NLP Project 3

Transformer

Chris Ding, Haoran Li and Dazhi Zhu

2025.11.20

Overview

- There will be 4 projects in total:
  - Project 1 and 2: Word2Vec (TA: Yue Lin)
    - Project 1: CBOW; Project 2: Skip-Gram
    - (We will continue to use materials from previous years, but GLoVe will not be included.)
  - Project 3: Transformer (TA: Haoran Li)
  - Project 4: BERT (TA: Haoran Li)

Transformer

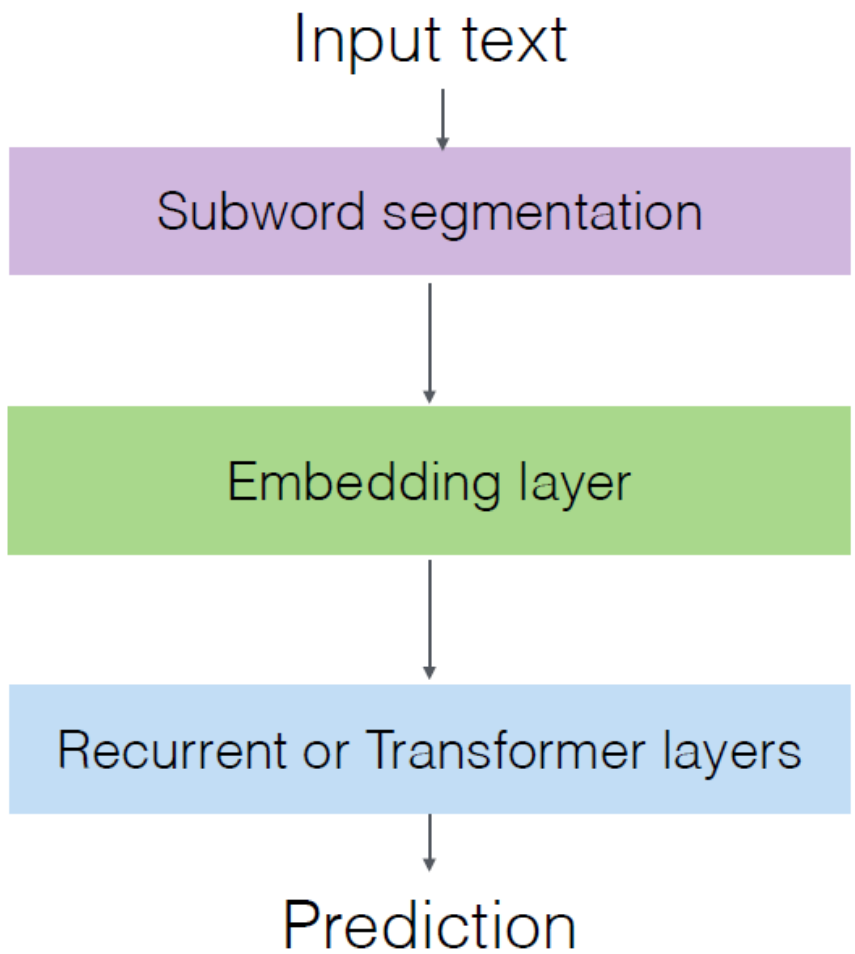
- The Transformer is a revolutionary deep learning architecture introduced in the 2017 paper Attention Is All You Need by Vaswani et al. from Google. It replaces traditional recurrent (RNN/LSTM) and convolutional (CNN) layers with a self-attention mechanism, enabling unprecedented performance in sequence modeling tasks like machine translation, text generation, and more.

Transformer

Can solve following NLP tasks

- Translation, e.g., Chinese to English
- Named entity recognition
- Text classification
- Text summarization
- Question and answering
- Dialogue generation

Standard pipeline for neural text processing



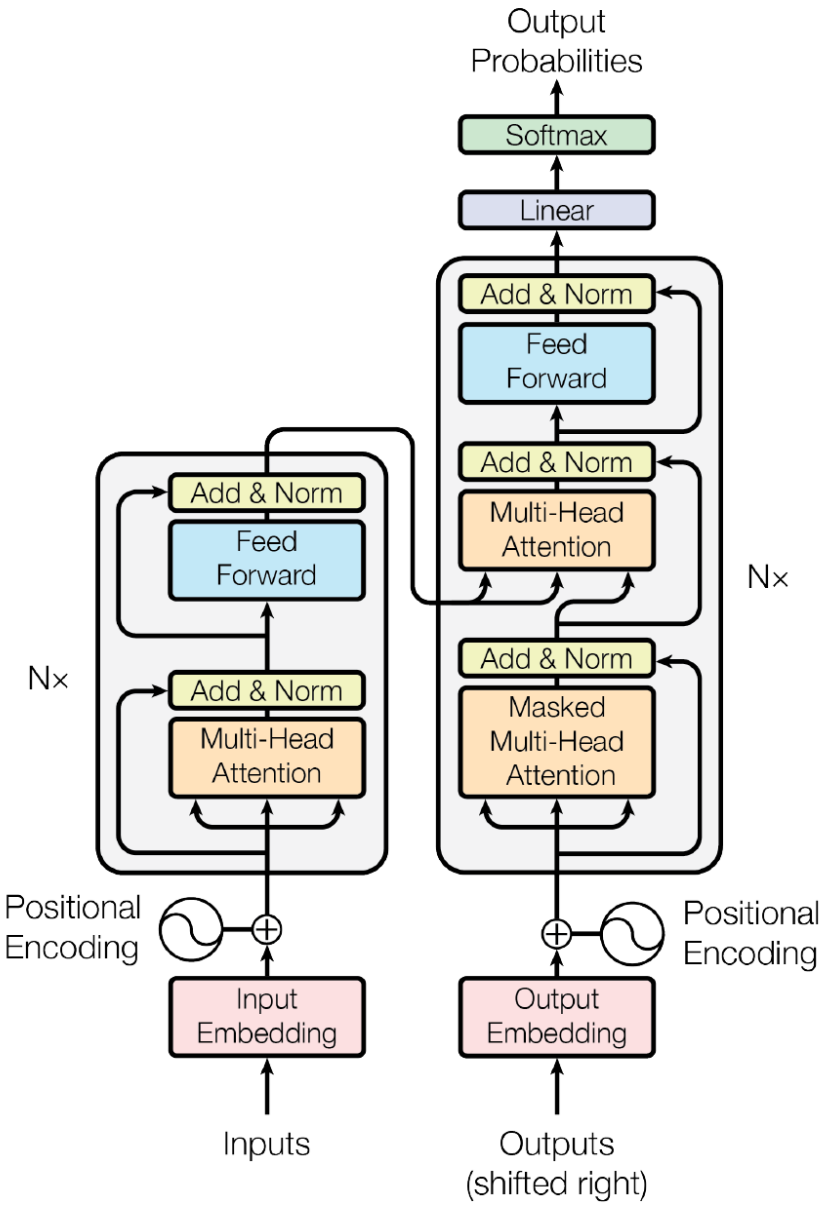


Figure 1: The Transformer - model architecture.

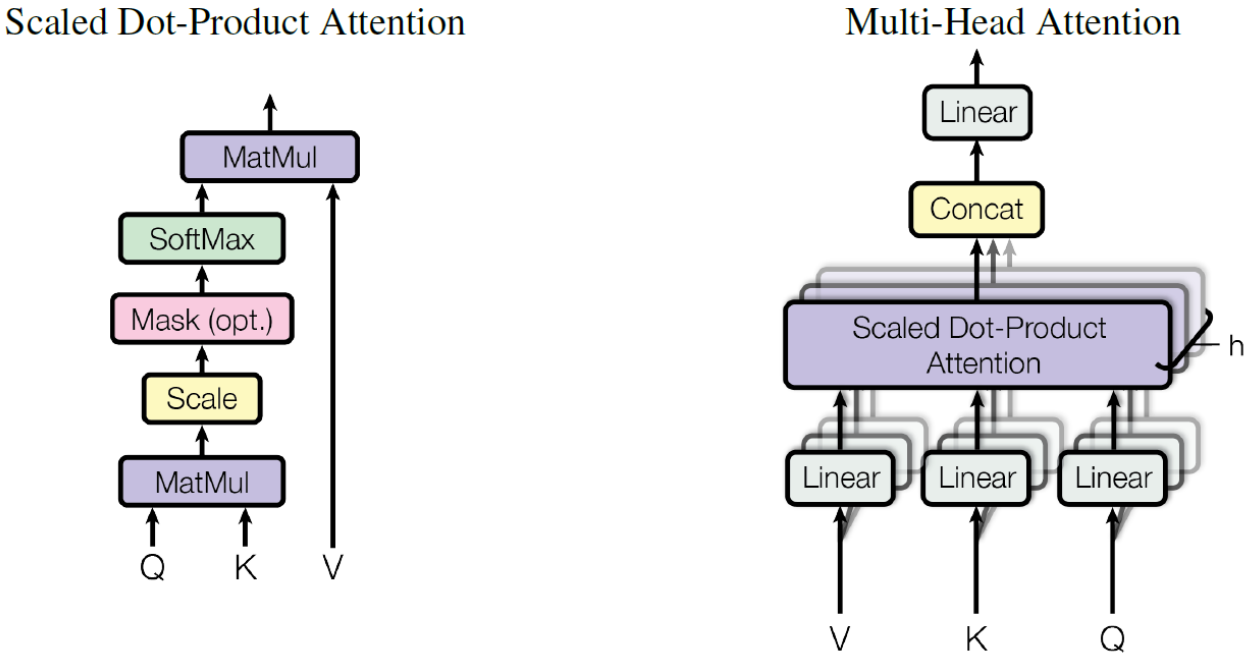
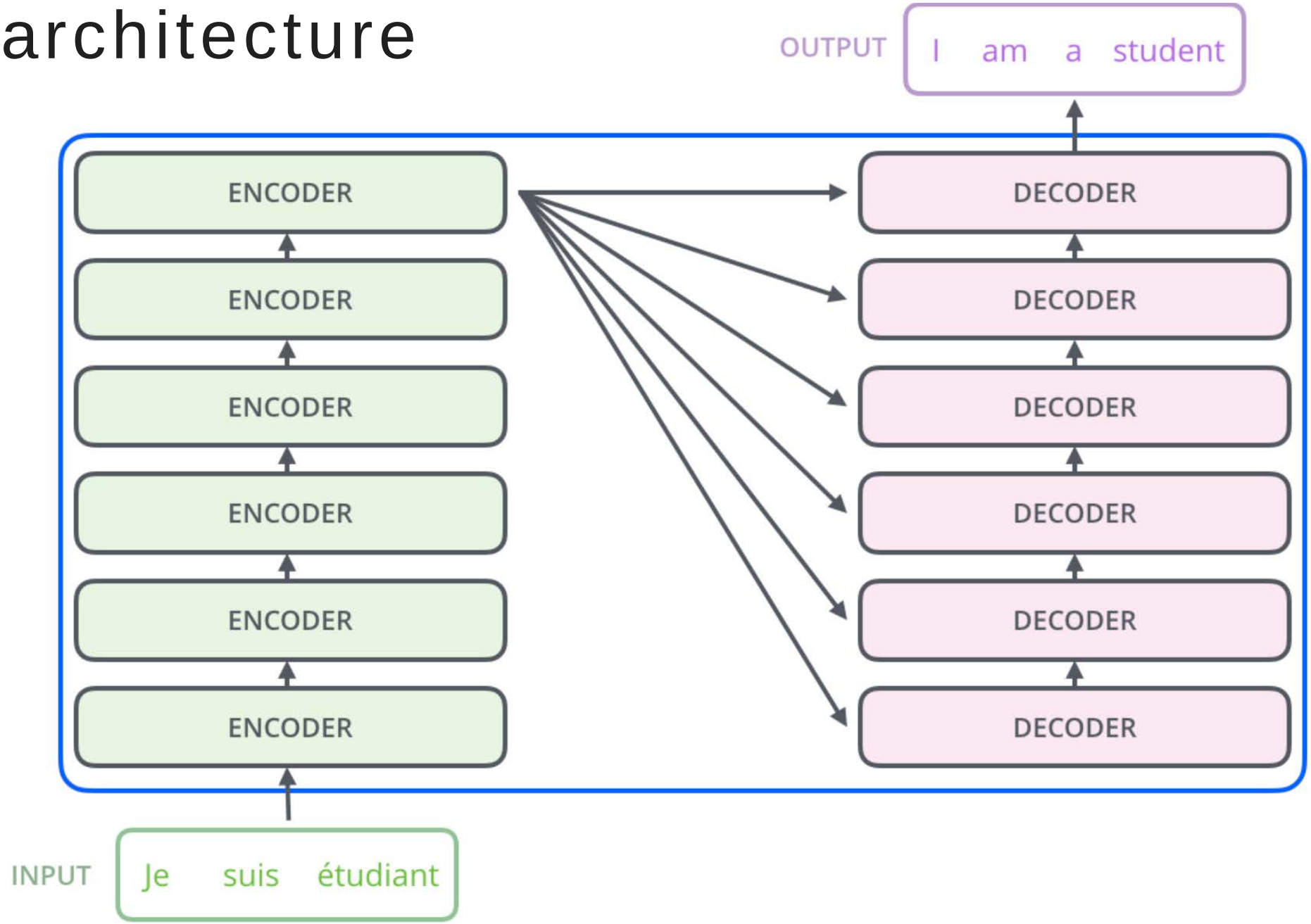


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

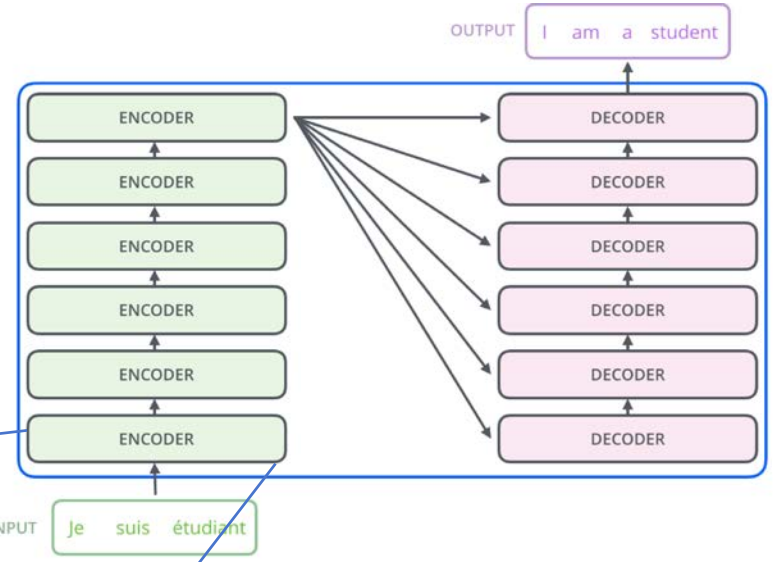
An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

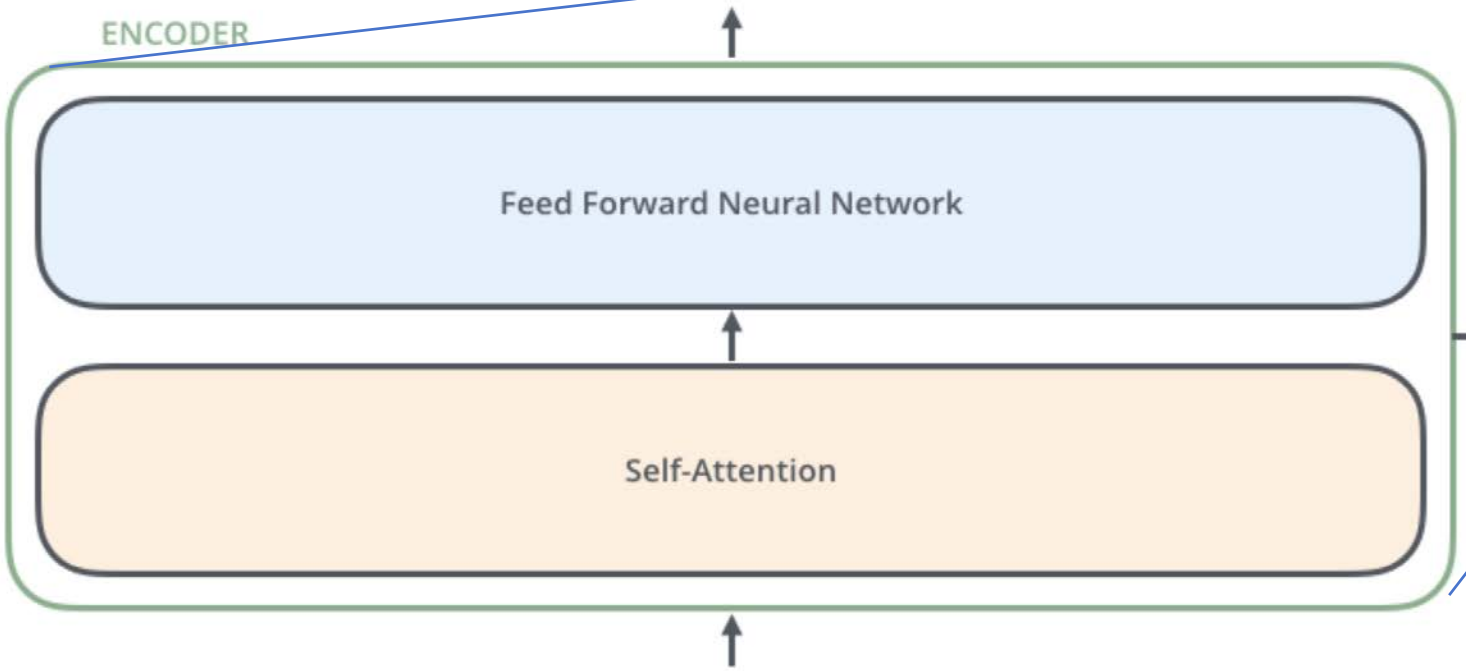
High-level architecture

- Will only look at the ENCODER(s) part in detail



The **encoders** are **all identical in structure** (yet they do not share weights). Each one is broken down into two sub-layers



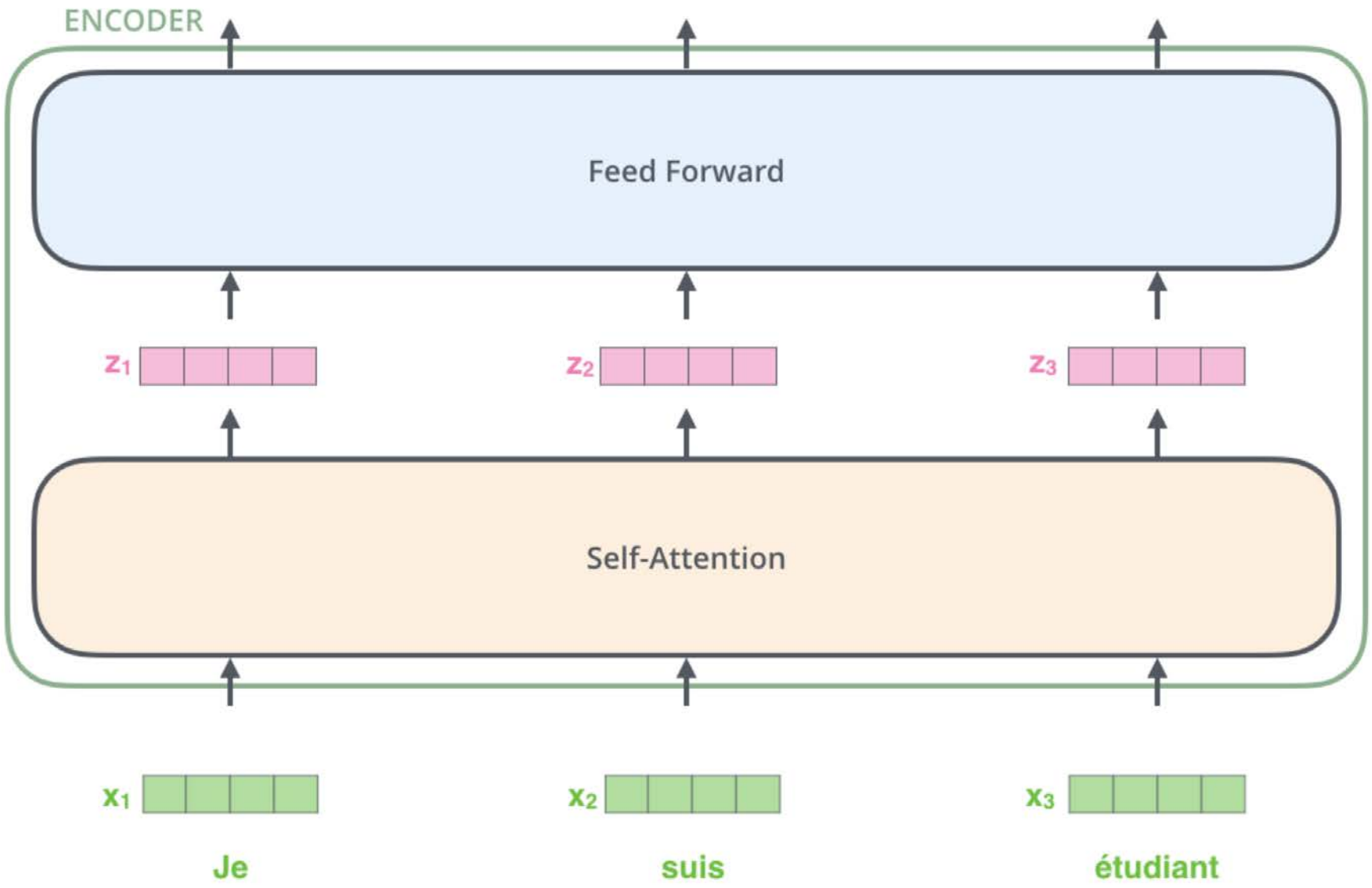


outputs of the self-attention are fed to a feed-forward neural network. The exact same one is independently applied to each position.

helps the encoder look at other words in the input sentence as it encodes a specific word.

Key property of Transformer: word in each position flows through its own path in the encoder.

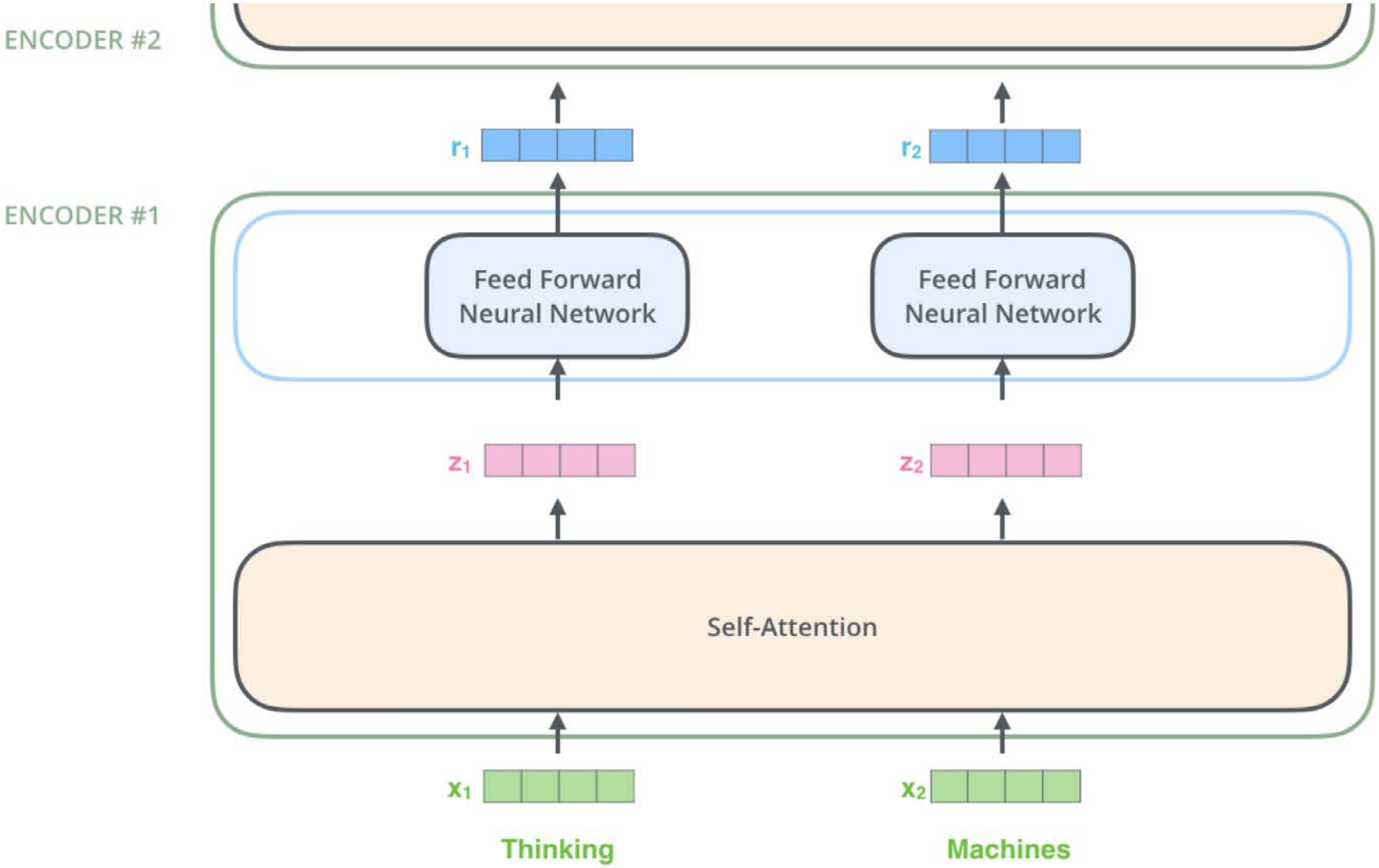
- There are dependencies between these paths in the self-attention layer.
- Feed-forward layer does not have those dependencies => various paths can be executed in parallel !



Word embeddings

Visually clearer on two words

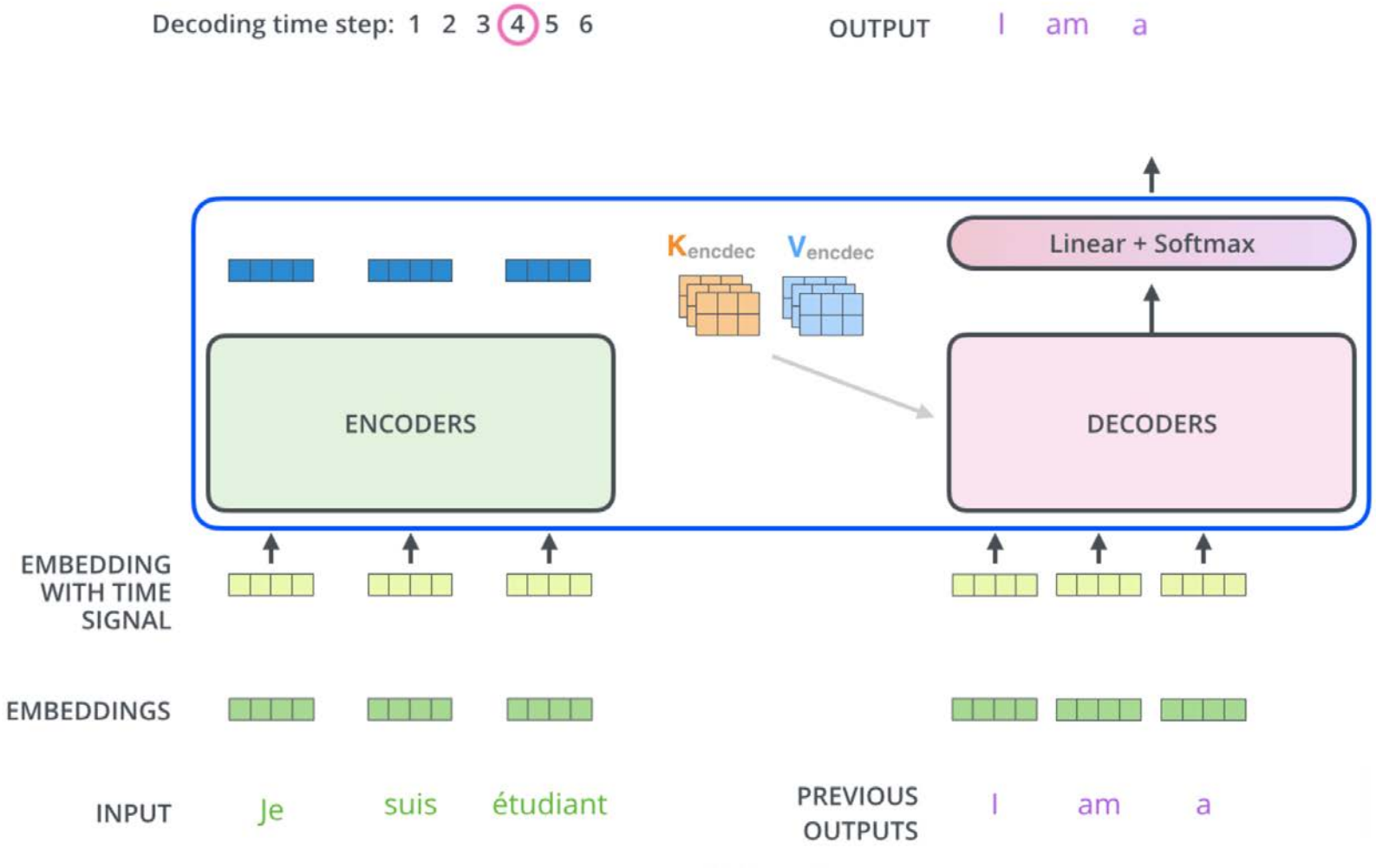
- dependencies in self-attention layer.
- No dependencies in Feed-forward layer



Word embeddings

The Decoder Side

- Relies on most of the concepts on the encoder side
- See animation on https://jalammar.github.io/illustrated-transformer/



Multi-head Attention

The Transformer uses multi-head attention in three different ways:

- In "encoder-decoder attention" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence. This mimics the typical encoder-decoder attention mechanisms in sequence-to-sequence models such as [38, 2, 9].
- The encoder contains self-attention layers. In a self-attention layer all of the **keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder**. Each position in the encoder can attend to all positions in the previous layer of the encoder.
- Similarly, self-attention layers in the decoder **allow each position in the decoder to attend to all positions in the decoder up to and including that position**. We need to prevent leftward information flow in the decoder to preserve the **auto-regressive** property. We implement this inside of scaled dot-product attention by masking out (setting to infinity) all values in the input of the softmax which correspond to illegal connections. See Figure 2.

At each step the model is auto-regressive [10], consuming the previously generated symbols as additional input when generating the next. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term);

3.3 Position-wise Feed-Forward Networks

In addition to attention sub-layers, each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2}$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{\text{model}} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$.

Positional Encodings

we add "positional encodings" to the input embeddings at the bottoms of the encoder and decoder stacks. The positional encodings have the same dimension d_{model} as the embeddings, so that the two can be summed. There are many choices of positional encodings, learned and fixed [9].

In this work, we use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

where pos is the position and i is the dimension. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. We chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}.

We also experimented with using learned positional embeddings [9] instead, and found that the two versions produced nearly identical results (see Table 3 row (E)). We chose the sinusoidal version because it may allow the model to extrapolate to sequence lengths longer than the ones encountered during training.

Key Innovations

- Self-Attention Mechanism
  - Dynamically weighs relationships between all words in a sequence, capturing long-range dependencies regardless of distance (e.g., linking pronouns to their referents).
  - Computed as Scaled Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

    where Q (Query), K (Key), and V (Value) are learned matrices.
- Multi-Head Attention
  - Parallel attention "heads" learn diverse contextual relationships, improving model expressiveness.
- Positional Encoding
  - Injects positional information into input embeddings (via sine/cosine functions) since Transformers lack innate sequence awareness.

Core Components

- Layer Normalization and Residual Connections: Stabilize training in deep networks.
- Feed-Forward Networks: Applied per position for non-linear feature transformation.
- Masking: Prevents the decoder from "peeking" at future tokens during autoregressive generation.

Advantages Over RNNs/CNNs

- Long-range dependency modeling (no vanishing gradients).
- Superior computational efficiency (parallelizable).
- Scalability: Foundation for models like GPT (decoder-only), BERT (encoder-only), and multimodal systems (e.g., Vision Transformers).

Dataset used in Project 3

- **THUCNews** is a Chinese news text classification dataset developed by Tsinghua University's NLP Lab (THUNLP). It contains approximately 740,000 news articles from Sina News, categorized into 10 classes (e.g., sports, finance, technology). Widely used as a benchmark for Chinese NLP tasks, it provides balanced, high-quality labeled data for training and evaluating text classification models (e.g., BERT, CNN). Its standardized structure makes it a popular choice for both research and practical applications.
- The task in this project is to classify sentences of news in Chinese into 10 categories namely 财经, 房产, 股票, 教育, 科技, 社会, 时政, 体育, 游戏, and 娱乐. Each of the categories is represented by an integer from 0 to 9 in the given order.

Training Task

- The transformer model will be trained, evaluated, and tested on a truncated version of the THUCNews dataset with 50000 sentences of news and their categories, where 90% are in the training set and 5% are in the evaluation set (also known as the validation or development set) and test set respectively. To ensure the validity of evaluation and test results, these three sets should not have any shared content.

Training Task

- To run the transformer model, simply run the run.py file in the directory. The program will automatically train the model for 6 epochs and evaluate its performance (loss and accuracy) on the evaluation set at the end of each epoch. The evaluation can provide a reference for when to terminate training. After the training terminates, the model will run on the test set for a final evaluation. There are also some classification cases at the end of the program to provide an intuitive perception of the model's performance. You should include a screenshot of the whole running outputs and the model's loss and total accuracy on the test set in your report.

Evaluation

- The model evaluation criteria for this project are the classification accuracy and confusion matrix of various types of data on the test set.

```
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

          财经     0.6858    0.8697    0.7669       261
          房产     0.8800    0.8907    0.8853       247
          股票     0.5180    0.8471    0.6429       255
          教育     0.9216    0.8969    0.9091       262
          科技     0.7966    0.5975    0.6828       236
          社会     0.7430    0.9231    0.8233       260
          时政     0.9128    0.6157    0.7354       255
          体育     0.9919    0.4822    0.6489       253
          游戏     0.8196    0.8008    0.8101       261
          娱乐     0.7962    0.7590    0.7772       278

    accuracy                         0.7702      2568
   macro avg     0.8065    0.7683    0.7682      2568
weighted avg     0.8061    0.7702    0.7692      2568

Confusion Matrix...
[[227   1  26   1   2   2   0   0   0   2]
 [  9 220  12   1   1   1   0   0   3   0]
 [ 29   3 216   1   2   1   2   0   0   1]
 [  6   2   4 235   3   3   3   0   2   4]
 [ 18   2  33   7 141   9   3   0  23   0]
 [  0   4   7   5   1 240   3   0   0   0]
 [ 11   4  43   1   4  26 157   0   2   7]
 [ 18   5  37   2   5  21   3 122   2  38]
 [  5   2  22   1  15   5   0   0 209   2]
 [  8   7  17   1   3  15   1   1  14 211]]
Time usage: 0:00:00
text:10基金赎益平安深发展重组 深证100指数被遗忘     label:财经
text:高考要签诚信承诺书 认定考生作弊范围扩大     label:教育
text:基金募资额创5年次新低    label:财经
text:考研路上相逢勇者胜 英语词汇是关键     label:教育
text:通州月亮湾皖镇小产权现房均价4300元(图)    label:房产
```

What you need to do?

- Using various fine-tuning techniques to improve the accuracy of your model
- Method
  - Data Preprocessing(**Required**)
    - Task A: use original data without preprocessing, do training and testing get accuracy.
    - Task B: do text preprocessing(remove puctuation etc.), then do the training and testing.
    - Compare the test accuracy in task a and task b.
    - output taskA.misclassified(all sentences that been misclassfied)
    - output taskB.misclassified
    - finding 3 sentences in taskA.misclassified but not in taskB.misclassifieied, Expalin why?
    - finding 3 sentences in taskB.misclassified but not in taskA.misclassifieied, Expalin why?
  - Architecture Adjustments(**Optional**)
  - For example:
    - Classifier Head: Replace the default head with a Dropout + Linear layer for 10-class output.
    - Pooling: Use [CLS] token output or mean/max pooling of token embeddings.
  - (**Optional**)Training Strategies(for example): linear decay with warmup,Early Stopping
  - (**Optional**)Advanced Tricks(for example):Layer-wise LR Decay, Focal Loss, Data Augmentation

中华女子学院：本科层次仅1专业招男生　3
两天价网站背后重重迷雾：做个网站究竟要多少钱　4
东5环海棠公社230-290平2居准现房98折优惠　1
卡佩罗：告诉你德国脚生猛的原因 不希望英德战踢点球　7
82岁老太为学生做饭扫地44年获授港大荣誉院士　5
记者回访地震中可乐男孩：将受邀赴美国参观　5
冯德伦徐若 隔空传情 默认其是女友　9
传郭晶晶欲落户香港战伦敦奥运 装修别墅当婚房1
《赤壁OL》攻城战诸侯战硝烟又起　8
“手机钱包”亮相科博会　4
上海2010上半年四六级考试报名4月8日前完成　3
李永波称李宗伟难阻林丹取胜 透露谢杏芳有望出战　7
3岁女童下体红肿 自称被幼儿园老师用尺子捅伤　5
金证顾问：过山车行情意味着什么　2
谁料地王如此虚　1
《光环5》Logo泄露 Kinect版几无悬念　8
海淀区领秀新硅谷宽景大宅预计10月底开盘　1
柴志坤：土地供应量不断从紧 地价难现07水平(图)　1
伊达传说EDDA Online　8
三联书店建起书香巷　4
宇航员尿液堵塞国际空间站水循环系统4
研究发现开车技术差或与基因相关　6
皇马输球替补席闹丑闻 队副女球迷公然调情(视频)　7
北京建工与市政府再度合作推出郭庄子限价房　1
组图：李欣汝素颜出镜拍低碳环保大片9
2008中文网志年会演讲人：庄秀丽　4
3000点之下是买入好时机　2
赵本山要追究法律责任：居然诽谤我打小沈阳　9
总有一款适合你 潮人必备多用途电视导购　4
万科退赛 恒大16.6亿抢下深圳建设集团　2
俄企业赴港上市热情不减 资源类企业积极性最高2
《非诚勿扰》单亲妈妈邂逅沧桑型男 推出微直播9
朱之文亮的都是绝活《GIGA SLAVE》 8
《非诚勿扰》“冯女郎”车晓带妈妈闯世界(图)　9
美弗吉尼亚大学访华太设计签实习基地协议（组图）　1
华中科技大学2010年考研成绩查询开通　3
陈小艺“激吻照”疑似炒作　9
90岁老太半世纪撮合200多对新人(图)　5

中华女子学院本科层次仅1专业招男生 3
两天价网站背后重重迷雾做个网站究竟要多少钱 4
东5环海棠公社230290平2居准现房98折优惠 1
卡佩罗告诉你德国脚生猛的原因不希望英德战踢点球 7
82岁老太为学生做饭扫地44年获授港大荣誉院士 5
记者回访地震中可乐男孩将受邀赴美国参观 5
冯德伦徐若隔空传情默认其是女友 9
传郭晶晶欲落户香港战伦敦奥运装修别墅当婚房 1
赤壁OL攻城战诸侯战硝烟又起 8
手机钱包亮相科博会 4
上海2010上半年四六级考试报名4月8日前完成 3
李永波称李宗伟难阻林丹取胜透露谢杏芳有望出战 7
3岁女童下体红肿自称被幼儿园老师用尺子捅伤 5
金证顾问过山车行情意味着什么 2
谁料地王如此虚 1
光环5Logo泄露Kinect版几无悬念 8
海淀区领秀新硅谷宽景大宅预计10月底开盘 1
柴志坤土地供应量不断从紧地价难现07水平图 1
伊达传说EDDAOnline 8
三联书店建起书香巷 4
宇航员尿液堵塞国际空间站水循环系统 4
研究发现开车技术差或与基因相关 6
皇马输球替补席闹丑闻队副女球迷公然调情视频 7
北京建工与市政府再度合作推出郭庄子限价房 1
组图李欣汝素颜出镜拍低碳环保大片 9
2008中文网志年会演讲人庄秀丽 4
3000点之下是买入好时机 2
赵本山要追究法律责任居然诽谤我打小沈阳 9
总有一款适合你潮人必备多用途电视导购 4
万科退赛恒大166亿抢下深圳建设集团 2
俄企业赴港上市热情不减资源类企业积极性最高 2
非诚勿扰单亲妈妈邂逅沧桑型男推出微直播 9
朱之文亮的都是绝活GIGASLAVE 8
非诚勿扰冯女郎车晓带妈妈闯世界图 9
美弗吉尼亚大学访华太设计签实习基地协议组图 1
华中科技大学2010年考研成绩查询开通 3
陈小艺激吻照疑似炒作 9
90岁老太半世纪撮合200多对新人图 5

1. **阔别 9 年：大一新生回母校看小学老师**
    - **Reason for misclassification in Task A:** This sentence contains a combination of phrases like "阔别 9 年" (being apart for 9 years) and "回母校看小学老师" (going back to the alma mater to visit a primary school teacher), which are quite unique. The model might misinterpret the underlying meaning due to the uncommon combination of terms.
    - **Reason for correct classification in Task B:** In Task B, preprocessing (like punctuation removal) might clean up punctuation marks like "：" (colon), making the model focus more on the core meaning of the sentence, improving its understanding of the context, which helps correctly classify it.

2. **姜建清：工行五年内境外利润占比实现 10%**

   - **Reason for misclassification in Task A:** This sentence involves some numbers and specific terms like "姜建清" (a person's name) and "工行五年内境外利润占比" (ICBC's foreign profit ratio in 5 years). The model might struggle with handling named entities (like a person's name) or specific percentages without appropriate contextual understanding, causing a misclassification.

   - **Reason for correct classification in Task B:** Preprocessing, especially removal of punctuation like the colon ": ", could help the model better handle the content by focusing on the key terms like "工行" (ICBC), and possibly understanding the context of the statement without being distracted by less relevant punctuation.

2. **香港政府呼吁菲律宾当局保证人质安全**

    - **Reason for misclassification in Task B:** The sentence talks about a sensitive situation involving a hostage and a government request. After preprocessing, the removal of punctuation could blur the intent or weaken the model's ability to correctly classify the sentence. The model may fail to capture the urgency or the political context without the punctuation marks separating key phrases.

    - **Reason for correct classification in Task A:** In Task A, the sentence remains in its raw form, with the punctuation aiding the model in parsing the meaning and context. The punctuation marks could highlight the urgency or political nature of the statement, leading to correct classification.

3-2 A fail B success

```
pos = np.where(misclassified_pos_0 * (~misclassified_pos_1))[0]

np.random.seed(123)
pos = np.random.choice(pos, 3)

for seq, label in zip(texts_all_0[pos].detach().cpu().numpy(), labels_all[pos]):
    print(''.join([rev_vocab_0[ss] for ss in seq]).ljust(50), label)
```

```
漫画书为伍兹立传 高尔夫巨星与政经界名人比肩                                  7
激萌国宝魅力 星辰变熊猫公仔装来袭                                        8
肖鹰：郭敬明活着就是为了被粉丝消费                                        9
```

`Analysis` : I think it is hard to explain why a model precisely succeeds on 3 particular samples and another model does not. Perhaps one model is better for recognizing certain classes, perhaps it is just coincidencial due to the intrinsic model/data limit.

3-3 B fail A success

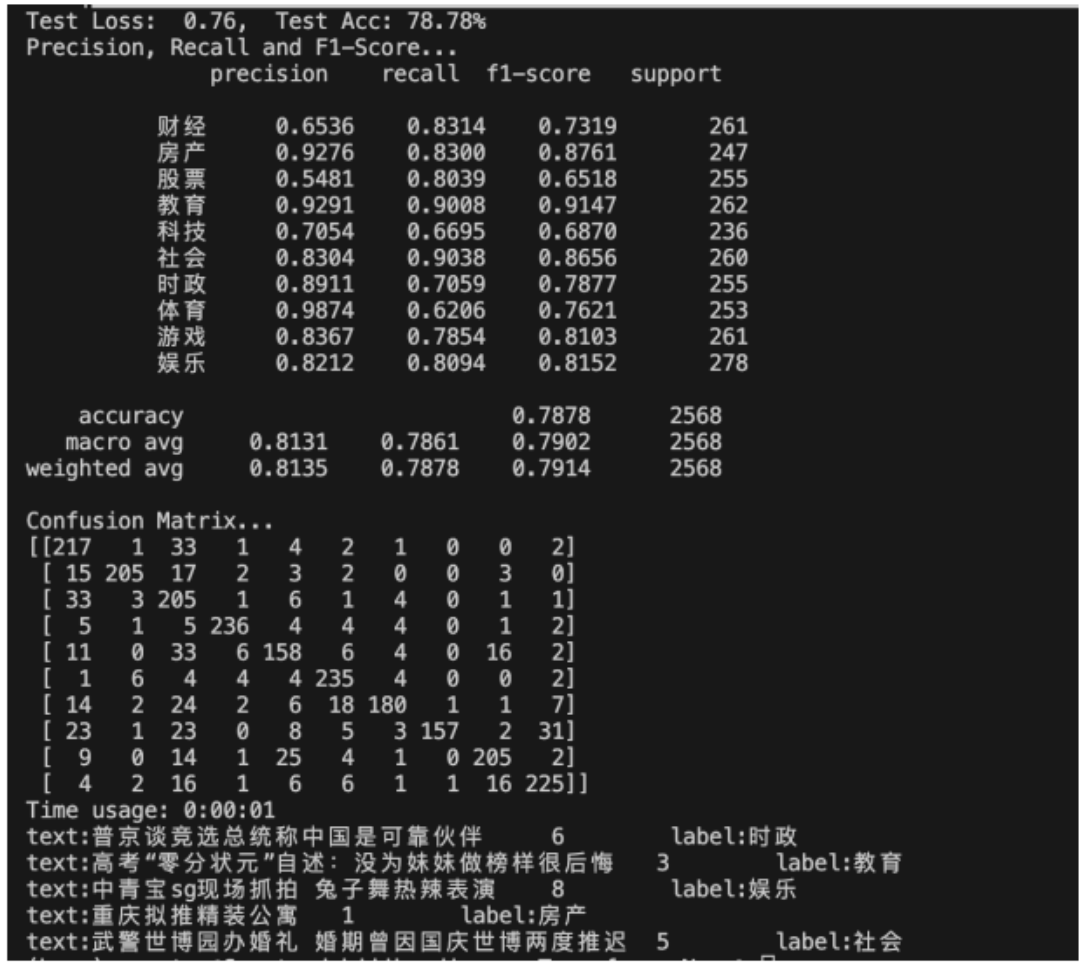
```
pos = np.where(misclassified_pos_1 * (~misclassified_pos_0))[0]

np.random.seed(121)
pos = np.random.choice(pos, 3)


for seq, label in zip(texts_all_0[pos].detach().cpu().numpy(), labels_all[pos]):
    print(''.join([rev_vocab_0[ss] for ss in seq]).ljust(50), label)
```

```
央视称张怡宁为自己划完美句号 复出可能几乎为零                                  7
少年发明家高考266分被破格录取(图)                                       5
中国石化和中国石油分别上涨1.20%和1.24%                                  2
```

`Analysis` : Same as above.



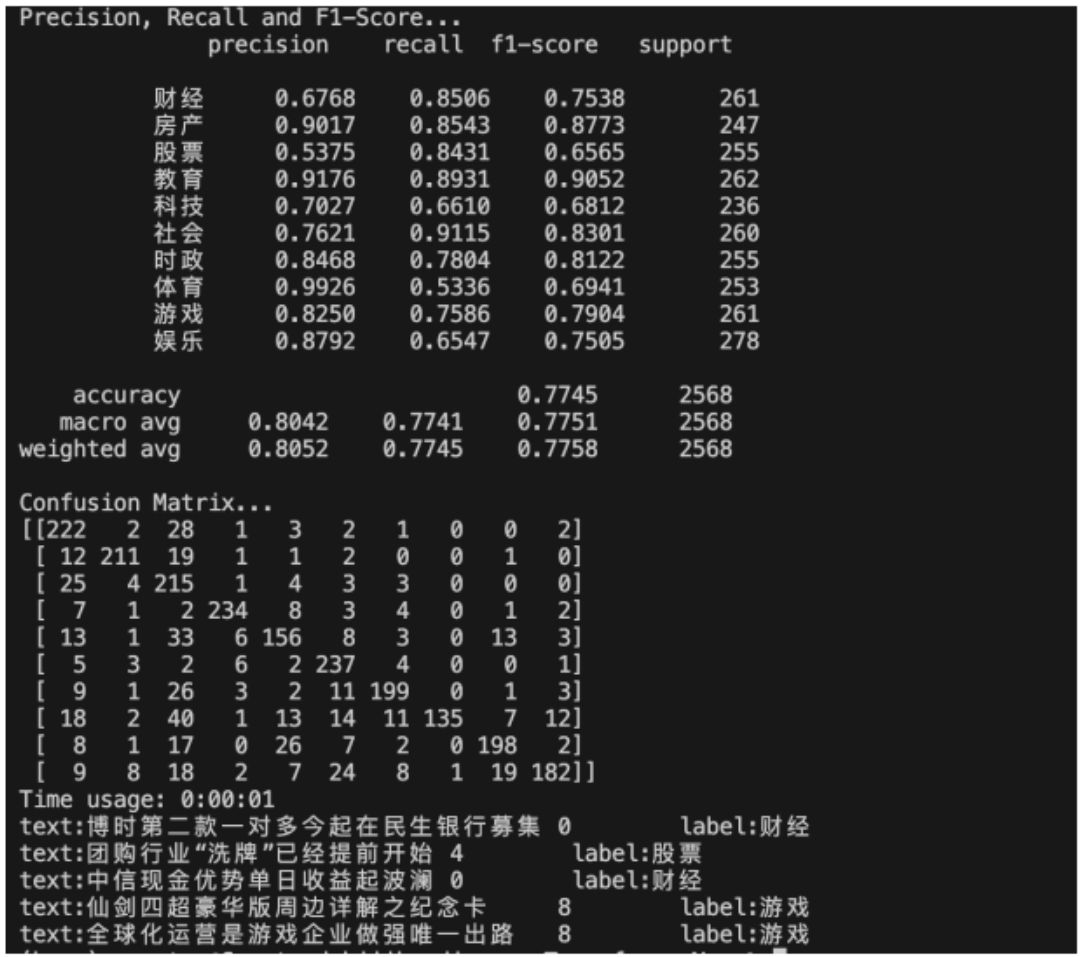
```
Test Loss:  0.76,  Test Acc: 78.78%
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

          财经     0.6536    0.8314    0.7319       261
          房产     0.9276    0.8300    0.8761       247
          股票     0.5481    0.8039    0.6518       255
          教育     0.9291    0.9008    0.9147       262
          科技     0.7054    0.6695    0.6870       236
          社会     0.8304    0.9038    0.8656       260
          时政     0.8911    0.7059    0.7877       255
          体育     0.9874    0.6206    0.7621       253
          游戏     0.8367    0.7854    0.8103       261
          娱乐     0.8212    0.8094    0.8152       278

    accuracy                         0.7878      2568
   macro avg     0.8131    0.7861    0.7902      2568
weighted avg     0.8135    0.7878    0.7914      2568

Confusion Matrix...
[[217   1  33   1   4   2   1   0   0   2]
 [ 15 205  17   2   3   2   0   0   3   0]
 [ 33   3 205   1   6   1   4   0   1   1]
 [  5   1   5 236   4   4   4   0   1   2]
 [ 11   0  33   6 158   6   4   0  16   2]
 [  1   6   4   4   4 235   4   0   0   2]
 [ 14   2  24   2   6  18 180   1   1   7]
 [ 23   1  23   0   8   5   3 157   2  31]
 [  9   0  14   1  25   4   1   0 205   2]
 [  4   2  16   1   6   6   1   1  16 225]]
Time usage: 0:00:01
text:普京谈竞选总统称中国是可靠伙伴      6        label:时政
text:高考“零分状元”自述：没为妹妹做榜样很后悔    3        label:教育
text:中青宝sg现场抓拍 兔子舞热辣表演     8        label:娱乐
text:重庆拟推精装公寓     1        label:房产
text:武警世博园办婚礼 婚期曾因国庆世博两度推迟   5        label:社会
```

(a) Task A



```
Precision, Recall and F1-Score...
              precision    recall  f1-score   support

          财经     0.6768    0.8506    0.7538       261
          房产     0.9017    0.8543    0.8773       247
          股票     0.5375    0.8431    0.6565       255
          教育     0.9176    0.8931    0.9052       262
          科技     0.7027    0.6610    0.6812       236
          社会     0.7621    0.9115    0.8301       260
          时政     0.8468    0.7804    0.8122       255
          体育     0.9926    0.5336    0.6941       253
          游戏     0.8250    0.7586    0.7904       261
          娱乐     0.8792    0.6547    0.7505       278

    accuracy                         0.7745      2568
   macro avg     0.8042    0.7741    0.7751      2568
weighted avg     0.8052    0.7745    0.7758      2568

Confusion Matrix...
[[222   2  28   1   3   2   1   0   0   2]
 [ 12 211  19   1   1   2   0   0   1   0]
 [ 25   4 215   1   4   3   3   0   0   0]
 [  7   1   2 234   8   3   4   0   1   2]
 [ 13   1  33   6 156   8   3   0  13   3]
 [  5   3   2   6   2 237   4   0   0   1]
 [  9   1  26   3   2  11 199   0   1   3]
 [ 18   2  40   1  13  14  11 135   7  12]
 [  8   1  17   0  26   7   2   0 198   2]
 [  9   8  18   2   7  24   8   1  19 182]]
Time usage: 0:00:01
text:博时第二款一对多今起在民生银行募集 0        label:财经
text:团购行业“洗牌”已经提前开始 4        label:股票
text:中信现金优势单日收益起波澜 0        label:财经
text:仙剑四超豪华版周边详解之纪念卡     8        label:游戏
text:全球化运营是游戏企业做强唯一出路    8        label:游戏
```

(b) Task B

(a) Task A

(b) Task B

Figure 2: Training Process

In the task details, we suggested students could do the final additional /optional tasks:

- Architecture Adjustments(**Optional**)
- For example:
  - Classifier Head: Replace the default head with a Dropout + Linear layer for 10-class output.
  - Pooling: Use [CLS] token output or mean/max pooling of token embeddings.
- (**Optional**)Training Strategies(for example): linear decay with warmup,Early Stopping
- (**Optional**)Advanced Tricks(for example):Layer-wise LR Decay, Focal Loss, Data Augmentation

One student did the following additional tasks

2.2 Architecture Adjustments

The second experiment modifies the classification block at its tail.

- **Head.** Replace the large "flatteneverything + Linear" head with `Dropout(0.5) -> Linear(dim_model, 10)`.
- **Pooling.** Feed the head with the encoder's first-token (`[CLS]`) embedding: `pooled = x[:,0,:]`.