

مدل سازی زبانی مبتنی بر شبکه های عصبی

محمد بحرانی^۱

چکیده

مدل های زبانی به انتساب احتمال به کلمات در دنباله های لغات می پردازند که از این احتمالات در پیش بینی کلمه بعدی در کاربردهای مختلف پردازش زبان استفاده می شود. ساده ترین مدل های زبانی یعنی مدل های π -تایی، وابستگی های دور میان کلمات را در نظر نمی گیرند. یکی از راه های رفع این مشکل، مدل سازی زبانی با شبکه عصبی است. از انواع مختلف شبکه های عصبی مانند شبکه های جلورو و شبکه های بازگشتی برای مدل سازی زبانی استفاده می شود. شبکه های عصبی از بردارهای حاصل از جاسازی کلمات در ورودی خود استفاده می کنند که قدرت تعمیم شبکه را در مدل سازی افزایش می دهد. در این پژوهش از نوع خاصی از شبکه های بازگشتی به نام شبکه های عصبی با حافظه کوتاه مدت ماندگار (LSTM) برای مدل سازی زبان فارسی استفاده شده است. در شبکه های LSTM، به جای نرون های معمولی از واحدهای حافظه استفاده می شود که مشکل محو شدن گرادیان را در شبکه های بازگشتی استاندارد حل می کند. شبکه های LSTM یک طرفه و دوطرفه با استفاده از یک پیکره متنی ۱۰۰ میلیون کلمه ای فارسی آموزش داده شده اند. این شبکه ها دارای ۳۰

۱. نویسنده مسئول: استادیار، دکترای مهندسی کامپیوتر، دانشکده آمار، ریاضی و رایانه، دانشگاه علامه طباطبائی
bahrani@atu.ac.ir

هزار نرون در لایه‌های ورودی و خروجی، به تعداد کلمات واژگان و دو لایه مخفی می‌باشند. مدل زبانی حاصل از این شبکه‌ها حدود ۵۷٪ کاهش در سرگشتگی را نسبت به مدل پایه سه‌تایی نشان داده‌اند.

کلمات کلیدی: مدل‌سازی زبانی، شبکه‌های عصبی، شبکه حافظه کوتاه‌مدت ماندگار

۱- مقدمه

نقش مدل زبانی انتساب احتمال به یک کلمه با فرض رخ دادن یک دنباله از لغات قبلی است. با استفاده از این احتمالات علاوه بر اینکه می‌توان به پیش‌بینی کلمه بعدی پرداخت، می‌توان به یک دنباله از لغات $W = w_1 w_2 \dots w_m$ نیز یک احتمال متناسب نمود. احتمالات حاصل از مدل زبانی در کاربردهایی مانند بازشناسی گفتار، ترجمه ماشینی، شناسایی نویسه‌های نوری، فهم زبان طبیعی و هر کاربرد دیگری که در آن به پیش‌بینی کلمه نیاز است، استفاده می‌شوند.

ساده‌ترین و پرکاربردترین نوع مدل‌های زبانی مدل‌های n -تایی^۱ هستند. این مدل‌ها با استفاده از قانون زنجیره‌ای و مدل زنجیره مارکوف، احتمال شرطی یک کلمه پس از دنباله‌ای از کلمات را محاسبه می‌کنند و بدین طریق کلمه بعدی را پیش‌بینی می‌نمایند. در این مدل‌ها احتمال رخداد یک لغت، تنها به شرط رخ دادن $n-1$ کلمه قبلی در نظر گرفته می‌شود و درواقع تاریخچه یک کلمه به $n-1$ کلمه قبلی محدود می‌شود (رابطه ۱).

$$P(W) = P(w_1 w_2 \dots w_m) \cong \prod_{i=1}^m P(w_i | w_{i-(n-1)} \dots w_{i-1}) \quad (1)$$

در حالت‌های خاص $n=2$ و $n=3$ به این مدل به ترتیب دوتایی^۲ و سه‌تایی^۳ گفته می‌شود. با افزایش n ، مدل n -تایی تاریخچه طولانی‌تر را در نظر می‌گیرد و در نتیجه از لحاظ تئوری احتمال دقیق‌تری به ما می‌دهد ولی در عمل با افزایش n ، تعداد پارامترهای مدل افزایش یافته و تخمین آن‌ها با مشکل مواجه می‌شود و علاوه بر اینکه تخمین این پارامترها نیازمند داشتن

1. N-gram
2. Bigram
3. Trigram

داده متنی بسیار بزرگی است، از لحاظ محاسباتی نیز عملاً بسیار پرهزینه بوده و غیرممکن می‌باشد؛ بنابراین در بسیاری از کاربردها، n به ۳ یا حداکثر ۴ محدود می‌شود؛ بنابراین عیب مدل‌های n -تایی این است که برای پیش‌بینی کلمات بعدی فقط از اطلاعات محلی (چند کلمه قبلی) استفاده می‌کنند و وابستگی دور میان کلمات را در نظر نمی‌گیرند. در عمل، زبان طبیعی دارای ساختارهای پیچیده‌تری (مانند روابط دستوری و معنایی) است که توسط مدل‌های n -تایی پوشش داده نمی‌شوند.

برای در نظر گرفتن وابستگی‌های دور میان کلمات، مدل‌های زبانی مختلفی پیشنهاد شده‌اند، مانند مدل‌های نحوی، مدل‌های معنایی و مدل‌های موضوع پنهان که هر کدام تنها بخشی از ساختارهای پیچیده موجود در زبان طبیعی را مدل‌سازی می‌کنند؛ اما یکی از روش‌هایی که می‌تواند با الهام از ساختار و نحوه یادگیری ذهن انسان، به نوعی همه ساختارهای زبان را با هم مدل‌سازی کند، استفاده از شبکه‌های عصبی مصنوعی است. در سال‌های اخیر با افزایش قدرت ماشین‌های محاسبه گر، استفاده از شبکه‌های عصبی مصنوعی، به خصوص شبکه‌های عمیق در کاربردهای مختلف پردازش زبان و شناسایی الگو افزایش چشم‌گیری یافته و به نتایج خوبی رسیده است.

در ادامه ابتدا به نحوه به کارگیری شبکه‌های عصبی در مدل‌سازی زبانی می‌پردازیم و سپس نتایج مدل‌سازی زبانی فارسی با استفاده از این روش را ذکر می‌کنیم.

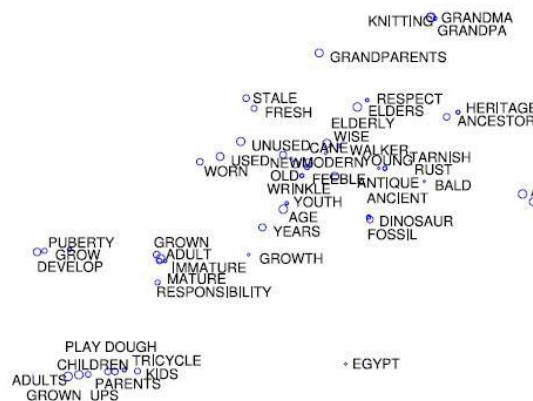
۲- مدل‌های زبانی مبتنی بر شبکه‌های عصبی

در مدل‌های زبانی n -تایی، کلمات به عنوان متغیرهای مجزا از هم هستند که رسیدن به یک تعمیم را بسیار دشوار می‌کند (آریسوی و همکاران، ۲۰۱۲). در خیلی از مواقع، توالی کلمات جدیدی در مرحله استفاده از مدل زبانی (مرحله آزمون) دیده می‌شود که در مرحله آموزش در داده‌ها دقیقاً مشاهده نشده ولی توالی مذکور از لحاظ ساختاری و معنایی با توالی‌هایی در داده آموزش مشابهت دارد. به عنوان مثال دو جمله زیر را در نظر بگیرید:

- مطالعات نشان می‌دهد که استفاده از رادیو در بزرگ سالان در حال افزایش است.

- مطالعات نشان می‌دهد که استفاده از رسانه در بزرگسالان در حال رشد است.

در این جملات، کلمات «رادیو» و «رسانه» و همچنین کلمات «افزایش» و «رشد» در ارتباط معنایی و دستوری با یکدیگر هستند. اگر جمله اول در داده آموزش و جمله دوم در داده آزمون رخ دهد، مدل زبانی n-تایی قادر به پیش‌بینی درست (انتساب یک احتمال بزرگ) به کلمات «رسانه» و «رشد» نیست. درواقع این مدل‌ها از لحاظ معنایی و مفهومی قابلیت تعمیم ندارند. شبکه‌های عصبی مصنوعی، برای حل این مشکل، از بردارهایی برای نمایش کلمات استفاده می‌کنند که به روش جاسازی کلمات^۱ به‌دست آمده‌اند. در جاسازی کلمات، هر کلمه در مجموعه واژگان، به یک فضای برداری با بعد کم نگاشت می‌شود که در آن فضا، کلمات با معانی نزدیک به هم در مجاورت یکدیگر قرار می‌گیرند (شکل ۱). این بردارها درواقع نقش بردار ویژگی را برای کلمات بازی می‌کنند (میکولوف و همکاران، ۲۰۱۳).



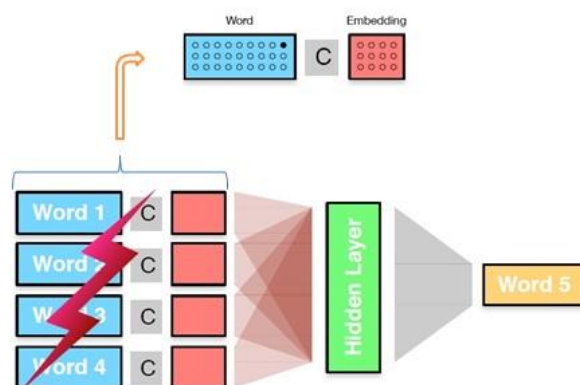
شکل ۱- جاسازی کلمات

در مرحله بعد، این بردارهای ویژگی به عنوان ورودی به شبکه های عصبی مصنوعی داده می شوند تا توالی واژگان دیده شده در داده آموزش، توسط شبکه یاد گرفته شود. برای این کار از روش ها و شبکه های عصبی مختلفی استفاده می شود که دو نوع کلی آن ها شبکه های

عصبی جلورو^۱ و شبکه‌های عصبی بازگشتی^۲ می‌باشند (ساندرمیر و همکاران، ۲۰۱۵).

۱-۲- مدل شبکه عصبی جلورو

مدل‌های زبانی مبتنی بر شبکه‌های عصبی جلورو درواقع شبیه مدل‌های زبانی π -تایی عمل می‌کنند، با این تفاوت که در مدل‌های زبانی عصبی، به دلیل استفاده از بردارهای حاصل از جاسازی کلمات به عنوان بردارهای بازنمایی کلمه، روابط معنایی و دستوری بین کلمات تا حد زیادی توسط شبکه یاد گرفته می‌شود و با اضافه شدن توان شبکه‌های عصبی در شناسایی الگوهای مشابه و نگاشت آن‌ها به یک مکان مشابه در یک فضای پیوسته، در عملکرد مدل‌زبانی بهبود زیادی حاصل خواهد شد (شکل ۲).



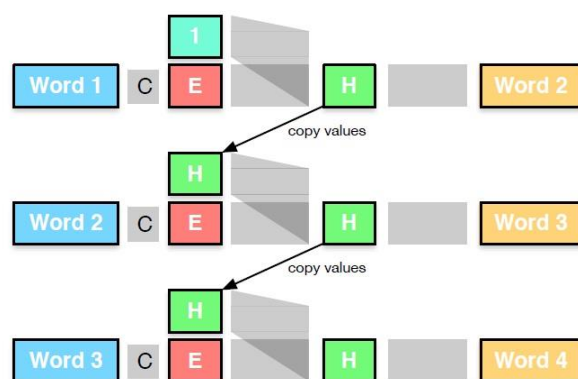
شکل ۲- معماری یک مدل مرتبه پنجم با استفاده از یک شبکه عصبی جلورو

همان‌طوری که در شکل ۲ مشاهده می‌شود، برای به دست آوردن یک مدل زبانی مرتبه پنجم با استفاده از یک شبکه عصبی جلورو، چهار کلمه قبل به عنوان پنجره ورودی به شبکه داده می‌شود و کلمه پنجم به عنوان خروجی از شبکه گرفته می‌شود.

-
1. Feed-forward Neural Networks
 2. Recurrent Neural Networks

۲-۲- مدل شبکه عصبی بازگشتی

در شبکه‌های عصبی بازگشتی، خروجی شبکه در گام‌های قبلی، به‌عنوان ورودی گام بعدی مورد استفاده قرار می‌گیرد. این نوع شبکه‌های عصبی در یادگیری توالی‌ها عملکرد خوبی دارند، بدین صورت که اطلاعات مربوط به گام‌های قبل را ذخیره کرده و در گام‌های بعد از آن استفاده می‌کنند. بدین ترتیب، شبکه‌های عصبی بازگشتی مشکل بافت ثابت شبکه‌های عصبی جلورو را حل می‌کنند (میکولوف و همکاران، ۲۰۱۰). همان‌طور که گفتیم، در شبکه‌های عصبی جلورو، یک بافت ثابت را که در مدل‌های n -تایی معادل $n-1$ کلمه قبل است، به‌عنوان ورودی به شبکه می‌دهیم و کلمه بعد را به‌عنوان خروجی دریافت می‌کنیم؛ در نتیجه، بافتی که از ابتدا تعیین می‌کنیم تا آخر ثابت خواهد ماند؛ اما شبکه‌های عصبی بازگشتی از تمامی گام‌های قبل به‌عنوان بافت برای پیش‌بینی کلمه بعد استفاده می‌کند (شکل ۳).



شکل ۳- ساختار یک مدل‌زبانی با استفاده از یک شبکه عصبی بازگشتی

در شکل ۳، E به جاسازی کلمات اشاره دارد و H محل ذخیره اطلاعات گام‌های قبل است که در هر مرحله از آن استفاده می‌شود. شبکه‌های عصبی بازگشتی می‌توانند یک توالی با هر طولی را پردازش کنند؛ بنابراین، این نوع شبکه‌های عصبی مصنوعی می‌توانند برای کار مدل‌سازی زبان که نوعی یادگیری توالی‌هاست مناسب باشند و عملکرد مدل زبانی را به

شکل چشمگیری افزایش دهند (میکولوف و همکاران، ۲۰۱۰).

۳-۲- شبکه عصبی با حافظه کوتاه‌مدت ماندگار (LSTM)

شبکه عصبی با حافظه کوتاه‌مدت ماندگار^۱ (LSTM)، نوع خاصی از شبکه‌های عصبی بازگشتی است که در سلول‌های آن از واحدهای حافظه استفاده شده است. در پژوهش حاضر، شبکه عصبی بازگشتی LSTM به صورت یک طرفه و دوطرفه برای مدل‌سازی زبان فارسی به کار رفته است.

یکی از مشکلات شبکه‌های عصبی بازگشتی، مشکل محوشدن گرادیان است. بدین معنی که اثر ورودی شبکه در لایه‌های مخفی و در نتیجه در خروجی شبکه به تدریج بسیار کوچک یا محو می‌شود. شبکه‌های LSTM با استفاده از واحدهای حافظه این مشکل را حل می‌کنند. هر واحد حافظه شامل یک یا چند سلول حافظه متصل به هم و سه واحد تکثیرپذیر است که دروازه‌های ورودی، خروجی و فراموشی نامیده می‌شوند. این دروازه‌ها عملیاتی مانند نوشتن، خواندن و مرتب‌سازی مجدد سلول‌ها را فراهم می‌کنند. یک شبکه عصبی LSTM مانند یک شبکه بازگشتی استاندارد است با این تفاوت که واحدهای بازگشتی در لایه مخفی با واحدهای حافظه جایگزین شده‌اند.

۳- نتایج مدل‌سازی زبانی با شبکه‌های عصبی

در این پژوهش از شبکه عصبی بازگشتی LSTM به صورت یک طرفه و دوطرفه برای مدل‌سازی زبان فارسی استفاده شده است. به منظور آموزش و آزمون شبکه از پیکره متنی پرسیکا (اقبال‌زاده و همکاران، ۲۰۱۲) استفاده شده و یک زیرمجموعه ۱۰۰ میلیون کلمه‌ای از این پیکره برای آموزش و آزمون انتخاب گشته است. داده آموزش شامل ۹۰ میلیون کلمه از این دادگان و داده‌های آزمون و اعتبارسنجی نیز هر کدام شامل ۵ میلیون کلمه می‌باشند. از ۳۰ هزار کلمه پرکاربرد بخش آموزش به عنوان مجموعه واژگان استفاده شده و سایر کلمات در داده آموزش با نماد UNK جایگزین شده‌اند.

1. Long Short-Term Memory

با توجه به اندازه مجموعه واژگان، تعداد نرون‌ها در لایه ورودی و لایه خروجی شبکه ۳۰ هزار نرون بوده که نمایش کلمات را به صورت one-hot انجام می‌دهند. لایه مخفی درواقع لایه جاسازی کلمات است که بردار ویژگی کلمه در آن لایه به صورت خودکار یاد گرفته می‌شود. در آزمایش‌های ما تعداد لایه‌ای مخفی، دولایه بوده و اندازه این لایه‌ها ۵۰۰ می‌باشد. در مرحله آموزش، نمایش one-hot یک کلمه در ورودی و نمایش one-hot کلمه بعدی در خروجی قرار داده می‌شود که این خروجی به نوبه خود ورودی مرحله بعد را تأمین می‌کند و بدین ترتیب طی چند تکرار شبکه آموزش داده می‌شود (ساندرمیر و همکاران، ۲۰۱۲). تابع فعال‌سازی در لایه خروجی، تابع softmax می‌باشد. در مرحله آزمون، یک کلمه در ورودی قرار می‌گیرد و به ازای آن، نرون‌های لایه خروجی فعال می‌شوند. میزان فعالیت هر کدام از نرون‌های لایه خروجی، احتمال پیش‌بینی شده برای کلمه مربوطه را مشخص می‌سازند. از این احتمالات به عنوان احتمالات مدل زبانی استفاده می‌شود. به منظور ارزیابی مدل زبانی حاصل از شبکه عصبی، به مانند مدل‌های زبانی دیگر از معیار سرگشتگی استفاده می‌شود. جدول زیر میزان سرگشتگی مدل‌های حاصل از شبکه عصبی LSTM یک‌طرفه و دوطرفه را بر روی داده آزمون نشان می‌دهد.

جدول ۱- سرگشتگی حاصل از مدل‌های زبانی عصبی در مقایسه با مدل n-تایی

سرگشتگی	مدل زبانی
۱۳۸	مدل سه تایی (مدل پایه)
۶۲	شبکه LSTM یک‌طرفه
۵۹	شبکه LSTM دوطرفه

همان‌طور که مشاهده می‌شود، چنان که انتظار می‌رفت، شبکه عصبی نسبت به مدل سه تایی، بهبود قابل توجهی را در سرگشتگی نتیجه داده است. همچنین شبکه LSTM دوطرفه که برای مدل‌سازی زبانی، علاوه بر بافت پیشین، بافت پسین را نیز در نظر می‌گیرد، نتیجه بهتری نسبت به شبکه یک‌طرفه داده است.

۴- جمع‌بندی و نتیجه‌گیری

شبکه‌های عصبی به صورت‌های مختلف برای مدل‌سازی زبانی به کار رفته‌اند. از میان انواع مختلف شبکه‌های عصبی که بدین منظور استفاده شده‌اند می‌توان به شبکه‌های جلورو، شبکه‌های بازگشتی استاندارد و شبکه‌های بازگشتی LSTM اشاره کرد. با توجه به جدید بودن پژوهش‌ها در زمینه استفاده از شبکه‌های عصبی LSTM در مدل‌سازی زبانی، در این پژوهش به مدل‌سازی زبان فارسی با این شبکه‌ها پرداختیم و از شبکه‌های LSTM یک‌طرفه و دوطرفه بدین منظور استفاده کردیم. استفاده از این نوع شبکه‌ها برای مدل‌سازی زبانی در زبان فارسی کاری نسبتاً نو به شمار می‌رود. در این زمینه در زبان فارسی کارهای اندکی انجام گرفته که همگی بر روی داده‌های کوچک‌تری بوده‌اند. همان‌طور که انتظار می‌رفت شبکه‌های عصبی نسبت به مدل پایه سه‌تایی نتیجه بسیار بهتری در سرگشتگی داده است. کاهش نسبی سرگشتگی در بهترین حالت حدود ۵۷٪ می‌باشد. در آینده برآنیم که از مدل‌های جدیدتر شبکه عصبی و همچنین از داده‌های زبانی بیشتری به منظور مدل‌سازی زبانی استفاده کنیم.

مراجع

- Arisoy, E. Sainath, T. N. Kingsbury, B. & Ramabhadran, B. (2012). "Deep neural network language models", In *Proceedings of the NAACL-HLT 2012 Workshop*.
- Eghbalzadeh, H. Hosseini, B. Khadivi, S. & Khodabakhsh, A. (2012). "Persica: A Persian corpus for multi-purpose text mining and Natural language processing", In *Telecommunications (IST), 2012 Sixth International Symposium on*, pp. 1207-1214.
- Mikolov, T. Chen, K. Corrado, G. & Dean, J. (2013). "Efficient estimation of word representations in vector space", *arXiv preprint*, arXiv:1301.3781.
- Mikolov, T. Karafiát, M. Burget, L. Cernocký, J. & Khudanpur, S. (2010). "Recurrent neural network based language model", In *Interspeech 2010*.

- Sundermeyer, M. Ney, H. & Schlüter, R. (2015). "From feedforward to recurrent LSTM neural networks for language modeling", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(3), pp. 517-529.
- Sundermeyer, M. Schlüter, R. & Ney, H. (2012). "LSTM neural networks for language modeling", In *Thirteenth Annual Conference of the International Speech Communication Association*.