دانشکده مهندسی کامپیوتر

استاد درس: دکتر اعتمادی

بهار ۱۴۰۰

# گزارش پروژه
## درس مبانی پردازش گفتار و زبان

**پریسا یل‌سوار**
**شماره دانشجویی: ۹۶۵۲۲۰۸۷**

# فهرست مطالب


۱ word2vec ۵

۲ language modeling ۶

۳ fine tuning ۷

۱.۳ مدل زبانی . . . . . ۷

۲.۳ رده بندی . . . . . ۷

# فهرست تصاویر

۱ عملکرد مدل رده بند روی داده تست . . . . . ۸

# فهرست جداول

۱ مقایسه perplexity مدل زبانی ژانرها . . . . . ۷
۲ مقایسه جملات تولید شده . . . . . ۷
۳ مقایسه جملات تولید شده . . . . . ۸

# ۱ word2vec

## ۲ language modeling

در این قسمت از ۳ لایه LSTM با اندازه ۱۲۸ و اندازه embedding ۱۲۸ به عنوان مدل زبانی استفاده شده است [۱]. مدل هر ژانر از مجموعه داده جملات تمیز شده استفاده می کند و به مدت ۵ دور با طول sequence ۱۵ آموزش می بیند.

| Perpelexity | Genre |
|---|---|
| 735.4264358106171 | Action |
| 850.6009242951887 | Adventure |
| 762.2966998333878 | Fantasy |
| 630.12191391778 | Drama |
| 621.3412662654786 | Comedy |
| 577.2775060373292 | Romance |

جدول ۱: مقایسه perplexity مدل زبانی ژانرها

| Sentence | Genre |
|---|---|
| an accident NUM year ago left mrunas | Action |
| an accident dimension finally meet fate new member group come | Adventure |
| an accident happens people think wrongperson named ik | Fantasy |
| an accident set fire inside roomtakanagi | Drama |
| an accident also happens accidentally end yukichi girl | Comedy |
| an accident dimension world called nanamiunusual day | Romance |

جدول ۲: مقایسه جملات تولید شده

# ۳ fine tuning

## ۱.۳ مدل زبانی

از بین ۱۲ ژانر موجود؛ ۶ ژانر اول با بیشترین تعداد جمله برای تسک fine-tune انتخاب شدند. در این قسمت از کد huggingface [۲] و مدل Distil GPT2 برای fine-tune روی داده هر کلاس استفاده شده است. مدل زبانی هر ژانر از نوع causal است و برای ۶ دور با batch size ۱۰۰۰ fine-tune می شود.

همچنین داده هر ژانر به نسبت ۸۰ به ۲۰ به دو قسمت آموزش و ارزیابی تقسیم می شود. از داده ارزیابی در زمان fine-tune و در پایان برای اندازه گیری perplexity استفاده می شود. از دیگر تنظیمات مدل می توان به اعمال weight decay و توقف زودهنگام (early stopping) اشاره کرد. مقایسه perplexity مدل ها در جدول ۱ آورده شده است. از آنجایی که زمان fine-tune کوتاه بوده و تعداد جملات برای هر ژانر زیاد نیست، perplexity ها مقادیر خوبی ندارند.

تعدادی از جملات تولید شده توسط مدل ها در جدول ۲ بررسی شده اند. کلمات آبی شده توسط مدل تولید شده اند. کلمات ناقص به دلیل حذف کردن stopwords از داده اصلی به وجود آمده اند. در این مثال ژانرهای Action، Adventure و Drama جملات مناسب و مرتبطی تولید کرده اند.

در مثال دوم در جدول ۳ جمله تولید شده توسط ژانرهای Action، Adventure و Fantasy کاملا حال و فضای ژانر را دارد و بقیه ژانرها نیز به نسبت کلمات مناسبی تولید کرده اند.

## ۲.۳ رده بندی

چالش رده بندی ژانر بر اساس خلاصه انیمه multilabel classification نام دارد. در این قسمت هر خلاصه ۱۲ برچسب دارد و بر پایه اینکه یک ژانر به این خلاصه تعلق داشته باشد مقدار صفر یا یک می گیرد [۳]. مدل

| Sentence | Genre |
|---|---|
| he was from far away girl come new world | Action |
| he was from the previous generation ai japan | Adventure |
| he was from another dimension time travel another dimension new | Fantasy |
| he was from hino one day find dead another | Drama |
| he was from home village made thing even worsehow | Comedy |
| he was from heaven told never say anything one day | Romance |

جدول ۳: مقایسه جملات تولید شده

bert-base-uncased برای ۴ دور با تکنیک cross validation و تابع خطا binary cross entropy آموزش می بیند. در پایان آموزش دقت مدل روی داده validation 18.52 و F1 score آن 58.67 است.

شکل ۱ عملکرد مدل را روی داده تست نشان می دهد. پارامتر support نشان دهنده تعداد خلاصه ها با یک ژانر مشخص هستند. بیشترین دقت متعلق به ژانر Romance است چون این ژانر بیشترین تعداد خلاصه را دارد. ژانر های School life و Fantasy نیز دقت بالایی دارند که می تواند به دلیل شباهت یا همراه بودن با ژانر Romance باشد (انیمه هایی که ژانر Romance دارند اکثر ژانرهای School life و Fantasy نیز دارند). ژانر Drama دقت و recall پایینی دارد. دلیل آن می تواند برجسته نبودن ژانر Drama باشد، به عبارت دیگر در مقایسه با ژانر Action یا Romance که کلمات مشخصی از قبیل fight یا love دارند، این ژانر کلمات مشخصی ندارد به همین دلیل تشخیص آن سخت تر است. ژانر Sci Fi نیز به دلیل کم بودن تعداد داکیومنت ها recall پایینی دارد که با افزایش داده می توان این مشکل را حل کرد. فایل *comparisons.csv* تمام پیش بینی ها و برچسب های درست برای داده تست نشان می دهد.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Action | 0.67 | 0.65 | 0.66 | 873 |
| Adventure | 0.62 | 0.42 | 0.50 | 462 |
| Comedy | 0.65 | 0.48 | 0.55 | 1145 |
| Drama | 0.55 | 0.40 | 0.46 | 1111 |
| Fantasy | 0.72 | 0.65 | 0.68 | 819 |
| Romance | 0.74 | 0.72 | 0.73 | 1560 |
| School Life | 0.71 | 0.46 | 0.56 | 657 |
| Sci Fi | 0.61 | 0.39 | 0.47 | 312 |
| Shoujo | 0.59 | 0.45 | 0.51 | 476 |
| Shounen | 0.68 | 0.18 | 0.29 | 489 |
| Supernatural | 0.66 | 0.38 | 0.48 | 468 |
| Yaoi | 0.69 | 0.69 | 0.69 | 505 |
| micro avg | 0.67 | 0.52 | 0.59 | 8877 |
| macro avg | 0.66 | 0.49 | 0.55 | 8877 |
| weighted avg | 0.66 | 0.52 | 0.58 | 8877 |
| samples avg | 0.64 | 0.54 | 0.55 | 8877 |

شکل ۱: عملکرد مدل رده بند روی داده تست

# مراجع

[1] D. Bitvinskas, *Pytorch lstm: Text generation tutorial*, Web Page. [Online]. Available: https://closeheat.com/blog/pytorch-lstm-text-generation-tutorial.

[2] huggingface, *Language_modeling*, Web Page. [Online]. Available: https://github.com/huggingface/notebooks/blob/master/examples/language_modeling.ipynb.

[3] R. Patel, *Transformers for multi-label classification made simple.* Web Page. [Online]. Available: https://towardsdatascience.com/transformers-for-multilabel-classification-71a1a0daf5e1.