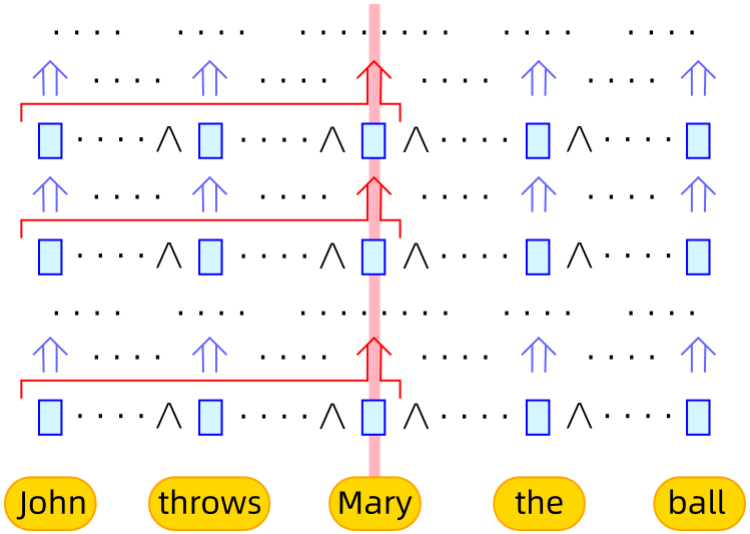
Further Interpreting the Transformer

This is an illustration of how the Transformer processes tokens:



(1)

- Each pass of Self-Attention has an axis (thick red line), where a pivot token is combined with other tokens via super-position.
  Why does it
- Each token seems to be the "summary" of the incoming tokens up to time t. Such a summary seems to be a distributive representation made of superpositions, but due to softmax, it may also be a disentangled representation closer to logic.