

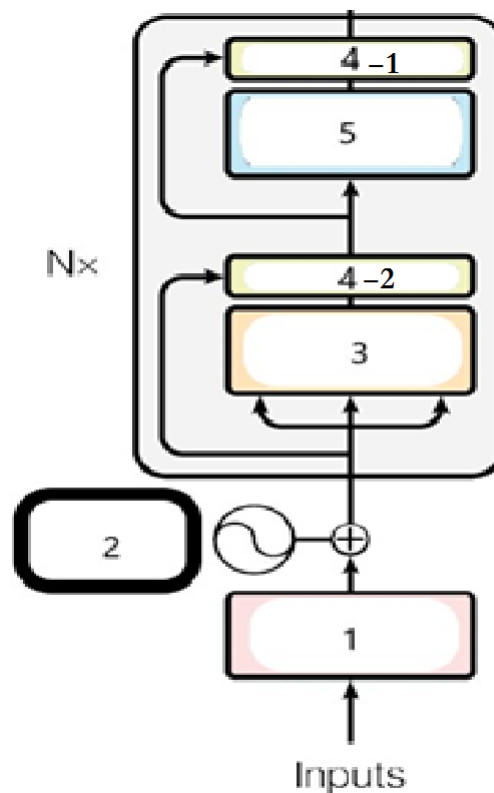
**سوال ۱)** در الگوریتم Beam-Search اگر سایز beam را افزایش بدهیم، کدام موارد اتفاق می افتد؟ (جواب می تواند چند گزینه باشد). (۴ نمره- پاسخ نادرست منفی یک نمره) **گزینه a,b,c**

- این الگوریتم کندتر اجرا می شود.
- این الگوریتم به حافظه بیشتری نیاز خواهد داشت.
- این الگوریتم بطور کلی جواب بهتری را پیدا خواهد کرد.
- این الگوریتم با تعداد مراحل کمتری همگرا می شود.

**سوال ۲)** برای کدام کاربردهای زیر معماری Encoder-Decoder مناسب است؟ (جواب می تواند چند گزینه باشد). (۴ نمره- پاسخ نادرست منفی یک نمره) **گزینه a,d**

- Summarization
- Image captioning
- Sentiment Analysis
- Machine translation

**سوال ۳)** تصویر زیر معماری یک بلاک از بخش انکودر ترنسفرمر را نشان می دهد. جاهای خالی را پر کرده و کاربرد آن ها را بطور مختصر توضیح دهید. (۲۰ نمره)



مثال:

۱- امبدینگ توکن ها: این بخش متن ورودی را به یک بازنمایی عددی نگاشت می کند تا شبکه بتواند با اعداد کار کند.

## پاسخ

2- **Positional encoding/embedding**: it provides the transformer model with information about where the words are in the input sequence.

3- **Multi-headed Self-attention**: where the representation of a sequence (or sentence) is computed by relating different words in the same sequence.

4-2- **Residual Connections**: Residual connections alleviate unstable gradient and aids in preserving information across different layers of the network.

4-1- **Layer Normalization**: layer norm just normalizes each layer with the mean and variance of each activation. Layer Normalization improves both the training time and the generalization performance

سوال (۴) کدام ماژول یا ماژول‌ها از معماری ترنسفرمر مشابه مکانیسم gating (مشابه آنچه در lstm و gru دیدیم) عمل می‌کند؟ چرا؟ (۹ نمره)

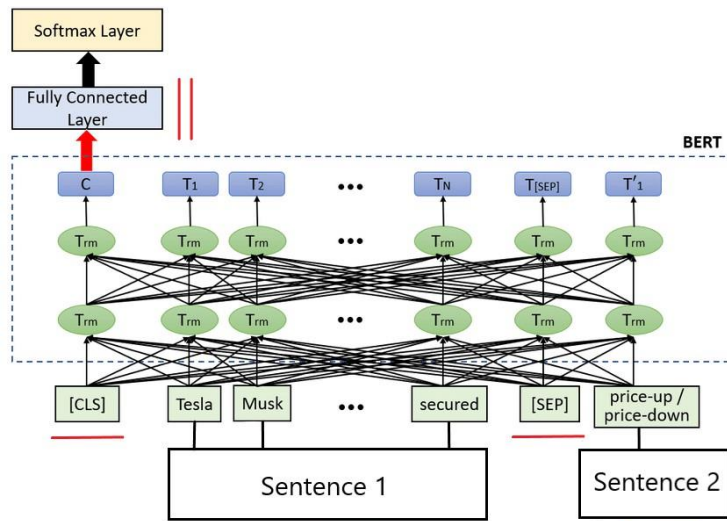
In the Transformer model, residual connections act as a gating mechanism, similar to those in GRUs and LSTMs. These connections allow the model to pass a portion of the input directly to deeper layers, effectively bypassing certain transformations. This mechanism helps in mitigating the vanishing gradient problem and aids in preserving information across different layers of the network.

اگر فقط نام ماژول را ذکر گردد ۲۰ درصد نمره داده خواهد شد. اگر دلیل درست نوشته شود، حتی اگر نام ماژول ذکر نگردد، کل نمره داده خواهد شد.

سوال (۵) چگونه می‌توان دو تسک NLI و NER را به وسیله‌ی مدل برت انجام داد؟ ترجیحا با رسم شکل پاسخ دهید. (۱۵ نمره)

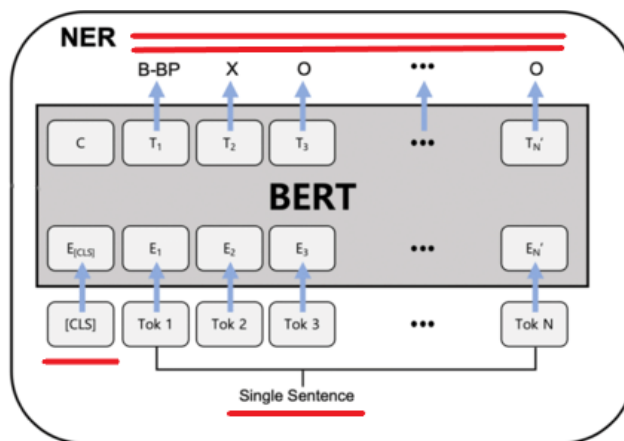
NLI یا Natural Language Inference تسکی است که در آن دو جمله ارائه می‌گردد و می‌بایست نسبت جمله‌ی دوم با جمله‌ی اول تشخیص داده شود. مثلا تشخیص داده شود که آیا از جمله‌ی اول می‌توان جمله‌ی دوم را نتیجه گرفت یا خیر).

A- NLI:



اگر لایه feed forward روی خروجی توکن‌ها (T ها یا C با بدون C) اعمال شود نیز پاسخ درست لحاظ می‌گردد.

## B- NER:



بخش متناظر با هر خط قرمز ۱۰ درصد نمره را دارد.

**سوال ۶)** به سوالات زیر پاسخ کوتاه بدهید. هر کدام ۶ نمره

a. Context vector در معماری Encoder-Decoder چیست؟

چکیده ورودی را به دیکودر منتقل می‌کند. به عبارت دیگر contextualized representation از کل ورودی است که در آخرین وضعیت مخفی قرار دارد.

b. چرا نمونه برداری از توزیع softmax برای تولید خروجی در Decoder مناسب نیست؟

دو دلیل میتوان متصور شد: ۱. این روش حریصانه است و شاید خروجی آخر بهترین نباشد. ۲. ممکن است جواب خیلی بد هم تولید شود. چون به هر حال شانس انتخاب بد، هر چند کم وجود دارد.

c. در الگوریتم Beam-Search وقتی یکی از خروجی‌ها به توکن </S> می‌رسد الگوریتم در ادامه چی کار می‌کند؟ در اینجا، این دنباله از frontier حذف می‌شود و سائز beam یکی کاهش پیدا می‌کند. و این دنباله به عنوان یک کاندید نهایی انتخاب می‌شود. در آخر بهترین کاندید نهایی به عنوان جواب انتخاب خواهد شد.

f) تفاوت self attention و cross attention در چیست و در ترنسفرمر ها از کدام نوع attention استفاده می‌گردد؟

e- Self-Attention: Focuses on relations within the same sequence. Cross-Attention: Consider the relations between different sequences – Both

g) دلیل استفاده از Muti head self attention چیست؟

g- Each head can learn to focus on different aspects, e.g., subject and object of a sentence.

h) دلیل استفاده از ماسک در مکانیسم attention، هنگام آموزش دیکدر در معماری ترنسفورمر چیست؟

h- The attention mask in the Transformer decoder makes sure that during the self-attention part, each token can only attend on tokens that come before it in the sequence.

پاسخ