Xét Decoder dịch "Xin chào" thành "Hello", sử dụng output của Encoder đã tính ở

Thiết lập tham số:

* Output Encoder (​): (Shape: 2x4)

* Đầu vào: <SOS>
* Tham số: .
* Tham số LayerNorm: Giả sử và .
* Từ vựng:

Bước A: Tạo Token thứ nhất ("Hello")

Đầu vào: Token <start>

* B1: Target Input Embedding: =
* B2: Target Positional Encoding (PE):
* B3: Masked Multi-Head Self-Attention:
  + Head 1:

Giả sử các ma trận trọng số

* + - (Sau mask và softmax)
  + Head 2:

Giả sử các ma trận trọng số

* + - (Sau mask và softmax)
  + Concatenate & Final Linear:
* B4: Add & Norm 1:
* B5: Multi-Head Cross-Attention:
  + Head 1:

Giả sử các ma trận trọng số

* + - (Từ X\_norm1\_dec)
    - (Từ E\_out)
    - (Từ E\_out)
    - (Softmax)
  + Head 2:

Giả sử các ma trận trọng số

* + - (Từ X\_norm1\_dec)
    - (Từ E\_out)
    - (Từ E\_out)
    - (Softmax)
  + Concatenate & Final Linear:
* B6: Add & Norm 2:
* B7: Feed-Forward Network (FFN):

Giả sử các ma trận tham số như sau

* B8: Add & Norm 3:
* B9: Final Linear & Softmax:

Giả sử các ma trận tham số như sau

* + Input:
  + (Shape 1x10).
* Logits cao nhất ở index 5
* Probabilities cao nhất ở index 5

Từ vựng:

* + Dự đoán: Token "Hello".

Bước B: Tạo Token thứ hai (<EOS>)

Thiết lập:

* Output Encoder (Shape: 2x4)
* Chuỗi đã tạo: <SOS> Hello
* Tham số: .
* Embeddings:
* Các ma trận trọng số và bias của Decoder: Sử dụng lại các trọng số đã giả định ở bước dự đoán "Hello".
* Tham số LayerNorm: Giả sử và .
* Từ vựng: (Mục tiêu là index 2).

Đầu vào: Chuỗi <SOS> Hello

* B1: Target Input Embedding:
* B2: Target Positional Encoding (PE):
* B3: Masked Multi-Head Self-Attention: (Xử lý cả <start> và "Hello")
  + Head 1:

Giả sử các ma trận tham số như sau

* + - Tính (Từ )
    - Tính (Từ )
    - Tính (Từ )

    - Áp dụng Mask :
    - (Softmax theo hàng)
  + Head 2:

Giả sử các ma trận trọng số

* + - Tính
    - Tính
    - V2 khoảng [[1.5, 2.0], [0.74, 2.31]]
    - khoảng [[2.25, 3.66], [2.37, 4.05]]
    - Áp dụng Mask :
  + Concatenate & Final Linear:

    - (Do là ma trận đơn vị)
* B4: Add & Norm 1:

* + Lấy vector cuối cùng (ứng với "Hello"):
* B5: Multi-Head Cross-Attention: (Chú ý đến , Query từ )
  + Head 1:

Giả sử các ma trận trọng số

* + - (Chiếu từ )
    - (Chiếu từ )
    - (Chiếu từ )
    - (Softmax)
  + Head 2:

Giả sử các ma trận trọng số

* + - (Chiếu từ last\_vec)
    - (Chiếu từ )
    - (Chiếu từ )
    - (Softmax)
  + Concatenate & Final Linear:
    - (Do là ma trận đơn vị)
* B6: Add & Norm 2:
  + (Sau LayerNorm)
* B7: Feed-Forward Network (FFN):

Giả sử các ma trận tham số như sau

* B8: Add & Norm 3:
  + (Sau LayerNorm)
* B9: Final Linear & Softmax:

Giả sử các ma trận tham số như sau

* + Input:
  + (Shape 1x10).
* Logit cao nhất ở index 2
  + Probabilities = softmax(Logits) (Shape 1x10).
* Xác suất cao nhất ở index 2

Từ vựng:

* + Dự đoán: Token <EOS>.