

Semantic Textual Similarity

فرید نظری منتظر

960122681008

هدف از این پروژه یافتن شباهت معنایی بین دو جمله ورودی است، برای این کار ما از شبکه عصبی siamese استفاده می کنیم و برای درک بهتر چگونگی انجام این کار ابتدا با فرمت داده های ورودی ما آشنا می شویم.

لینک به ویدیو توضیحات: [CI Project.mkv](#)

داده های ورودی:

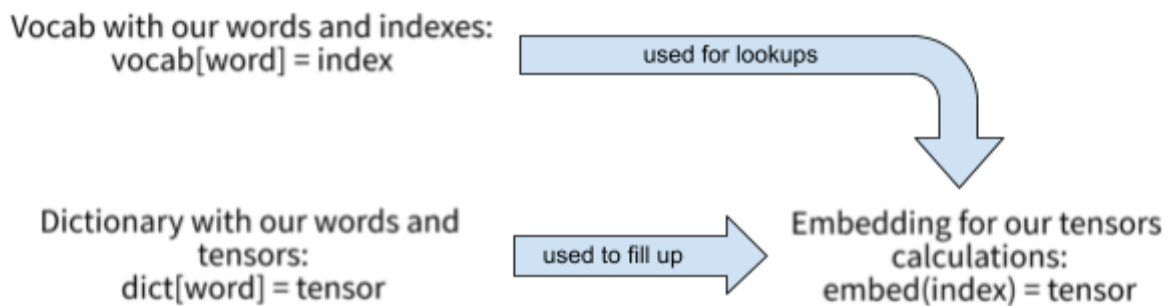
داده هایی که ما در شبکه خود سروکار داریم دارای دو جمله و یک برچسب برای شباهت دو جمله هستند از آنجایی که در داده اصلی برچسب ها به صورت کلمات هستند برای محاسبات آن ها را به عدد تبدیل می کنیم و بعد از حذف کردن ستون های اضافی به یک دیتاست این شکلی می رسمیم.

Sentence 1	Sentence 2	Label
Two dogs play	Two cats are playing	0
The skier is doing a flip through some trees	The skier is winning a competition	1
A young boy giving his brother a piggyback ride	a boy is giving his brother a ride	2

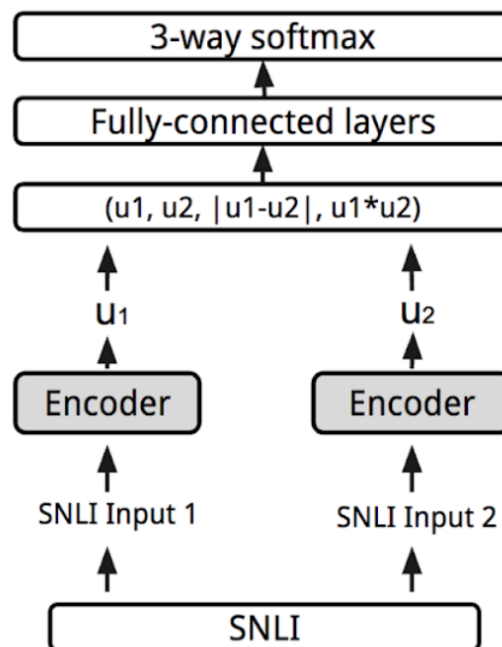
در قدم بعدی نیاز داریم که هر کلمه ی خاص یک tensor خاص برای نمایش خود و انجام محاسبات داشته باشد،

برای این کار یک set از کلمات درست کرده و سپس از آن یک vocabulary از کلمات ایجاد می کنیم که در آن ایندکس هر کلمه قرار می گیرد.

سپس از vocabulary یک dictionary درست کرده که دارای tensor های خود درون خود است در قدم بعد از embedding درون torch استفاده می کنیم برای محاسبات خود گرچه بدون gradient.

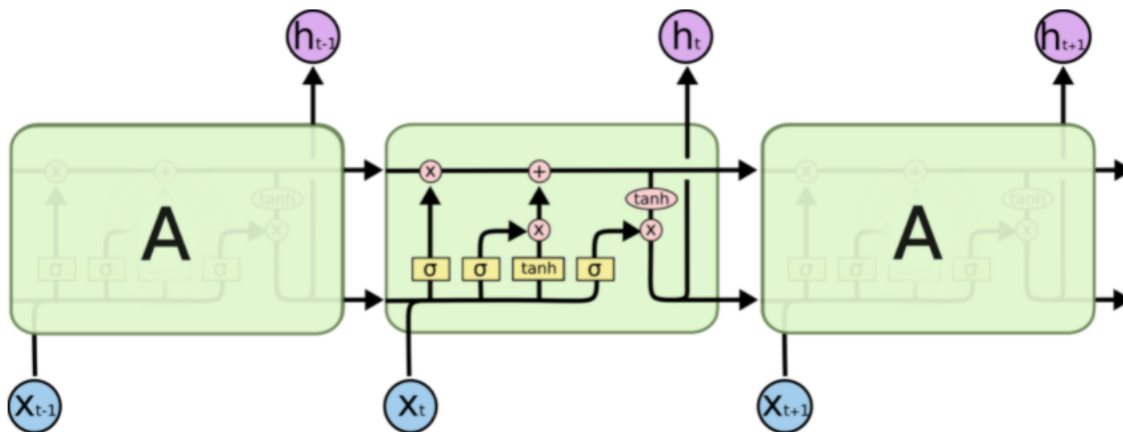


مدل شبکه:



طرز کار کلی دادن کلمات هر جمله به یک شبکه LSTM برای به دست آوردن یک tensor برای هر جمله سپس ترکیب دو tensor برای به دست آوردن یک tensor features به فرم $(u_1, u_2, |u_1 - u_2|, u_1 * u_2)$ و با گذراندن آن ها از layer های شبکه و در آخر با گذراندن از یک activation softmax به جواب مورد نظر رسیدن.

شبکه LSTM:



شبکه LSTM یک شبکه تکرار شونده یا RNN است که قابلیت به یاد آوردن داده ها با استفاده از یک cell state و یک hidden state.

که cell state تنها با مقداری تغییرات خطی از شبکه عبور می کند و امکان دارد در یک گذر تغییر خاصی نکند در حالی که hidden state با هر بار گذر تعدادی محاسبات روی آن انجام می گیرد که مقدار تاثیر گذاری آن از اکتیویشن sigmoid محاسبه می شود.

استفاده ما از LSTM برای این است که در ابتدا hidden state و cell state برای جمله می سازیم سپس با دادن هر کلمه این دو state برای محاسبه بعدی به جلو فرستاده می شوند که در آخر خروجی ما به عنوان tensor بیان گر جمله استفاده می شود.

```
def forward(self, input, hidden, cell):
    input_index = self.vocab[input]
    input = self.embeds(torch.tensor(input_index)).view(1, 1, -1)
    output, (hidden, cell) = self.lstm(input, (hidden, cell))
    return output, hidden, cell
```

در قدم بعدی با استفاده از concatenate یک tensor برای نمایش ویژگی های دو جمله و محاسبه یکی بودن، بی ربط بودن یا مخالف همدیگر بودن می سازیم.

```
features = torch.cat((vector1, vector2, torch.abs(vector1 - vector2), vector1 * vector2), 2)
```

و آن را از layer های خود عبور می دهیم که من تنها از دو linear layer استفاده کردم گرچه می توان بین آن ها از relu و sigmoid استفاده کرد.

```
layers = [
    nn.Linear(self.input_dim, int(self.input_dim / 2)),
    nn.Linear(int(self.input_dim / 2), 3)]
```

دیتاست استفاده شده:

دیتاست استفاده شده در این پروژه [snli 1.0](#) است که شامل 570 هزار جفت جمله است ولی از آنجایی که تعداد زیادی از آن ها 5 لیبل را ندارند من 20 هزار جفت آن را جدا کردم و استفاده کردم. دیتا های درون این دیتاست به شکل جمله اول به صورت بیان یک بحث فرضی و جمله دوم در تایید کامل، مخالفت یا خنثی با اولی است.

Hypothesis	Judgments	Text
The man is sleeping	contradiction C C C C C	A man inspects the uniform of a figure in some East Asian country.
Two men are smiling and laughing at the cats playing on the floor.	neutral N N E N N	An older and younger man smiling.
A man is driving down a lonely road.	contradiction C C C C C	A black race car starts up in front of a crowd of people.
Some men are playing a sport.	entailment E E E E E	A soccer game with multiple males playing.
A happy woman in a fairy costume holds an umbrella.	neutral N N E C N	A smiling costumed woman is holding an umbrella.

در ابتدا من برای استفاده از Mean Square Error loss به نسبت 5 برچسب امتیاز دهی یک امتیاز از 0 تا 5 دادم ولی از آنجایی که در آخر از Cross Entropy loss استفاده کردم تصمیمات نهایی را تبدیل به امتیاز 0 برای مخالفت، 1 برای خنثی و 2 برای تایید استفاده کردم گرچه جدا کردن تایید و خنثی از هم باعث کمتر شدن دقت ما در تعیین شباهت جملات می شود ولی سعی کردم پروژه کمی متفاوت باشد.

```
dataset['score'] = (0 +
((dataset['gold_label'] == 'neutral').astype(int) * 1) +
((dataset['gold_label'] == 'entailment').astype(int) * 2))
```

برای تقسیم داده به سه قسمت train, dev و test من از نسبت های 20، 60 و 20 استفاده کردم گرچه در زمان آدرس دهی به دیتا در فایل Jupyter Notebook می توان مقدار آن ها را تغییر داد.

```
def __init__(self, data_path, config, validation_ratio = 0.2, test_ratio = 0.2):
```

ساخت دیکشنری لغات:

برای ساخت دیکشنری ابتدا از تمامی لغات یک ست می سازیم از آنجایی که داده های درون ست تنها خاص هستند از آن برای تک بودن لغات استفاده می کنیم سپس یک دایره لغات به صورت دیکشنری پایتون از آن ست می سازیم و برای هر کدام یک ایندکس فرار می دهیم.

```
for i in range(len(sentence_pairs)):
    words = sentence_pairs[i][0].split()
    words += sentence_pairs[i][1].split()
    for j in range(len(words)):
        vocab_set.add(words[j])
vocab = {}
i = 0
for word in vocab_set:
    vocab[word] = i
    i += 1
```

سپس از روی دیکشنری دایره لغات یک دیکشنری tensor تولید می کنیم، برای این کار من از numpy استفاده کردم و embedding_dim نیز قابل تنظیم در فایل Jupyter Notebook می باشد.

```
current_vectors = {}
for word in word_vocab:
    if word not in mother_vectors:
        mother_vectors[word] = np.array(np.random.uniform(-1.0, 1.0, embedding_dim))
    current_vectors[word] = mother_vectors[word]
```

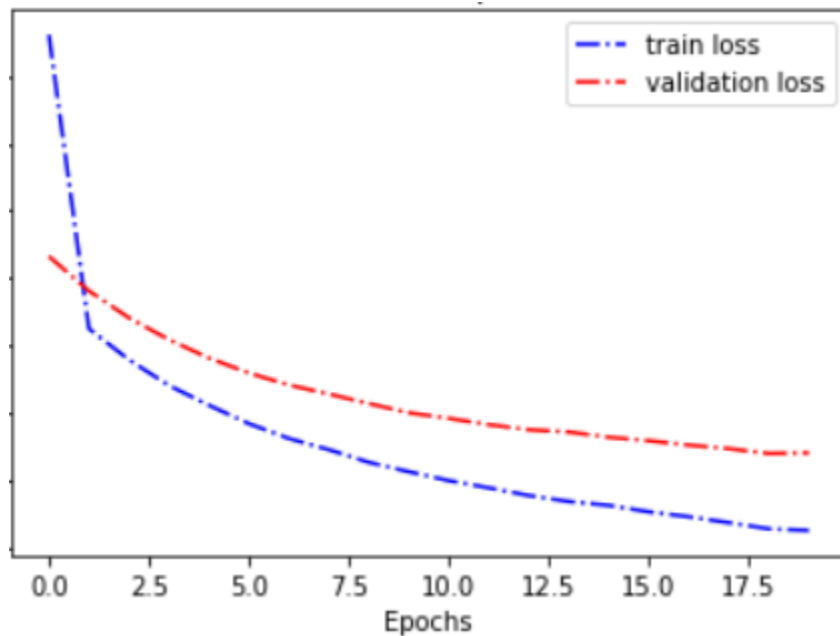
در پایان نیز آرایه رندوم 300 tensor بین منفی 1 تا 1 را درون یک embedding قرار می دهیم.

```
weights = []
for word in vocab:
    weights.append(vectors[word])
embedding = nn.Embedding(vocab_size, config['model']['embed_size'])
weights_matrix = torch.from_numpy(np.array(weights)).type(torch.FloatTensor)
embedding.weight = nn.Parameter(weights_matrix, requires_grad = False)

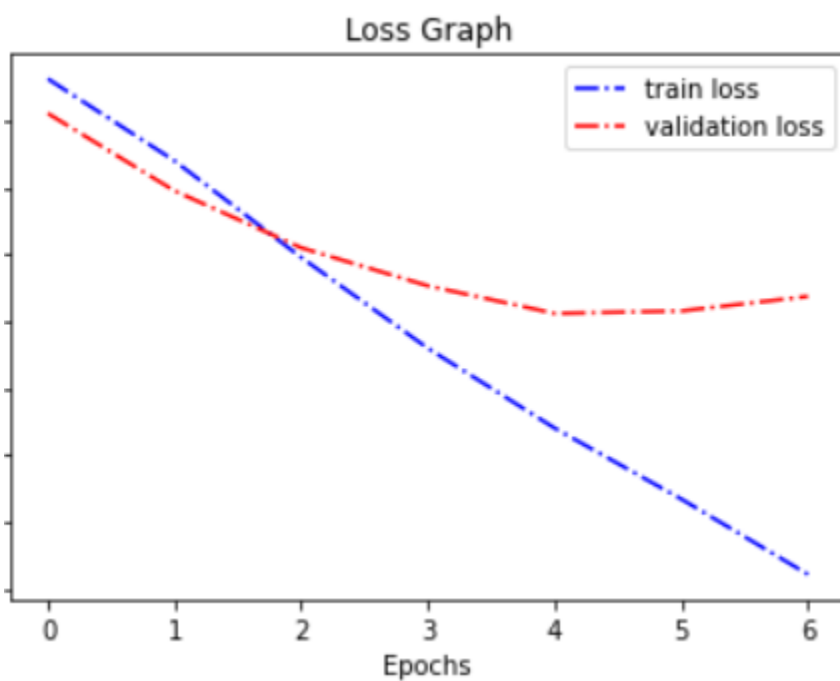
config['vocab'] = vocab
config['embedding'] = embedding
```

نمودار Loss:

گرچه با mse به مقدار بهتری از Loss می توان رسید اما از آنجایی که در تعداد epoch های پایین بیشتر به میانگین نزدیک میشود (90% پیش بینی ها بین 0.7 تا 1.3)



با optimizer adam در epoch های پایین می توان به پیش بینی های تقریباً منطقی رسید.



طریق کار Cross Entropy Loss:

محاسبات کلی آن به شکل زیر می باشد.

$$\text{loss}(x, \text{class}) = -\log \left(\frac{\exp(x[\text{class}])}{\sum_j \exp(x[j])} \right) = -x[\text{class}] + \log \left(\sum_j \exp(x[j]) \right)$$

و ایده کلی آن این است که یک tensor برای target به آن می دهیم و یک تانسور برای نشان دادن شانس هر عدد long که target ما یکی از این اعداد است و کار CEL اینگونه است که loss فاصله با جواب درست را محاسبه کند و جواب پایانی را نیز از این loss از طریق اکتیویشن softmax می توان بدست آورد.

پیش بینی های پایانی:

برای پیش بینی های پایانی من از activation Softmax استفاده کردم که مقدار شانس هر امتیاز را به ما باز می گرداند و با ضرب دادن به نسبت امتیاز می توان تقریباً به پیش بینی درستی رسید.

```
softmax = nn.Softmax(dim = 0)
output = softmax(output)
output = output[1] * 1 + output[2] * 2
```

پاسخ های نهایی ما گرچه دقت زیاد بالایی ندارند ولی در مقایسه با پروژه های مشابه که تنها از دو حالت مخالف یا موافق برای امتیاز دهی استفاده می کنند و دارای loss میانگین 0.5 تا 0.6 هستند، می توان گفت پیش بینی های ما مقدار قابل قبولی دارند.

sentence 1 ▼	sentence 2 ▲ ▼	score	prediction	loss
A black dog is digging in the snow.	A cat running down a beach	0	0.47	0.4
A dog with a ball in his mouth running down a road covered in leaves.	A cat running up a tree	0	1.02	1.02
A furry dog is jumping over a hurdle.	A cat sleeps in the window.	0	0.43	0.35
An older man is kneeling and talking to a group of younger children, all in soccer uniforms.	a cat sleeps soundly	0	0.58	0.65
A chef in a kitchen cuts a piece off of a large slab of rotating meat on a stick, preparing a meal.	A chef is cooking up a 6-course menu for his close friends.	1	0.82	0.26
A gentleman cooks multiple dishes while preparing extra ingredients in his kitchen.	A chef prepares a meal for a food critic.	1	0.92	0.68
A chef speaks into a microphone about a mixed salad contained in a bowl with a fork.	A chef speaks	2	0.83	1.01
An adorable child leaping to block a soccer ball from entering the goal.	A child blocks a goal with his hockey stick.	0	0.87	0.81
A child and small dog running.	A child chases his dog back into the house.	1	0.72	0.51
A young boy in red leaping into sand at a playground.	A child does a cannonball into a pool.	0	0.8	0.7
A young child with dark shoulder length hair, in a yellow shirt and pink pants, is sitting on a chair.	a child eats cheerios	2	1.26	0.61
A child dressed in purple clothing and a hat runs in splashing water while a man watches.	a child enjoys summer vacation	1	1.1	1.23
A little child is looking towards someone while their head is in a painting designed by a man.	A child has fallen at an art museum as a man helps him up.	0	0.83	1.04
A boy wearing a yellow shirt, khaki pants, and blue and yellow shoes is hitting a ball.	A child in a colorful outfit, holds a piece of sporting equipment and uses it to play.	2	0.8	1.72
A child with blond-hair looking out of a screen door.	A child is being held by his mother at the mall.	0	0.5	0.6
A child in a red jacket, waist deep in a pit on the beach.	A child is building a sandcastle on the beach.	0	1.16	1.31
A boy toddler is holding a bat he is about to swing at a pinata.	A child is crying because he didn't get cake	0	0.78	0.76
A young blond child eats chocolate cake with his fingers while holding the spoon.	a child is eating vegetables	0	0.85	0.68
A wet child stands in chest deep ocean water.	A child is in the ocean	2	0.97	1
A child, wearing a tie, jumps off steps.	A child is jumping off the step and into a puddle.	1	1.15	0.71
Young girl in an bathing suit falling off the end of a slide into a pool.	A child is moving swiftly toward a body of water.	2	1.24	0.74
A little child is looking towards someone while their head is in a painting designed by a man.	A child is observing an artist doing his work.	1	0.88	0.44
Under a blue sky with white clouds, a child reaches up to touch the propeller of a plane.	A child is playing with a ball.	0	1.36	1.4
A small child playing in a dusty square.	A child is playing with a doll.	1	1.4	1.76
The blond-haired and blue eyed child holds the wooden airplane in his hands.	A child is playing with a puppy.	0	1.06	0.84
The blond-haired and blue eyed child holds the wooden airplane in his hands.	A child is playing with a toy airplane.	2	1.28	0.61
A small child in a coat looks out over the lake of a park in the fall.	A child is playing with his legos in his warm room.	0	0.72	0.82
A child in a blue shirt is sitting on a bench.	A child is sitting on a bench.	0	1.55	0.86

با تشکر
پایان