各単語の予測確率は、Transformerが最終層の位置$i-1$において計算した単語埋め込み$H_k^{(L)} \in \mathbb{R}^d$を用いて

$$
\begin{aligned}
H^{(0)} &= WX + P \\
H^{(l)} &= \mathrm{transformer\_block}(H^{(l-1)}), \forall l \in \{1, \ldots, L\} \\
P(x_i | x_{i-k}, x_{i-k+1}, \ldots, x_{i-1}) &= \mathrm{softmax}_{x_i}(W^\top H_k^{(L)})
\end{aligned}
$$

と計算する。ここで、$X \in \mathbb{R}^{|V| \times k}$は文脈単語$x_{i-k}, x_{i-k+1}, \ldots, x_{i-1}$のワンホットベクトルを横方向に並べた行列、$W \in \mathbb{R}^{d \times |V|}$は単語埋め込み行列、$P \in \mathbb{R}^{d \times k}$は位置埋め込み行列、$H^{(l)} \in \mathbb{R}^{d \times k}$ $(l \in \{0, \ldots, L\})$は、Transformerの第l層における文脈単語の埋め込み表現を横方向に並べた行列である。