چالش پیش‌بینی میزان مصرف اینترنت مشترکین همراه اوّل

Order of the Phoenix

(دانشگاه صنعتی خواجه نصیرالدّین طوسی)

سالار حسینی شمعچی

احمد احمدی

محمد حسین‌زاده

روش حلّ مسئله:

شبکه‌های عصبی حافظه‌ی کوتاه-بلند مدّت[1] به دلیل عملکرد پرقدرت در تحلیل داده‌های ترتیب‌دار، به خصوص سری‌های زمانی که در آن‌ها اثر طولانی مدت نیز حائز اهمیت است، از جایگاه ویژه‌ای برخودار گشتند. در این چالش از آنجا که نوع داده‌های مورد استفاده بر اساس سری زمانی در نظر گرفته شده است، این نوع شبکه به عنوان ساختار مطلوب برای حل مسئله انتخاب شده است.

برای رسیدن به بهترین نتایج ممکن، ابتدا روش مناسب پیش پردازش داده‌ها بررسی و انتخاب شده، سپس با تحلیل و آزمون و خطای مکرر، ساختار و پارامترهای مناسب برای شبکه‌ی LSTM برگزیده شده‌اند. سپس نتایج شبکه با روش‌های مختلف سنجیده شده و با استفاده از بهترین شبکه‌ی آموزش دیده، خروجی‌های لازم پیش‌بینی شده‌اند. در ادامه جزئیات این روند مورد بحث قرار می‌گیرد.

پیش پردازش داده‌ها:

ابتدا درون یک حلقه داده‌های مربوط به هر شخص از فایل آموزش و فایل هفته‌ی اول جدا شده و جای خالی داده‌های گم شده با استفاده از میانگین گیری از ویژگی مربوط به آن داده پر می‌شود. سپس داده‌های ورودی مربوط به آن شخص به صورت یک سری از داده‌های چند روز متوالی انتخاب می‌شود. این داده‌ها شامل تمام ویژگی‌های داده شده مربوط به هر روز بوده که حجم اینترنت مصرفی هر روز را نیز در بر می‌گیرد، اما حجم اینترنت مصرفی روز آخر هر سری صفر قرار داده شده و مقدار واقعی آن به عنوان خروجی آن سری در نظر گرفته شده است. تعداد روزهای هر سری نیز از روز اول تا ۵۰ روز مانده به روز آخر تا تمامی روزهای هر شخص می‌تواند متغیر باشد. نهایتا ویژگی‌های مربوط به روزهای پیشین در سری‌های کوتاه همگی به عنوان صفر در نظر گرفته شده‌اند، تا تمامی سری‌های آموزشی طول یکسان داشته باشند.

تمامی داده‌ها پس از گذراندن مراحل پیش پردازش و آماده سازی سری‌ها ابتدا به سه دسته‌ی آموزش، صحه گذاری و آزمایش تقسیم شده‌اند تا معیار خوبی برای سنجش عملکرد شبکه داشته باشیم. پس از رسیدن به ساختار و پارامترهای مطلوب برای شبکه، این سه دسته به دو دسته‌ی آموزش و صحه گذاری کاهش داده شده‌اند، تا از نهایت پتانسیل داده‌های موجود استفاده شده و نتیجه‌ی مطلوب

---

[1] Long-short Term Memory (LSTM)

حاصل شود. نهایتا این دادها بر اساس مقادیر کمینه و بیشینه‌ی داده‌های آموزش نرمالیزه شده و این مقادیر نیز برای استفاده‌های آتی ذخیره می‌شوند.

پارامترها و ساختار شبکه:

به عنوان ساختار از یک شبکه‌ی نسبتا ساده‌ی LSTM با دو لایه‌ی میانی که ۱۶ واحد دارند استفاده شده است. دلیل استفاده از این ساختار ساده نیز جلوگیری از مشکل overfit شدن می‌باشد. از آنجا که توابع فعالساز لایه‌ی LSTM مخصوص سری‌های مرتب بهینه سازی شده‌اند، دخالتی در تغییر این توابع صورت نگرفته است. در لایه‌ی آخر نیز یک ساختار ساده‌ی dense با یک نورون و یک تابع فعالساز خطی قرار گرفته تا حجم اینترنت مصرفی مربوط به روز آخر سری را پیش‌بینی کند.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, None, 32)]        0

 lstm (LSTM)                 (None, None, 16)          3136

 lstm_1 (LSTM)               (None, 16)                2112

 dense (Dense)               (None, 1)                 17

=================================================================
Total params: 5,265
Trainable params: 5,265
Non-trainable params: 0
_________________________________________________________________
```

شکل ۱- ساختار شبکه

برای آموزش شبکه، تابع هزینه‌ی انتخاب شده میانگین مربع خطاها[۲] است که با استفاده از تابع بهینه‌ساز Adam به مقدار کمینه می‌رسد. سپس داده‌ها به صورت دسته‌های[۳] ۵۱۲ تایی با تکرار ۳۰۰۰ مرتبه به شبکه داده شده و با تنظیم وزن‌ها روند آموزش پیش می‌رود. در طول آموزش نیز هزینه‌ی ناشی از داده‌های صحه گذاری مورد بررسی قرار می‌گیرند. تمامی این پارامترها نیز با آزمون و خطای بسیار بهینه سازی شده‌اند.

---

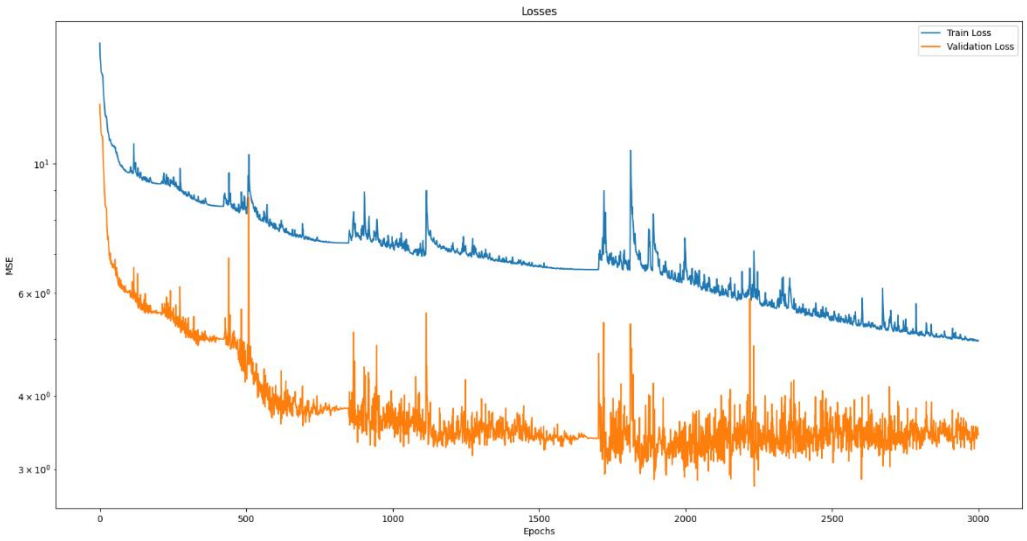
[۲] Mean Squared Error

[۳] Batches

برای جلوگیری بیشتر از مشکل overft شدن و هدایت روند آموزش در راستای مطلوب، از ترفندهایی همچون تنظیم کننده‌ی وزن‌های شبکه[۴]، کاهش نرخ یادگیری کسینوسی[۵] و ذخیره‌ی بهترین وزن‌ها در طول آموزش استفاده شده است، که هر یک تاثیر مثبت خود را در مراحل آزمون و خطا نشان داده‌اند.

نتایج:

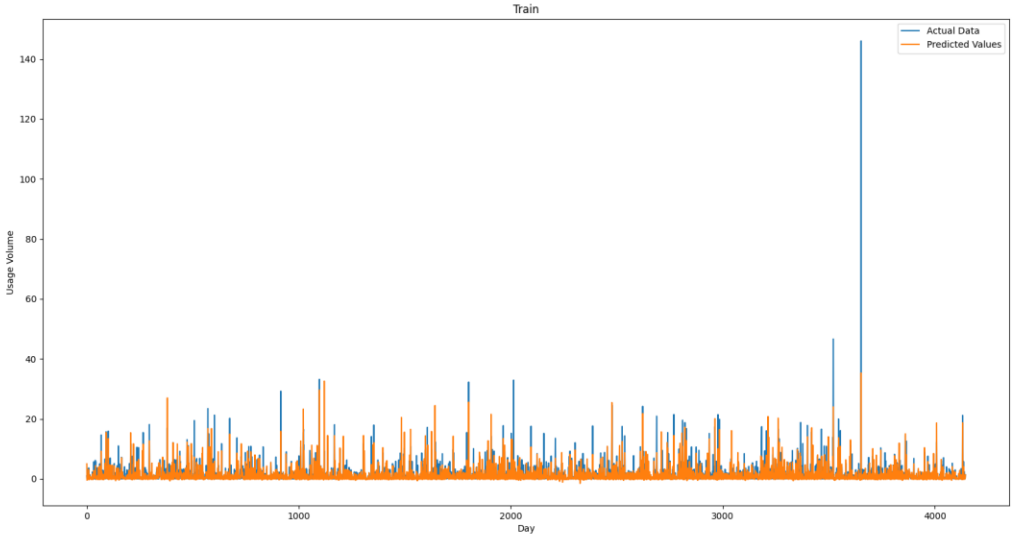
پس از اتمام یادگیری ابتدا برای بررسی عملکرد شبکه، آن را با استفاده از داده‌های صحه گذاری و آموزشی ارزیابی کرده، امتیاز R2 آن را محاسبه می‌کنیم، که برای داده‌های آموزش و صحه گذاری به ترتیب ۰.۵۷۷۳ و ۰.۷۶۵۷ به دست می‌آید. علاوه بر آن نمودار تغییرات تابع هزینه برای داده‌های آموزش و صحه گذاری را نیز برای ارزیابی روند آموزش رسم می‌کنیم که در شکل ۲ قابل مشاهده است. نهایتا مقادیر پیش‌بینی شده توسط شبکه برای داده‌های آموزش و صحه گذاری را در برابر مقادیر واقعی رسم می‌کنیم تا عملکرد شبکه به صورت بصری مشخص شود. این نمودارها نیز در شکل‌های ۳ و ۴ قابل مشاهده هستند.



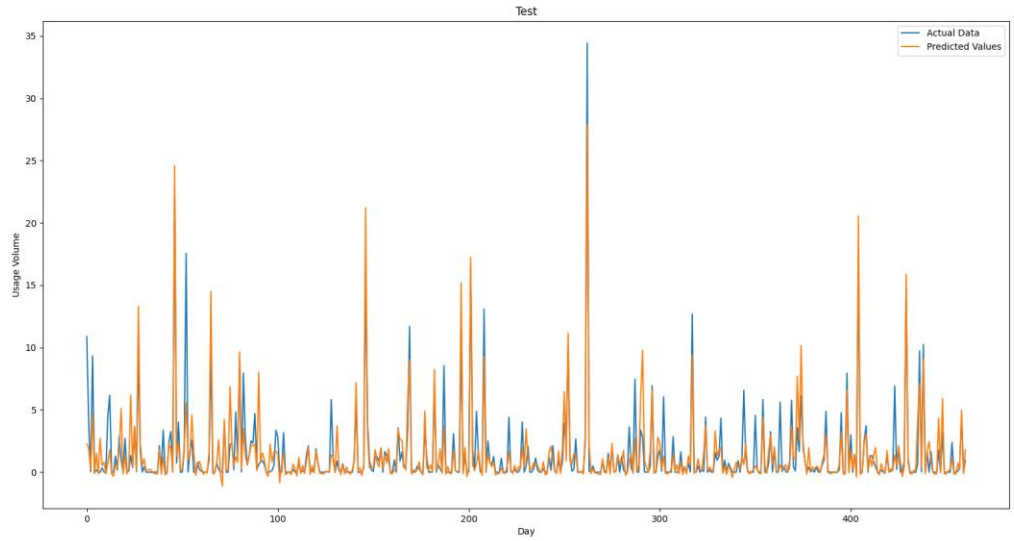
شکل ۲- روند تابع هزینه برای داده‌های آموزش و صحه گذاری

[۴] Kernel Regularizer

[۵] Cosine Learning Rate Decay



شکل ۳- مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های آموزش



شکل ۴- مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری

پس از ارزیابی عملکرد شبکه با روش‌های ذکر شده، داده‌های هفته‌ی دوم به عنوان بخشی از ورودی به شبکه داده شده و حجم اینترنت مصرفی کاربران به عنوان خروجی شبکه گرفته شده و ذخیره می‌شود. این نتایج در فایل week2_results.csv قابل دسترسی هستند.

لازم به ذکر است برای ساخت داده‌های خروجی باید از فایل main_load.py استفاده شود.