

به نام خدا



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)

درس شبکه‌های عصبی  
استاد صفابخش

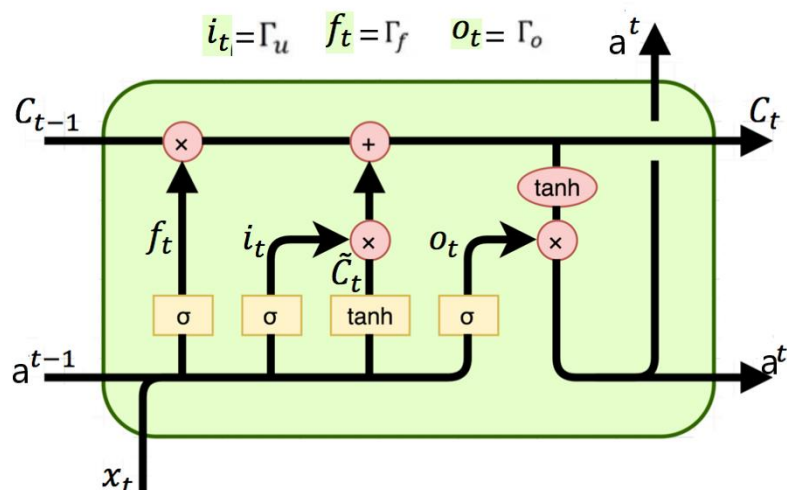
تمرین ششم

علیرضا مازوچی

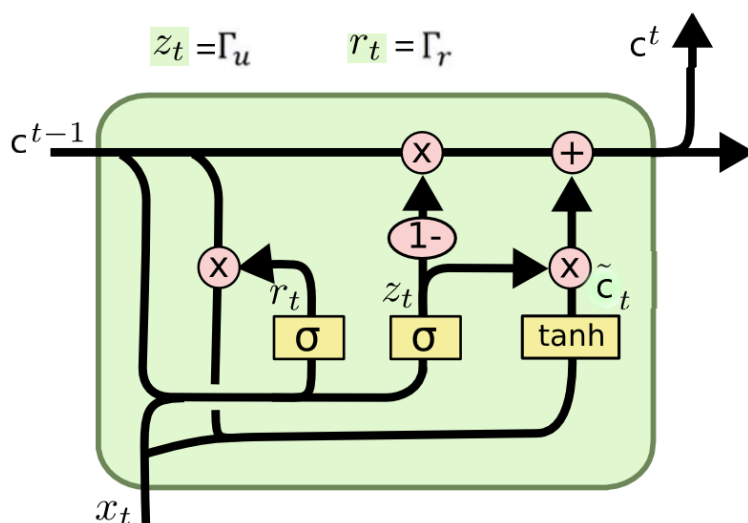
۴۰۰۱۳۱۰۷۵

## سوال ۱

ساختار شبکه LSTM به شکل زیر است:



ساختار شبکه GRU به شکل زیر است:



در RNN و با گذشت هر گام زمانی اطلاعات قدیمی‌تر تا حد زیادی از بین می‌رود و دانش فعلی بیشتر باقی می‌ماند. شباهت هر دو شبکه در این است که هر دو با بهبودی بر RNN توانسته‌اند اطلاعات مفیدتر از زمان‌های قدیمی‌تر را حفظ کنند. حفظ کردن اطلاعات قدیمی از طریق یک واحد حافظه صورت می‌گیرد. طبیعتاً باید نوشتن و خواندن از این حافظه توسط شبکه یاد گرفته شود. این منطق اگرچه بین دو شبکه یکسان است ولی جزئیات آن تا حدی متفاوت است.

در LSTM سه گیت با نام‌های input، output و forget طراحی شده است. گیت forget تعیین می‌کند که چه میزان از اطلاعات وضعیت قبلی باید نگهداشته شود و چه میزان از اطلاعات جدید باید حفظ شود. گیت input میزان تاثیرگذاری ورودی را مشخص می‌کند و نهایتاً گیت output مقدار خروجی بر اساس وضعیت فعلی را کنترل می‌کند. در GRU سعی شده است از یک معماری ساده‌تر با حفظ کارایی حدودی LSTM استفاده شود. در GRU تنها دو گیت به نام‌های reset و update وجود دارد؛ گیت update مشخص می‌کند که تا چه میزان اطلاعات جدید اهمیت دارد و باید در وضعیت سلول جایگزین شود و گیت reset تعیین می‌کند که تا چه میزان اطلاعات قبلی اهمیت دارد و باید نگهداشته شود.

باتوجه به موارد بیان‌شده بالا به نظر می‌رسد تفاوت‌ها چندان جدی نیست و بسته به تسک و مجموعه داده ممکن است هر یک تا حد کمی از دیگری بهتر باشد ولی اگر بخواهیم خیلی جزئی نگاه کنیم، به نظر می‌رسد که شبکه GRU به دلیل ساختار ساده‌تر توان پردازشی کمتری را می‌طلبد و استعداد آن برای بیش‌برازش کمتر است؛ اما این شبکه کنترل کمتری بر روی داده‌ها دارد و طبیعتاً قدرت آن از LSTM کمتر خواهد بود. نهایتاً اگر بخواهیم به سوال مطرح شده یعنی «استفاده از هر یک از این واحدها را در چه شرایطی توصیه می‌کنید؟» پاسخ دهیم، باید گفت که بهتر است هر دو شبکه تست شود و آن شبکه با نتایج بهتر حفظ شود ولی اگر قرار به انتخاب باشد، GRU بر روی حجم عظیم از داده‌ها به دلیل مشکلات زمان اجرا مناسب‌تر است و LSTM برای داده‌های با اندازه معمولی و کاربردهایی که مستلزم استخراج روابط پیچیده‌تری در طول زمان است، مناسب‌تر است.

## سوال ۲

### ملاحظات کلی

ابتدا نکات کلی در مورد مجموعه داده را بیان می‌کنم:

- برای ساخت مجموعه داده تنها از داده‌های یک سال اخیر استفاده شده است و از هر جدول ستون‌های <OPEN>، <HIGH>، <LOW>، <CLOSE>، <VOL>، <OPENINT>، <OPENINT>1 و <OPENINT>2 انتخاب شده است.
- تعداد کل داده‌ها در یک سال اخیر تنها ۲۳۷ تا بوده است. برای آنکه فرمت داده مناسب برای پیش‌بینی سری زمانی شود، یک پنجره لغزان با اندازه ده روز در نظر گرفتم و آن را روی داده‌ها لغزاندم. بدین ترتیب هر داده جدید شامل اندازه شاخص‌های مختلف برای ده روز و یک برچسب که تعیین‌کننده صعودی یا نزولی بودن شاخص برای روز بعدی است خواهد بود.
- داده‌ها به صورت تصادفی شافل شده‌اند و با نمونه‌برداری طبقه‌ای به سه دسته آموزشی، اعتبارسنجی و ارزیابی مطابق با نسبت‌ها ذکر شده در صورت سوال تقسیم شده‌اند. در هر سه مجموعه داده حدوداً ۵۶٪ داده‌ها برچسب یک، به معنای صعودی بودن شاخص، را دارند.

حال نکاتی درباره مدل و آموزش آن بیان می‌کنم:

- برای آموزش بهتر از لایه Batch Normalization در تمام مدل‌ها استفاده شده است. باتوجه به رنج متفاوت مقادیر ویژگی‌های مختلف وجود این لایه کمک زیادی به آموزش مناسب مدل می‌کند.
- یک لایه Dropout در لایه قبل از خروجی استفاده شده است. برای بررسی تاثیر آن سه مقدار مختلف ۰، ۰/۵ و ۰/۸ استفاده شده است که مقدار ۰ عملاً به معنای نبودن این لایه است.
- از یک کالبد Early Stopping برای جلوگیری از بیش‌برازش استفاده کردم و تعداد گام را بر روی ۳۰۰ تنظیم کرده‌ام که بدین ترتیب بر اساس خطای اعتبارسنجی و در جای مناسب آموزش خاتمه پیدا خواهد کرد.

- از دو بهینه‌ساز مختلف Adam و SGD استفاده شده است.
- برای هر دو شبکه GRU و LSTM تعداد سلول‌های مختلف ۸، ۳۲، ۶۴ و ۲۵۶ ارزیابی شده است.
- برای آنکه دقت‌های بدست آمده عادلانه باشد، بر روی هر تنظیم نتایج پنج بار اجرا درنظر گرفته شده است و میانگین آن به عنوان نتایج نهایی گزارش شده است.

اگرچه چنین چیزی همواره صادق نیست ولی به طور کلی مشاهده کردم که بهینه‌ساز Adam از بهینه‌ساز SGD نتایج بهتری گرفته است و Dropout با نرخ ۰/۸ بهترین بوده است. به عنوان مثال برای اولین دسته از نتایج که مربوط به یک شبکه با یک لایه هشت سلولی LSTM است، در جدول زیر دقت‌های ارزیابی آورده شده است:

Dropout 0.8	Dropout 0.5	Dropout 0	
۶۰/۴٪	۵۸/۳٪	۵۴/۳٪	بهینه‌ساز Adam
۵۳/۰٪	۵۲/۶٪	۴۶/۵٪	بهینه‌ساز SGD

بر اساس این مشاهدات و برای سهولت در ادامه تنها نتایج مربوط به بهینه‌ساز Adam و شبکه با یک لایه Dropout با نرخ ۰/۸ را گزارش کرده‌ام.

## نتایج شبکه تک لایه

نتایج مربوط به شبکه یک لایه از سلول‌های LSTM بدین شکل است:

تعداد سلول	دقت مجموعه آموزشی	دقت مجموعه اعتبارسنجی	دقت مجموعه ارزیابی
۸	۵۹/۴٪	۵۳/۰٪	۶۰/۴٪
۳۲	۵۵/۲٪	۵۸/۳٪	۵۴/۳٪
۶۴	۵۳/۵٪	۵۹/۱٪	۵۳/۰٪
۲۵۶	۵۳/۹٪	۵۳/۹٪	۵۳/۹٪



نتایج مربوط به شبکه یک لایه از سلول‌های GRU نیز در ادامه آورده شده است:

تعداد سلول	دقت مجموعه آموزشی	دقت مجموعه اعتبارسنجی	دقت مجموعه ارزیابی
۸	۵۹/۹٪	۵۵/۷٪	۵۵/۲٪
۳۲	۶۱/۳٪	۵۵/۷٪	۵۱/۷٪
۶۴	۵۲/۹٪	۵۳/۹٪	۴۷/۸٪
۲۵۶	۶۳/۲٪	۵۵/۷٪	۵۳/۹٪

هم برای شبکه LSTM و هم برای شبکه GRU می‌بینیم که به طور کلی افزایش تعداد سلول منجر به کاهش دقت شده است و بهترین نتایج روی مجموعه ارزیابی با یک لایه هشت سلولی رخ داده است. این موضوع به این دلیل است که تعداد داده‌های مسئله بسیار کم است و تعداد ویژگی‌ها به نسبت زیاد. در این شرایط یک شبکه ساده بهتر می‌تواند آموزش ببینید و بیش‌برازش نشود. در این شرایط و برای یک مدل پیچیده اگر کالبد Early Stopping استفاده نشود، مدل دقت بالا روی مجموعه آموزش و پایین روی مجموعه ارزیابی خواهد داشت و اگر مطابق رویکرد فعلی استفاده شود، جلوی آموزش بیشتر مدل گرفته می‌شود ولی خاتمه زودهنگام هم نمی‌تواند تمام مشکلات را حل کند.

در مقایسه بین دقت بهترین تنظیم LSTM (۶۰/۴٪) و بهترین تنظیم GRU (۵۵/۲٪) مشاهده می‌شود که شبکه LSTM نتایج بهتری داشته است؛ نه تنها بهترین تنظیم که سایر تنظیم‌های LSTM هم به طور کلی از GRU بهتر بوده است. باتوجه به تعداد کم داده من پیش از آزمایش انتظار داشتم GRU بهتر باشد، ولی بهتر بودن LSTM هم بی‌دلیل نیست؛ یکی از مزایای GRU نسبت به LSTM سرعت پردازشی بهتر آن است که اینجا مدنظر نیست و مزیت دیگر آن که عیب آن هم هست، ساده بودن مدل است؛ در اینجا احتمالاً روابط پیچیدگی بالایی دارند و به این واسطه LSTM در استخراج این‌ها بهتر عمل کرده است.

نکته‌ای که باید به آن توجه کرد این است که دقت تمام مدل‌ها با هر تنظیمی اصلاً قابل قبول نیست! باتوجه به آنکه ۵۶٪ داده‌ها دارای برچسب یک هستند، یک

مدل پایه بدون هیچ آموزشی می‌تواند به این دقت برسد. در میان مدل‌های موجود، تعداد مدل کمی به این دقت رسیده است؛ این نشان می‌دهد که پیشبینی من خوب نیست. به نظر می‌رسد که کم بودن داده و پیچیده بودن ذاتی حوزه کاربرد که داده‌های مالی باشد علت اصلی پایین بودن دقت است. تلاش‌های ناموفق زیادی برای بهتر شدن دقت انجام شده است که مهم‌ترین آن‌ها عبارت است از:

- تغییر تعداد سلول، تعداد لایه و نوع سلول (مطابق درخواست سوال)
- استفاده از لایه Dropout در انتها و در میان لایه‌ها با مقادیر مختلف
- استفاده از بهینه‌سازهای مختلف
- استفاده از مقادیر یادگیری مختلف در بهینه‌ساز
- مشابه کردن فراوانی برجسب‌های هدف در هر سه مجموعه داده
- تغییر تعداد روز پیشین در هر داده برای پیش‌بینی با اطلاعات قبلی بیشتر یا کمتر
- استفاده از تنها شاخص کل و حذف سایر شاخص‌ها برای سادگی مدل

لازم به ذکر است که امکان ساخت مجموعه داده‌ی مختلف به گونه‌ای که مدل بتواند روی آن دقت ۷۰٪ بگیرد امکان‌پذیر است! ولی باتوجه به اینکه در بیشتر مجموعه‌های داده چنین دقتی بدست نیامد، من نیز آن دقت‌ها را گزارش نکردم و دقت را روی یک مجموعه داده مشابه سایر مجموعه داده‌ها گزارش دادم. نهایتاً آنکه به واسطه اجرای چندباره شاید می‌توانستیم دقت‌های خوب تصادفی بگیریم ولی من برای نتایج عادلانه میانگین پنج بار اجرا را در نظر گرفتم.

### سوال ۳

نتایج شبکه LSTM در جدول زیر آورده شده است:

تعداد لایه‌های بازگشتی	تعداد سلول	دقت آموزشی مجموعه	دقت اعتبارسنجی مجموعه	دقت ارزیابی مجموعه
۲	۸	۶۰/۰٪	۵۶/۵٪	۵۷/۰٪
	۳۲	۵۸/۰٪	۵۸/۳٪	۵۶/۵٪
	۶۴	۵۶/۸٪	۵۸/۳٪	۵۳/۵٪
	۲۵۶	۵۸/۱٪	۵۶/۵٪	۵۶/۵٪
۳	۸	۵۶/۲٪	۵۶/۵٪	۵۳/۵٪
	۳۲	۵۹/۶٪	۵۶/۵٪	۵۶/۱٪
	۶۴	۵۶/۸٪	۵۷/۴٪	۵۵/۲٪
	۲۵۶	۵۸/۰٪	۵۷/۴٪	۵۶/۵٪
۴	۸	۵۷/۵٪	۵۶/۵٪	۵۴/۸٪
	۳۲	۵۷/۶٪	۵۶/۵٪	۵۶/۵٪
	۶۴	۵۷/۵٪	۵۷/۴٪	۵۵/۲٪
	۲۵۶	۵۷/۱٪	۵۶/۵٪	۵۷/۰٪

نتایج مربوط به شبکه GRU نیز بدین شکل است:

تعداد لایه‌های بازگشتی	تعداد سلول	دقت آموزشی مجموعه	دقت اعتبارسنجی مجموعه	دقت ارزیابی مجموعه
۲	۸	۵۴/۴٪	۵۸/۳٪	۵۲/۲٪
	۳۲	۵۵/۲٪	۵۹/۱٪	۵۲/۶٪
	۶۴	۵۸/۰٪	۵۶/۵٪	۵۳/۵٪
	۲۵۶	۶۰/۳٪	۵۳/۹٪	۵۴/۳٪
۳	۸	۵۵/۳٪	۵۸/۳٪	۵۴/۳٪
	۳۲	۵۷/۸٪	۵۷/۴٪	۵۶/۵٪
	۶۴	۵۸/۰٪	۵۸/۳٪	۵۷/۰٪
	۲۵۶	۵۷/۶٪	۵۳/۰٪	۵۴/۸٪



۵۶/۵٪	۵۶/۵٪	۵۷/۶٪	۸	۴
۵۴/۸٪	۵۷/۴٪	۵۷/۶٪	۳۲	
۵۵/۲٪	۵۳/۹٪	۵۵/۴٪	۶۴	
۵۴/۸٪	۵۳/۹٪	۵۹/۴٪	۲۵۶	

اولین چیزی که می‌توان دید این است که هیچ مدلی و در هیچ تنظیمی نتیجه مناسبی نداشته است و خروجی همگی به هم شبیه است و تفاوت زیادی با حالت تصادفی ندارند. در نتیجه می‌توان گفت که پشته کردن لایه‌ها در پیاده‌سازی من و مجموعه داده موجود به طور کلی سودمند نیست. اگر دقیق‌تر بررسی کنیم، می‌توان دید که پشته کردن لایه‌های GRU کمی دقت را بهتر کرده است.

باتوجه به مشابه بودن نتایج چندان نمی‌توان نظر دقیقی داد ولی به نظر می‌رسد GRU را بیشتر می‌توان عمیق کرد؛ علت هم احتمالاً به دلیل سادگی شبکه است؛ چراکه پشته کردن باعث افزایش پیچیدگی مدل می‌شود و اگر هر لایه خود به اندازه کافی پیچیده باشد، آموزش دچار مشکل می‌شود.