بسم الله الرحمن الرحیم

نام دانشجو: سید محمد علی رضایی

شماره دانشجویی: 400131020

استاد درس: دکتر صفابخش

گزارش تمرین شماره ۸

کدهای گزارش در فایل زیپ موجود می‌باشد.

1- نمایش داده‌های موجود در مجموعه دادگان:

```
e quando melhoramos a procura , tiramos a única vantagem da impressão , que é a serendipidade .
mas e se estes fatores fossem ativos ?
mas eles não tinham a curiosidade de me testar .
e esta rebeldia consciente é a razão pela qual eu , como agnóstica , posso ainda ter fé .
`` `` '' podem usar tudo sobre a mesa no meu corpo . ''

and when you improve searchability , you actually take away the one advantage of print , which is serendipity .
but what if it were active ?
but they did n't test for curiosity .
and this conscious defiance is why i , as an agnostic , can still have faith .
you can use everything on the table on me .
```

تصویر ۱

پنج خط اول تصویر ۱ شامل جملات موجود در دیتا ست و پنج خط بعدی شامل ترجمه این جملات از زبان پرتغالی به انگلیسی می‌باشد. از آنجایی که نمی‌توان مدل را مستقیما روی متن اموزش داد، به همین خاطر متن را به یک نمایش عددی تبدیل می‌کنیم که در اینجا متن را به دنباله‌هایی از شناسه‌های نشانه‌دار تبدیل می‌کنیم. که این اعداد به عنوان شاخص در Embedding استفاده می‌شوند. برای توکن کردن جملات موجود در این دیتا ست مطابق با اموزش موجود در تنسورفلو با توجه به بلاک کد زیر از یک مدل بهینه برای این کار استفاده شده است. اندازه vocab size = 2^13=8192، در نظر می‌گیریم در این جا از صفر تا ۸۱۹۲ آی دی های مربوطه به هر کلمه موجود در دیکشنری می‌باشد. و در صورت نبود این کلمه در مجموعه دیکشنری مربوطه، ان را به کلماتی که در مجموعه دیکشنری باشند می‌شکند.

```
1 tokenizer_en = tfds.deprecated.text.SubwordTextEncoder.build_from_corpus((en.numpy() for pt, en in train_examples), target_vocab_size=2**13)
2 tokenizer_pt = tfds.deprecated.text.SubwordTextEncoder.build_from_corpus((pt.numpy() for pt, en in train_examples), target_vocab_size=2**13)
```

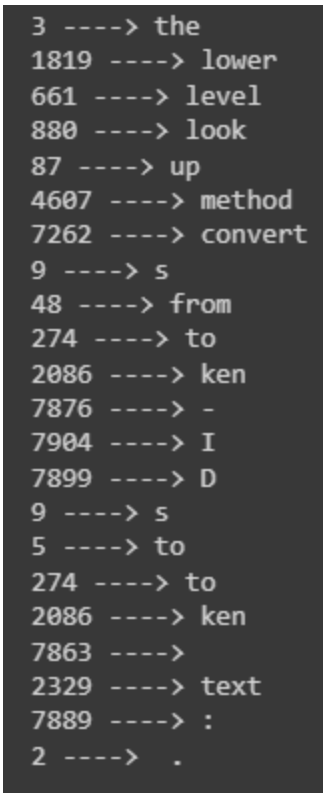
تصویر۲

تصویر۲ بلاک کد لازم برای توکن بندی و اختصاص دادن آی دی مناسب به هر کلمه برای زبان انگلیسی و پرتغالی می‌باشد.

به عنوان مثال رشته ورودی به زبان انگلیسی به صورت زیر:

“The lower level lookup method converts from token-IDs to token text:”

به صورت تصویر ۳ خواهد شد:



تصویر۳

از دیگر پیش پردازش‌های صورت گرفته اضافه کردن کلمه start و end به ابتدا و انتهای هر جمله با استفاده از تابع encode می‌باشد. همچنین برای جلوگیری از بار محاسباتی زیاد، جملاتی که طول ان ها از ۸۰ توکن بیشتر می‌باشد با استفاده از تابع فیلتر از مجموعه اموزش حذف کرده‌ایم. سپس با استفاده از تابع cache مجموعه دادگان را برای افزایش در سرعت محاسبات به داخل رم انتقال می‌دهیم و با استفاده از تابع shuffle مجموعه دادگان را بر می‌زنیم.

در نهایت مجموعه دادگان مورد نظر به صورت زیر می‌باشد:

از آنجایی که batch size را برابر با ۶۴ و ماکزیمم طول جملات برای توکن بندی برابر با ۸۰ در نظر گرفته شده است، داده‌ها به صورت دسته‌های ۶۴ تایی با یکدیگر تا نهایت طول ۸۰ کنار یکدیگر به صورت تصویر۴ قرار گرفته‌اند:

```
<tf.Tensor: shape=(64, 54), dtype=int64, numpy=
array([[8214, 6744,   12, ...,    0,    0,    0],
       [8214,  368,    1, ...,    0,    0,    0],
       [8214,  119,    1, ...,    0,    0,    0],
       ...,
       [8214,  533,  106, ...,    0,    0,    0],
       [8214,  123,    7, ...,    0,    0,    0],
       [8214,    3,    4, ...,    0,    0,    0]])>,
<tf.Tensor: shape=(64, 70), dtype=int64, numpy=
array([[8214, 2398,    1, ...,    0,    0,    0],
       [8214,   77,  142, ...,    0,    0,    0],
       [8214,   11,   79, ...,    0,    0,    0],
       ...,
       [8214,    8,   23, ...,    0,    0,    0],
       [8214,   25,   47, ...,    0,    0,    0],
       [8214,    6, 2433, ...,    0,    0,    0]])>,
```

تصویر۴

۲- محاسبه تعبیه مکانی:

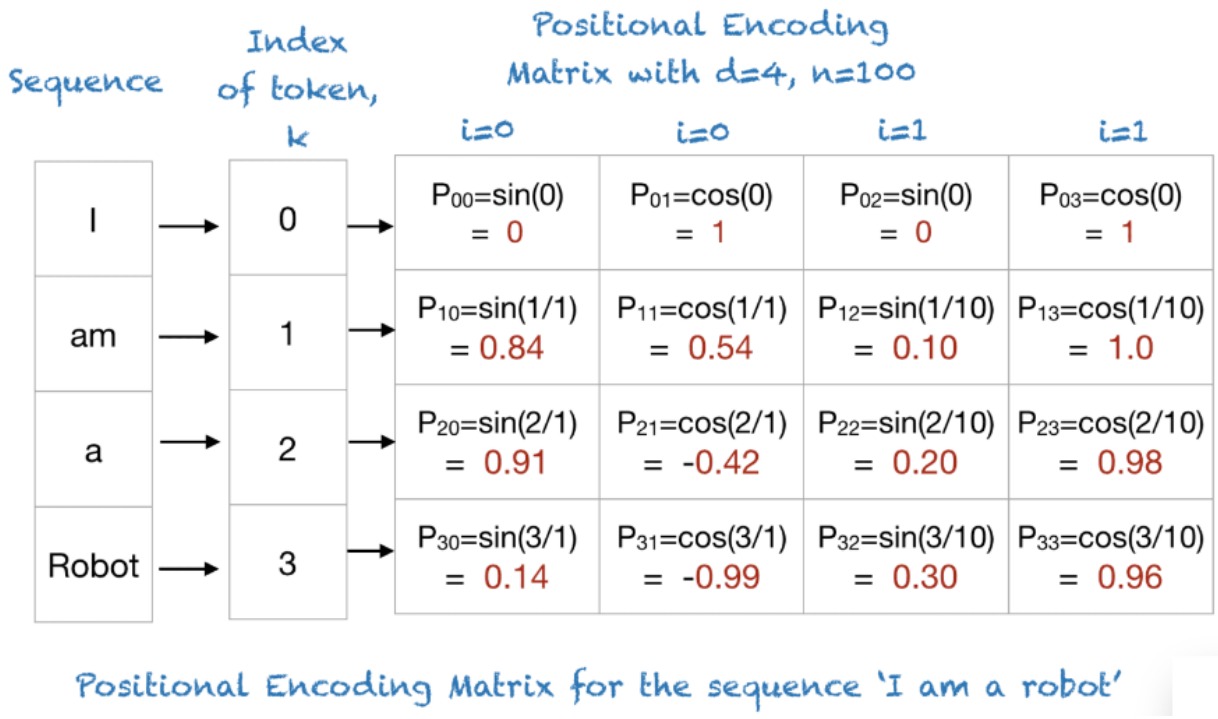
در اموزش شبکه ترنسفورمر پس از محاسبه embedding کلمات ورودی، اطلاعات مکانی کلمات ورودی را با embedding هایی که در مرحله قبل بدست اورده‌ایم اضافه می‌کنیم. با استفاده از فرمول زیر برای پوزیشن‌های زوج از رابطه sin و پوزیشن‌های فرد از رابطه cos استفاده می‌کنیم. عددی که این روابط بدست می‌اورند با بردار embedding متناظر با ان کلمه جمع می‌شوند، با این کار در هنگام اموزش، شبکه متوجه می‌شود این اطلاعات مکانی نیز اضافه شده است.

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

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

با توجه به رابطه بالا ابتدا به محاسبه مقدار درجه موجود در رابطه می‌پردازیم. در این رابطه pos برابر موقعیت کلمه در جمله مورد نظر می‌باشد. d_{model} ابعاد مسئله که در ان بردارهای embedding قرار دارند می‌باشد. i نیز بر حسب ابعاد مسئله می‌باشد. همان طور که هر کلمه به صورت یک بردار پس از embedding می‌باشد برای لحاظ کردن این موقعیت به کلمات از رابطه بالا استفاده می‌کنیم به گونه‌ای که برای مولفه‌های زوج بردار ار تابع سینوس و مولفه‌های فرد، از تابع کسینوس استفاده می‌کنیم.

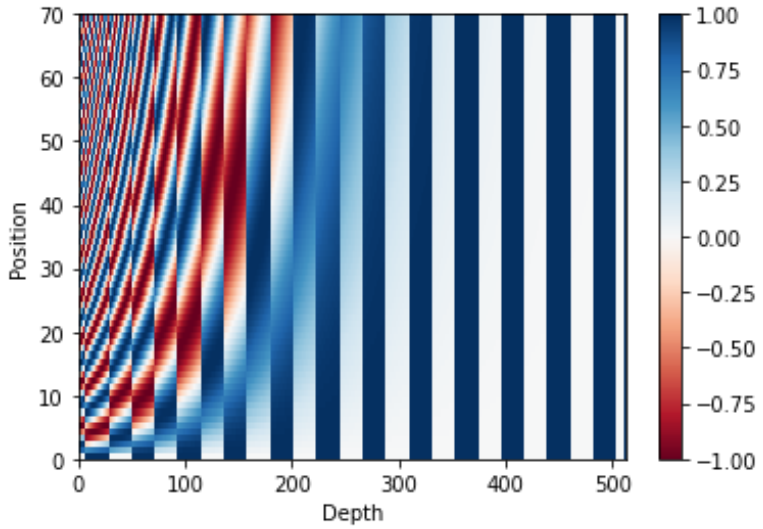
تصویر زیر نحوه محاسبه بردار مطلوب در positional encoding برای هر کلمه را نمایش می‌دهد.



Positional Encoding Matrix for the sequence 'I am a robot'

همان طور که مشاهده می‌شود برای مثال کلمه اول یعنی " I " در موقعیت اول جمله قرار گرفته است پس مقدار pos ان برابر با صفر می‌باشد سپس با توجه به اینکه d_model برابر با ۴ می‌باشد پس این بردار نیز دارای طول ۴ می‌باشد که برای محاسبه مولفه‌های فرد ان از تابع sin و مولفه‌های زوج ان از تابع cos استفاده می‌شود. برای کلمه اول بردار بدست امده طبق رابطه‌ها برابر با [0,1,0,1] می‌باشد که این بردار با بردار بدست امده از مرحله embedding برای کلمه اول جمع خواهد شد. برای کلمات دیگر نیز به همین منوال محاسبه خواهد شد.

تصویر زیر نمایشی از بردار های بدست امده برای تعبیه مکانی صورت گرفته به اندازه بردار ۵۱۲ و اندازه توکن ۷۰ می‌باشد. همان طور که مشخص است، بردارهای اختصاص داده شده به هر توکن به صورت منحصر به فرد می‌باشد و علت اینکه هر بردار در انتها دارای شباهت با بردارهای دیگر است استفاده از padding می‌باشد.



۳-شبکه ترنسفورمر :

برای اموزش این شبکه پارامترهایی که در مقاله اصلی پیش‌نهاد شده است عبارت اند از :

تعداد لایه ها =۶ ، ابعاد مدل=۵۱۲ ، تعداد نرون‌های لایه تمام متصل=۲۰۴۸ ، تعداد مجموعه ماتریس‌های قابل یادگیری برای محاسبه key,value,query=۸ می‌باشد. اما برای سرعت بخشیدن و کوچک‌تر کردن این تمرین در اموزش تنسورفلو پیش‌نهاد شده است که این پارامترها را تغییر داده و برابر با تعداد لایه ها =۶۴، ابعاد مدل=۱۲۸ ، تعداد نرون‌های لایه تمام متصل=۵۱۲ ، تعداد مجموعه ماتریس‌های قابل یادگیری برای محاسبه key,value,query=۸ در نظر بگیریم. همچنین اموزش مدل را با epoch =100 بر روی کل مجموعه داده train اموزش داده‌ایم. شکل زیر مقدار صحت و loss بدست امده در تلاش ۱۰۰ ام را نشان می‌دهد.

```
Epoch 100 Batch 0 Loss 0.5248 Accuracy 0.2612
Epoch 100 Batch 50 Loss 0.5032 Accuracy 0.2874
Epoch 100 Batch 100 Loss 0.5028 Accuracy 0.2885
Epoch 100 Batch 150 Loss 0.5013 Accuracy 0.2891
Epoch 100 Batch 200 Loss 0.5043 Accuracy 0.2869
Epoch 100 Batch 250 Loss 0.5077 Accuracy 0.2866
Epoch 100 Batch 300 Loss 0.5121 Accuracy 0.2858
Epoch 100 Batch 350 Loss 0.5153 Accuracy 0.2851
Epoch 100 Batch 400 Loss 0.5174 Accuracy 0.2852
Epoch 100 Batch 450 Loss 0.5206 Accuracy 0.2852
Epoch 100 Batch 500 Loss 0.5222 Accuracy 0.2851
Epoch 100 Batch 550 Loss 0.5248 Accuracy 0.2850
Epoch 100 Batch 600 Loss 0.5273 Accuracy 0.2845
Epoch 100 Batch 650 Loss 0.5298 Accuracy 0.2844
Epoch 100 Batch 700 Loss 0.5318 Accuracy 0.2843
Epoch 100 Batch 750 Loss 0.5344 Accuracy 0.2841
Saving checkpoint for epoch 100 at /content/drive/MyDrive/mamad/ckpt-24
Epoch 100 Loss 0.5365 Accuracy 0.2835
Time taken for 1 epoch: 73.44700527191162 secs
```

قسمت اول:

بهینه ساز مناسب برای اموزش این مدل بر اساس بهینه ساز معرفی شده در مقاله attention is all you need(2017)، از بهینه ساز آدام استفاده شده است با این تفاوت که نرخ یادگیری با استفاده از رابطه موجود در تصویر ۵ در طول اموزش تغییر خواهد کرد. بر اساس اطلاعات این مقاله پارامترهای بهینه ساز ادام از قبیل beta1 که معرف نرخ فروپاشی نمایی برای تخمین‌های گشتاور اول است برابر با 0.9 و beta2 که معرف نرخ فروپاشی نمایی برای تخمین‌های لحظه دوم است برابر با 0.98 و همچنین مقدار اپسیلون که برای جلوگیری از تقسیم بر صفر در پیاده‌سازی‌ها استفاده می‌شود برابر با 10^{-9} در نظر گرفته شده‌اند.

$$lrate = d_{\text{model}}^{-0.5} \cdot \min(step\_num^{-0.5}, step\_num \cdot warmup\_steps^{-1.5})$$

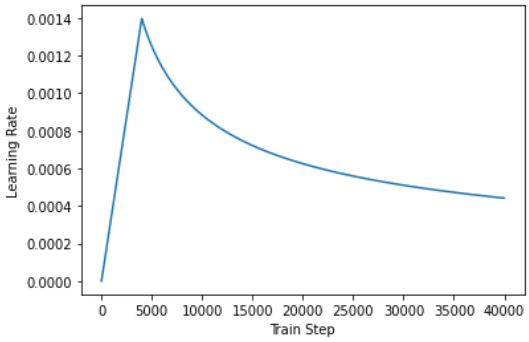
Figure ۵

در این رابطه step_num برابر با شماره epoch مربوطه در طول اموزش و step_num.warmup_step یک هایپر پارامتر که در مقاله مقدار پیش‌نهادی برای ان را ۴۰۰۰ معرفی می‌کنند. که با توجه به رفتار تابع در ابتدا مقدار نرخ یادگیری به صورت خطی افزایش پیدا می‌کند و سپس نسبت به جذر معکوس عدد گام کاهش پیدا خواهد کرد.

تصویر زیر نحوه پیاده سازی کلاس مربوط به بهینه ساز را نمایش می‌دهد.

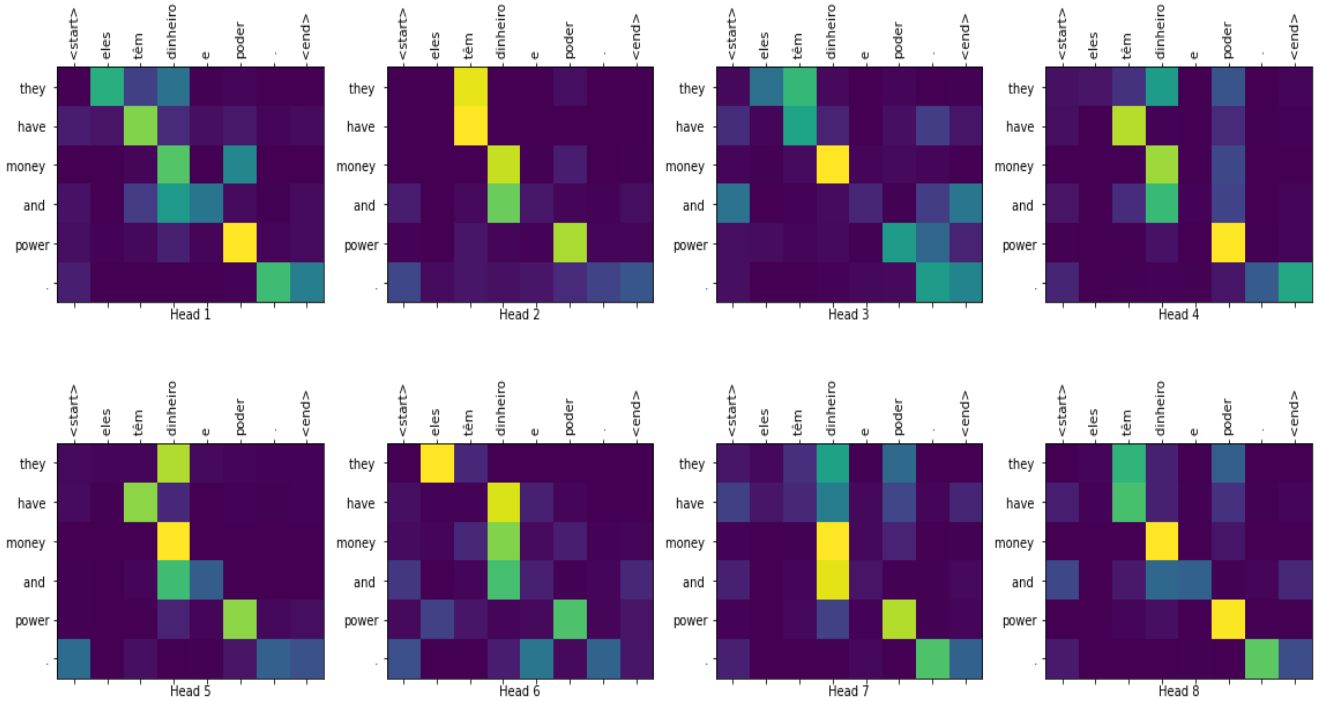
```
class CustomSchedule(tf.keras.optimizers.schedules.LearningRateSchedule):
  def __init__(self, d_model, warmup_steps=4000):
    super(CustomSchedule, self).__init__()

    self.d_model = d_model
    self.d_model = tf.cast(self.d_model, tf.float32)

    self.warmup_steps = warmup_steps

  def __call__(self, step):
    arg1 = tf.math.rsqrt(step)
    arg2 = step * (self.warmup_steps ** -1.5)

    return tf.math.rsqrt(self.d_model) * tf.math.minimum(arg1, arg2)
```

شکل زیر نمودار تغییرات نرخ یادگیری را نمایش می‌دهد:



قسمت دوم:

تصویر زیر نقشه حرارتی توجه برای یک جمله و میزان توجه هر توکن به بخش‌های مختلف در جمله را در head های متفاوت نمایش می‌دهد. در این نقشه، داده ورودی عبارت است از: Input: eles têm dinheiro e poder . و جمله ترجمه شده برابر است با: they have money and power. و جمله ترجمه اصلی نیز برابر است با : they have money and power. که نشان می‌دهد مدل به درستی توانسته است ترجمه جمله از پرتغالی به انگلیسی را انجام دهد.



یکی از اصلی ترین ایرادهای self-attention می‌توان به این مهم اشاره کرد، که این مکانیزم بیشترین توجه را به خود کلمه دارد انجام می‌دهد، به همین دلیل ماژول multi head attention معرفی شده است که هدف ان با ایجاد ماتریس‌های مختلف برای query و key و value این اجازه را می‌دهد تا مدل فضاهای representation مختلفی برای هر کلمه در نظر بگیرد. در این قسمت ما به جای اینکه برای هر کلمه از یک ماتریس w^Q و w^K و w^V استفاده کنیم(در این پروژه-مقاله اصلی) از ۸ سری از این ماتریس‌ها استفاده کرده‌ایم و مدل به جای اینکه یک ماتریس را یاد بگیرد باید ۸ سری از این ماتریس‌ها را یاد بگیرد. در انتها نیز به جای اینکه برای هر کلمه فقط یک بردار z(representation) داشته باشیم ۸ بردار خواهیم داشت که این ۸ بردار با هم الحاق خواهند شد؛ بدین‌گونه وقتی ۸ بردار z برای هر کلمه تولید کردیم در هر یک از فضاها کلمه مورد نظر به یکی از کلمات توجه می‌کند برای مثال کلمه poder در پرتغالی که معادل کلمهpower در انگلیسی می‌باشد در هر یک از head های تصویر بالا که بیانگر فضاهای توضیح داده شده می‌باشد میزان توجه خودش را به سایر کلمات نشان می‌دهد که بیشترین توجه مربوط به رنگ نارنجی و کمترین توجه مربوط به رنگ سرمه‌ای می‌باشد.

قسمت سوم:

۲۰ جمله از مجموعه دادگان تست و نمایش خروجی پیش‌بینی شده و خروجی مطلوب :

1)

Input: (risos) parece-me que todos vocês são tocs , astrofísicos e ultramaratonistas .

Predicted translation: (laughter) it seems to me all of you are interplanets , astrophysical and marbles .

Real translation:(laughter) you 're all cfo , astrophysicists , ultra-marathoners , it turns out .

2)

Input: sabemos mais do que eles ?

Predicted translation: do we know more than them ?

Real translation:do we know better than them ?

3)

Input: e isso afeta-nos a todos . ?

Predicted translation: and does that affect everyone . ?

Real translation:and that affects all of us .

4)

Input: eu não sou o meu pai .

Predicted translation: i 'm not my dad .

Real translation:i am not my father .

5)

Input: eles têm dinheiro e poder .

Predicted translation: they have money and power .

Real translation:they have money and power .

6)

Input: neste vídeo , podemos ver como um cateter muito fino leva a bobina até ao coração .

Predicted translation: in this video , we can see how a catheter get too fine to bina by putting it up to the heart .

Real translation:in this video , we can see how a very tiny catheter takes the coil to the heart .

7)

Input: depois , podem fazer-se e testar-se previsões .

Predicted translation: then they can do it and test if possible .

Real translation:then , predictions can be made and tested .

8)

Input: forçou a parar múltiplos laboratórios que ofereciam testes brca .

Predicted translation: it forced him to stop multiple laboratories that have offered bage tests .

Real translation:it had forced multiple labs that were offering brca testing to stop .

9)

Input: as formigas são um exemplo clássico ; as operárias trabalham para as rainhas e vice-versa .

Predicted translation: ig is a classic example ; opposed to quee and vice versal .

Real translation:ants are a classic example ; workers work for queens and queens work for workers .

10)

Input: uma em cada cem crianças no mundo nascem com uma doença cardíaca .

Predicted translation: one in every 100 children in the world are born with an heart disease .

Real translation:one of every hundred children born worldwide has some kind of heart disease .

11)

Input: neste momento da sua vida , ela está a sofrer de sida no seu expoente máximo e tinha pneumonia .

Predicted translation: at this moment of her life , she 's being suffered from aids in its exhibition and she had neural directon .

Real translation:at this point in her life , she 's suffering with full-blown aids and had pneumonia .

12)

Input: onde estão as redes económicas ?

Predicted translation: where are the economic networks ?

Real translation:where are economic networks ?

13)

Input: a partir daquele momento , comecei a pensar .

Predicted translation: from that moment i started thinking .

Real translation: at that moment , i started thinking .

14)

Input: a luz nunca desaparece .

Predicted translation: light never goes away .

Real translation: the light never goes out .

15)

Input: é um museu muito popular agora , e criei um grande monumento para o governo .

Predicted translation: it 's a very popular museum now , and i created a big monument to government .

Real translation: it 's a very popular museum now , and i created a big monument for the government .

16)

Input: e , no entanto , a ironia é que a única maneira de podermos fazer qualquer coisa nova é dar um passo nessa direção .

Predicted translation: and yet , the irony is that the only way we can do anything new thing is to take that step in that direction .

Real translation: and yet , the irony is , the only way we can ever do anything new is to step into that space .

17)

Input: este é o primeiro livro que eu fiz.

Predicted translation: this is the first book that i 've ever done in front of you .

Real translation: this is the first book i've ever done.

18)

Input: os meus vizinhos ouviram sobre esta ideia.

Predicted translation: my neighbors has heard about this idea of all .

Real translation: and my neighboring homes heard about this idea .

19)

Input: vou então muito rapidamente partilhar convosco algumas histórias de algumas coisas mágicas que aconteceram.

Predicted translation: so i 'm very quickly to share with you some magic stories that have happened .

Real translation: so i 'll just share with you some stories very quickly of some magical things that have happened .

20)

Input: este é um problema que temos que resolver.

Predicted translation: this is a problem that we have to solve for a generation .

Real translation: this is a problem we have to solve .