attention Maskingÿ v„ep[f[žs°N;‰•u(NŽW(Transformerg¶g„N-R¨` c§R6N T tokenNK•ôv„N¤

1. **Q³{Vc©x u b **
   ˜„mKj!WW•“QúkÏN*tokenv„OÝuYi,s‡ \(\pi \in \mathbb{R}^{N \times 2}\)ÿ • •ÇGumbel-Softmax‘Çh7_—
   \[
   \hat{D} \leftarrow \hat{D} \odot D
   \]
   QvN- \(\odot\) N:• QC} NXlÕÿ xnOÝˆ«Rjg•v„tokenN Q•SÂN T ~í‹¡{—0

2. **lèa R›c©x wé–5g„^ú**
   [šNIN¤N’wé–5 \(G \in \mathbb{R}^{N \times N}\)ÿ c§R6token \(j\) f/T&_qTÍtoken \(i\)ÿ
   \[
   G_{ij} = \begin{cases}
   1 & \text{if } i=j \ (\text{•ê•Þc¥OÝuY}) \\
   \hat{D}_j & \text{if } i \neq j \ (\text{NÅOÝuYg*Rjg•tokenv„N¤N’})
   \end{cases}
   \]

3. **c©x lèa R›‹¡{—**
   W(SoftmaxRM^”u(c©x \(G\)ÿ \O…=ˆ«Rjg•tokenv„•!s.ÿ
   \[
   \tilde{A}_{ij} = \frac{\exp(P_{ij}) G_{ij}}{\sum_{k=1}^N \exp(P_{ik}) G_{ik}}, \quad P = QK^T / \sqrt{C}
   \]
   QvN- \(P\) N:SŸYËlèa R›R epÿ c©x T v„ \(\tilde{A}\) NÅS T+g eHtokenv„N¤N’0

O R¿ÿ
- **xlNöSËY}**ÿ OÝc _ ‘Ï_br¶N SØÿ \(N \times N\)ÿ ÿ e/c ^vˆL‹¡{—0
- **Sï_®R **ÿ • •ÇGumbel-SoftmaxTŒc©x g:R6[žs°zïR0zï‹-~Ã0
- **R¨` `'**ÿ •“QeO••Vv„Rjg•{VuecÐSG‹¡{—eHs‡ÿ Y,‹ºe‡N-QÏ\ 31%~37% FLOPsÿ 0

Y,— •ÛN keNãx [žs°~Æ„ÿ SïSÂ€ ‹ºe‡_ n•NÓ^“0