GPTの事前学習

GPTはTransformerのデコーダにより、入力されたテキストから次の単語を予測するという言語モデルの学習を行う。事前学習に用いる、長さがN単語のテキストの単語列を$s = x_1, x_2, \ldots, x_N$で表す。位置$i$の単語$x_i$を予測するとき、それよりも$k$個前に出現する単語列$x_{i-k}, x_{i-k+1}, \ldots, x_{i-1}$を文脈として用い、言語モデルの負の対数尤度

$$J = -\sum_{i=k+1}^{N} \log P(x_i | x_{i-k}, x_{i-k+1}, \ldots, x_{i-1})$$

を最小化するように、L層のTransformerのデコーダ部分を学習する。