دانشکده مهندسی کامپیوتر

بررسی مقاله یادگیری توالی‌به‌توالی با شبکه‌های عصبی

گزارش پروژه درس پردازش زبان‌های طبیعی

فاز سوم (نهایی)

دانشجو:

مرتضی ذاکری نصرآبادی

استاد:

دکتر بهروز مینایی

بهمن ۱۳۹۶


بسم الله الرحمن الرحیم

چکیده

یادگیری ژرف شاخه‌ای نسبتا جدید از یادگیری ماشین است که در آن توابع محاسباتی به شکل گراف‌های چند سطحی یا ژرف برای شناسایی و تخمین قانون حاکم بر حل یک مسئله پیچیده به‌کار بسته می‌شوند. شبکه‌های عصبی ژرف ابزاری برای طراحی و پیاده‌سازی این مدل یادگیری هستند. این شبکه‌ها در بسیاری از وظایف یادگیری ماشینی سخت، موفق ظاهر شده‌اند. به‌منظور استفاده از شبکه‌های ژرف در وظایفی که ترتیب ورودی داده در انجام آن مؤثر است مانند اکثر وظایف حوزه پردازش زبان طبیعی، شبکه‌های عصبی مکرر ابداع گشتند که بازنمایی مناسبی از مدل‌های زبانی ارایه می‌دهند. این مدل‌ها در حالت ساده برای همه وظیفه‌های یک مدل زبانی مناسب نیستند. در این گزارش مدل خاصی از شبکه‌های مکرر تحت عنوان مدل توالی‌به‌توالی یا کدگذار-گدگشا بررسی می‌شود که برای وظایفی که شامل توالی‌های ورودی و خروجی با طول متفاوت هستند؛ نظیر ترجمه ماشینی، توسعه داده شده و توانسته است نتایج قابل قبولی را در این زمینه تولید کند.

کلیدواژه‌ها: مدل توالی‌به‌توالی، شبکه عصبی مکرر، یادگیری ژرف، ترجمه ماشینی.

فهرست مــطالب

عنوان | صــفحه

فهرست شکل‌ها

عنوان **صـفحه**

فهرست جدول‌ها

عــنوان **صــفحه**

جدول واژگان و نمادهای اخــتصاری

مفهوم کوته‌نوشت	کوته‌نوشت
Convolutional Neural Networks	CNN
Deep Neural Network	DNN
Language Model	LM
Long-short term memory	LSTM
Natural Language Processing	NLP
Neural Language Models	NLM
Neural Machine Translation	NMT
Rectified Linear Unit	ReLU
Recurrent Neural Network	RNN
Statistical Machine Translation	SMT

ما ممکن است امیدوار باشیم که ماشین‌ها در نهایت در همه زمینه‌های هوشمند با انسان رقابت خواهند کرد. اما بهترین زمینه برای شروع کدام است؟!

آلِن تورینگ

۱ مقدمه

مدل‌ها و روش‌های یادگیری به‌کمک شبکه‌های عصبی ژرف (DNNs)[۱] اخیرا، با افزایش قدرت محاسباتی سخت‌افزارها و نیز حل برخی از چالش‌های اساسی موجود بر سر راه آموزش و یادگیری این شبکه‌ها، بسیار مورد توجه واقع شده‌اند. DNNها در انجام وظایف سخت یادگیری ماشین مانند تشخیص گفتار، تشخیص اشیاء و غیره، فوق‌العاده قدرت‌مند ظاهر شده‌اند و در مواردی روش‌های سنتی را کاملا کنار زده‌اند. قدرت بازنمایی زیاد DNNها به این دلیل است که قادر هستند محاسبات زیادی را به صورت موازی در چندین لایه انجام داده، با تعداد زیادی پارامتر پاسخ مسئله داده شده را تخمین زده و مدل مناسبی از آن ارایه دهند. درحال حاضر DNNهای بزرگ می‌توانند با استفاده از الگوریتم پس‌انتشار[۲] به‌صورت بانظارت[۳] روی یک مجموعه آموزش برچسب‌زده و به‌قدر کافی بزرگ آموزش ببینند. بنابراین در مواردی که ضابطه حاکم بر یک مسئله دارای پارامترهای بسیار زیادی است و یک مقدار بهینه از این پارامترها وجود دارد (صرفا با استناد به این که مغز انسان همین مسئله را خیلی سریع حل می‌کند)، روش یادگیری پس‌انتشار این تنظیم از پارامترها (مقدارهای بهینه) را یافته و مسئله را حل می‌کند [1].

---

[۱] deep neural networks

[۲] backpropagation

[۳] supervised

بسیاری از وظایف یادگیری ماشین به حوزه پردازش زبان طبیعی (NLP)[۴] مربوط می‌شوند؛ جایی که در آن معمولا ترتیب ورودی‌ها و خروجی‌های یک مسئله مهم است. برای مثال در ترجمه ماشینی دو جمله با واژه‌های یکسان ولی ترتیب متفاوت، معانی (خروجی‌های) مختلفی دارند. این وظایف اصطلاحا مبتنی بر توالی[۵] هستند. در واقع ورودی آنها به صورت یک توالی است. شبکه‌های عصبی رو به جلو ژرف[۶] برای این دسته از وظایف خوب عمل نمی‌کنند؛ چرا که قابلیتی برای به‌خاطر سپاری و مدل‌سازی ترتیب در آنها تعبیه نشده است.

شبکه‌های عصبی مکرر (RNNs)[۷] خانواده‌ای از شبکه‌های عصبی برای پردازش وظایف مبتنی بر توالی هستند. همانطور که شبکه‌های عصبی پیچشی (CNNs)[۸]، ویژه پردازش یک تور[۹] از مقادیر، برای مثال یک تصویر، طراحی شده‌اند؛ یک RNN نیز همسو با پردازش یک توالی از مقادیر ورودی $x = < x^{(1)}, x^{(2)}, \dots, x^{(n)} >$ ساخته شده است [2]. خروجی RNNها نیز مانند ورودی آنها در اغلب وظایف یک توالی است. این قابلیت پردازش توالی توسط شبکه‌های عصبی، آنها را برای استفاده در وظایف NLP، بسیار درخور ساخته است.

۱-۱- شرح مسئله و اهمیت موضوع

برخلاف انعطاف پذیری و قدرت بالای RNNها، در حالت ساده این شبکه‌ها یک توالی ورودی با طول ثابت را به یک توالی خروجی با همان طول نگاشت می‌کنند. این موضوع اما یک محدودیت جدی است؛ زیرا، بسیاری از مسائل مهم، در قالب توالی‌هایی که طولشان از قبل مشخص نیست، به بهترین شکل قابل بیان هستند و در نظر گرفتن یک طول ثابت از پیش تعیین شده برای ورودی و خروجی به خوبی مسئله را مدل نمی‌کند. برای مثال ترجمه ماشینی (MT)[۱۰] و تشخیص

[۴] natural language processing
[۵] sequence
[۶] deep feed-forward neural networks
[۷] recurrent neural networks
[۸] convolutional neural networks
[۹] grid
[۱۰] machine translation

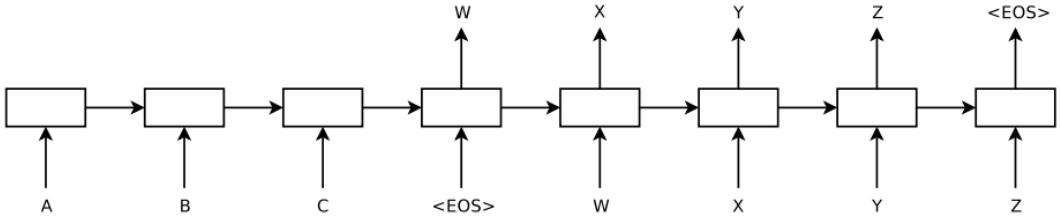
گفتار[11] مسائلی از این دست هستند. همچنین سیستم پرسش و پاسخ را نیز می‌توان به صورت نگاشت یک توالی از واژه‌ها به‌عنوان پرسش، به یک توالی دیگر از واژه‌ها به عنوان پاسخ، در نظر گرفت. بنابراین پُر واضح است که ایجاد یک روش مستقل از دامنه برای یادگـیری نگاشت توالی‌به‌تولی مفید و قابل توجیه خواهد بود [1].

۱-۲- اهداف و راهکارها

همانطور که دیدیم طیف وسیعی از وظایف NLP مبتنی بر نگاشت توالی‌های با طول نامشخص و متغیر به یکدیگر است. همچنین روش‌های سنتی مثل n-garm دارای محدودیت‌های خاص خود در حل این دسته مسائل هستند و استفاده از روش‌های یادگیری ژرف به وضوح امید بخش بوده است. بنابراین هدف ارایه یک مدل مبتنی بر RNNها جهت نگاشت توالی‌به‌توالی است. در این گـزارش راهکار مطرح شده در [1] و نتایج آن به‌تفصیل شرح داده می‌شود.

Sutskever و همکاران [1] نشان دادند که چگونه یک کاربرد ساده از شبکه با معماری حافظه کوتاه‌مدت بلند (LSTM)[12] می‌تواند مسائل نگاشت توالی‌به‌توالی را حل کند. ایده اصلی استفاده از یک LSTM برای خواندن توالی ورودی، به‌صورت یک نمونه در هر مرحله زمانی، جهت اقتباس برداری بزرگ با بعد ثابت و سپس استفاده از یک LSTM دیگر برای استخراج توالی خروجی از آن بردار است. LSTM دوم دقیقا یک مدل زبانی مبتنی بر RNN است با این تفاوت که حاوی احتمال شرطی نسبت به توالی ورودی نیز هست. قابلیت LSTM در یادگیری موفق وابستگی‌های مکانی طولانی مدت نهفته درون توالی‌ها، آن را برای استفاده در مدل پیشنهادی مناسب ساخته است. شکل (۱) یک طرح‌واره از این مدل را به صورت عام نشان می‌دهد.

[11] speech recognition
[12] long-short term memory



شکل (۱) یک طرح‌واره از مدل توالی‌به‌توالی متشکل از دو RNN. این مدل توالی ABC را به‌عنوان ورودی خوانده و توالی WXYZ را به‌عنوان خروجی تولید می‌کند. مدل پس از تولید نشانه <EOS> روند پیش‌بینی خود را متوقف می‌کند [1].

۱-۳- داده‌ها و نتایج

مدل پیشنهادی در بخش قبل، برروی وظیفه ترجمه ماشینی عصبی (NMT)[۱۳] مورد آزمایش قرار گرفته است. برای انجام آزمایش‌ها از مجموعه داده ترجمه انگلیسی به فرانسوی WMT'14 استفاده شده است [3]. همچنین مجموعه داده کوچکتری در [4] وجود دارد که برای آموزش مدل‌های کوچکتر مناسب است. این مجموعه شامل ترجمه‌های انگلیسی به فارسی نیز هست.

نتایج حاصل شده از این کار بدین قرار است. بر روی مجموعه داده WMT'14 با استخراج مستقیم ترجمه از پنج LSTM ژرف با ۳۸۰ میلیون پارامتر، در نهایت امتیاز BLEU معادل ۳۴٫۸۱ کسب گردیده است. این امتیاز بالاترین امتیازی است که تا زمان ارایه این مقاله از طریق NMT حاصل شده است. به‌عنوان مقایسه امتیاز BLEU برای ترجمه ماشینی آماری (SMT)[۱۴] برروی همین مجموعه داده برابر ۳۳٫۳۰ است. این درحالی است که امتیاز ۳۴٫۸۱ با احتساب اندازه واژه‌نامه ۸۰هزار کلمه به‌دست آمده و هرجا که کلمه ظاهر شده در ترجمه مرجع در واژه‌نامه نبوده این امتیاز جریمه شده است. بنابراین نتایج نشان می‌دهد که یک معماری مبتنی بر شبکه عصبی تقریبا غیر بهینه، که نقاط زیادی برای بهبود دارد، قادر است تا روش‌های سنتی مبتنی بر عبارتِ سیستم SMT را شکست دهد [1].

[۱۳] neural machine translation
[۱۴] statistical machine translation

۲ مفاهیم اولیه

در این قسمت پیرامون سه مفهوم اصلی گزارش پیشرو، یعنی مدل زبانی (LM)[۱۵]، شبکه‌های عصبی مکرر و ترجمه ماشینی عصبی، به‌صورت مختصر توضیحاتی ارایه می‌گردد.

۲-۱- مدل زبانی

LM یک مفهوم پایه در NLP است که امکان پیش‌بینی نشانه بعدی در یک توالی را فراهم می‌کند. به‌بیان دقیق‌تر LM عبارت است از یک توزیع احتمالی روی یک توالی از نشانه‌ها (اغلب واژه‌ها) که احتمال وقوع یک توالی داده شده را مشخص می‌کند. در نتیجه می‌توان بین چندین توالی داده شده برای مثال چند جمله، آن را که محتمل‌تر است، انتخاب کرد [5]. LM برای توالی $x = < x^{(1)}, x^{(2)}, \dots, x^{(n)} >$ عبارت است از:

$$p(x) = \prod_{t=1}^{n} p(x^{(t)} | x^{(<t)}) \qquad (۱)$$

مدل‌های سنتی n-gram برای غلبه بر چالش‌های محاسباتی، با استفاده از فرض مارکوف رابطه (۱) را به درنظر گرفتن تنها n-1 نشانه قبلی محدود می‌کنند. به‌همین دلیل برای توالی‌های طولانی (بیشتر از ۴ یا ۵ نشانه) و دیده نشده مناسب نیستند. مدل‌های زبانی عصبی (NLMs)[۱۶] که بر مبنای شبکه‌های عصبی عمل پیش‌بینی واژه بعدی را انجام می‌دهند، در ابتدا برای کمک به n-gramها با آنها ترکیب شدند که منجر به ایجاد پیچیدگی‌های زیادی شد؛ در حالی که مشکل توالی‌های طولانی همچنان وجود داشت [5]. اخیرا اما، معماری‌های جدیدی برای LM که کاملا بر اساس DNNها است، ایجاد شده‌اند. سنگ‌بنای این مجموعه معماری‌ها RNNها بوده که در بخش بعدی معرفی می‌شوند.

[۱۵] language model
[۱۶] neural language models

۲-۲- شبکه‌های عصبی مکرر

RNNها کلاسی از شبکه‌های عصبی هستند که به‌صورت یک ***گراف جهت‌دار دوری*** بیان می‌شوند. به‌عبارت دیگر ورودی هریک از لایه(های) پنهان یا خروجی علاوه بر خروجی لایه قبل، شامل ورودی از مرحله قبل به‌صورت بازخورد نیز می‌شود. شکل (۲) یک RNN را نشان می‌دهد. همانطور که پیداست، لایه پنهان از مراحل قبلی هم بازخورد می‌گیرد. در هر مرحله‌زمانی t از ($t=1$ تا $t=n$) یک بردار $x^{(t)}$ از توالی ورودی $x=<x^{(1)},x^{(2)},\dots,x^{(n)}>$ پردازش می‌شود. در حالت کلی معادله‌های بروزرسانی (گذرجلو[۱۷]) یک RNN در t عبارتند از [2]:

$$a^{(t)} = Ux^{(t)} + Wh^{(t-1)} + b, \quad (۲)$$

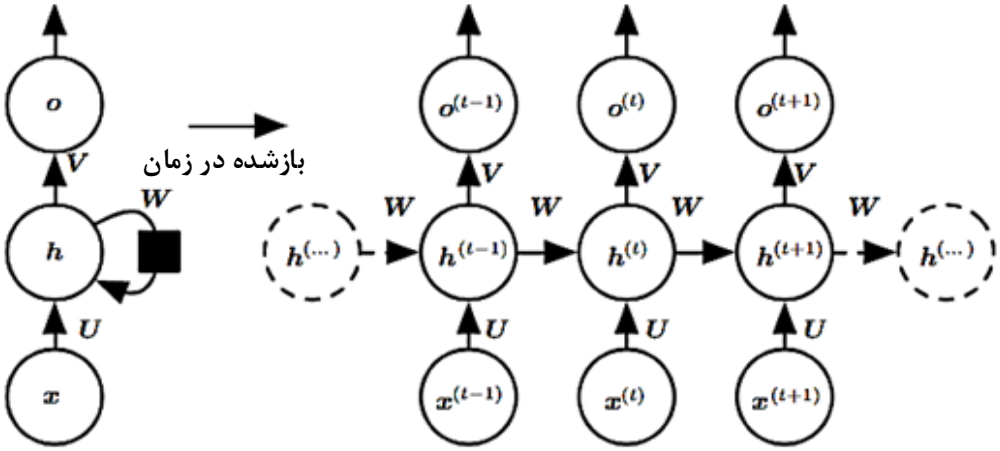
$$h^{(t)} = \Phi(a^{(t)}), \quad (۳)$$

$$o^{(t)} = Vh^{(t)} + c, \quad (۴)$$

$$\hat{y}^{(t)} = softmax\left(o^{(t)}\right), \quad (۵)$$

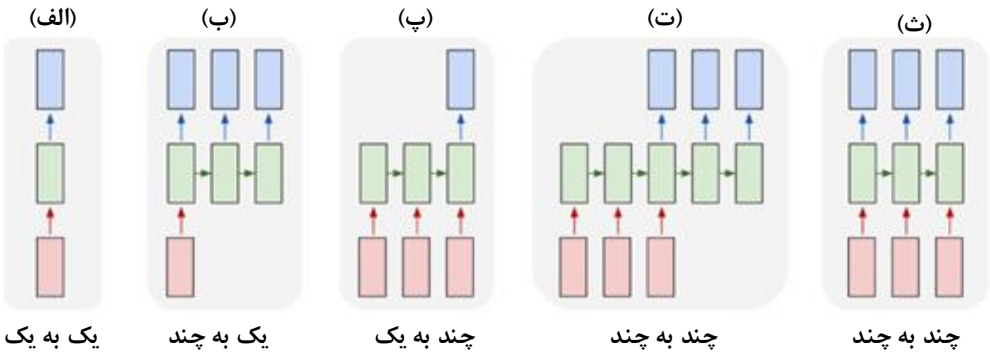
که در آن بردارهای b و c بایاس و ماتریس‌های U، V و W به‌ترتیب وزن یال‌های لایه ورودی به پنهان، پنهان به خروجی و پنهان به پنهان، تشکیل‌دهنده مجموعه پارامترهای شبکه هستند. Φ تابع انگیزش است که معمولا یکی از توابع ReLU[۱۸] یا سیگموید[۱۹] انتخاب می‌شود. لایه آخر را نیز تابع بیشینه هموار[۲۰] تشکیل می‌دهد که احتمال وقوع هر نشانه خروجی را مشخص می‌کند.

[۱۷] forward pass
[۱۸] rectified linear unit
[۱۹] sigmoid
[۲۰] softmax function



شکل (۲) گراف محاسباتی مربوط به یک نوع RNN که یک توالی ورودی از مقادیر x را به یک توالی خروجی از مقادیر o نگاشت می‌کند. فرض شده است که خروجی o احتمالات نرمال نشده است، بنابراین خروجی واقعی شبکه یعنی $\hat{y}$ از اعمال تابع بیشینه هموار روی o حاصل می‌شود. **چپ:** RNN به‌صورت یال بازگشتی. **راست:** همان شبکه به‌صورت باز شده در زمان، به‌نحوی که هر گره با یک برچسب زمانی مشخص شده است [2].

در شکل (۲)، RNN با یک لایه پنهان نشان داده شده است. اما می‌توان RNNژرف با چندین لایه پنهان نیز داشت. همچنین طول توالی‌های ورودی و خروجی می‌تواند بسته به مسئله مورد نظر متفاوت باشد. karpathy [6] RNNها را از منظر طول توالی ورودی و طول توالی خروجی به چند دسته تقسیم‌بندی کرده است. شکل (۳) این دسته‌بندی را نشان می‌دهد.



شکل (۳) طرح واره‌ای از حالت‌های مختلف RNN. **(الف):** شبکه عصبی استاندارد، **(ب):** شبکه یک به چند، **(پ):** شبکه چند به یک، **(ت)** و **(ث):** شبکه‌های چند به چند [6].

تصویر karpathy از حالت‌های مختلف RNN بعد از انتشار مقاله منتخب در این گزارش می‌باشد؛ با این حال در بخش ۴ خواهیم دید که چگونه می‌توان از ترکیب این طرح‌ها نیز برای ایده معماری توالی‌به‌توالی الهام گرفت.

۲-۳- ترجمه ماشینی عصبی

به‌طور کلی MT را می توان با یک LM که به جمله زبان مبدأ مشروط شده است، مدل‌سازی کرد. بر همین اساس NMT را می‌توان یک مدل زبانی مکرر در نظر گرفت که مستقیما احتمال شرطی $p(y|x)$ را در ترجمه جمله زبان مبدأ $x = < x^{(1)}, x^{(2)}, \dots, x^{(n)} >$ به جمله زبان مقصد $y = < y^{(1)}, y^{(2)}, \dots, y^{(m)} >$ مدل می‌کند. دقت شود که طول جمله مبدأ یعنی n و جمله مقصد یعنی m الزاما برابر نیست. بنابراین در NMT هدف محاسبه این احتمال و سپس استفاده از آن در تولید جمله به زبان مقصد، هر دو به کمک DNNها است [5].

۳ کارهای مرتبط

کارهای زیادی در زمینه NLMs انجام شده است. در بیشتر این کارها از شبکه‌های عصبی روبه‌جلو یا مکرر استفاده شده و کاربرد آن معمولا در یک وظیفه MT با امتیازدهی مجدد n فهرست بهتر[۲۱]، اعمال شده و نتایج آن معمولا نشان از بهبود امتیازهای قبلی داشته است [1].

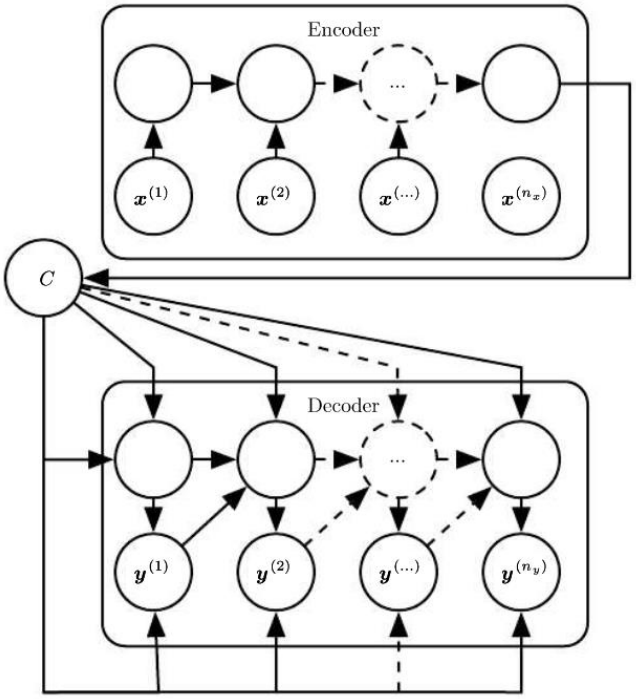
اخیرا کارهایی در زمینه فشردن اطلاعات زبان مبدأ در NLM انجام شده است. برای نمونه Auli و همکاران [7] NLM را با مدل عنوان[۲۲] جمله ورودی ترکیب کرده‌اند که نتایج بهبود بخشی داشته است. کار انجام شده در مقاله [1] به کار [8] بسیار نزدیک است. در مقاله [8] نویسندگان برای اولین بار توالی ورودی را در یک بردار فشرده کرده و سپس آن را به توالی خروجی تبدیل کردند. البته در این کار، برای تبدیل توالی به بردار، از CNNs استفاده شده که

[۲۱] n-best list
[۲۲] topic model

ترتیب واژه‌ها را حفظ نمی‌کند. چـــو و همکاران [9] یک معماری شبهِ LSTM را برای نگاشت توالی ورودی به بردار و سپس استخراج توالی خروجی و نهایتا ترکیب آن با SMT استفاده کرده‌اند. معماری آنها از دو RNN با عنوان‌های کدگذار[۲۳] و کدگشا[۲۴] تشکیل شده که RNN اول وظیفه تبدیل یک توالی با طول متغیر به یک بردار با طول ثابت را در قالب یک سلول زمینه c دارد و RNN دوم وظیفه تولید توالی خروجی را با لحاظ کردن c و نماد شروع جمله مقصد بر عهده دارد. معماری پیشنهادی آنها تحت عنوان کلی RNN کدگذار-کدگشا در شکل (۴) نشان داده شده است. چون آنها از LSTM استفاده نکرده و بیشتر تلاش خود را معطوف به ترکیب این روش با مدل‌های قبلی SMT کرده‌اند، برای توالی‌های ورودی و خروجی طولانی همچنان مشکل عدم حفظ حافظه وجود دارد. این معماری در [2] به صورت مختصر توضیح داده شده است.

Bahdanau و همکاران [10] یک روش ترجمه مستقیم با استفاده از شبکه عصبی پیشنهاد داده‌اند که از سازوکار *attention* برای غلبه بر کارآمدی ضعیف روش [9] روی جملات طولانی استفاده می‌کند و به نتایج مطلوبی هم دست یافتند. نتایج [1] برای نمونه با نتایج حاصل از کار آنها مقایسه شده است.

[۲۳] encoder
[۲۴] decoder



شکل (۴) مدل RNN کدگذار-کدگشا، که برای یادگیری تولید توالی خروجی $< y^{(1)}, \dots, y^{(n_y)} >$ از روی توالی ورودی $< x^{(1)}, \dots, x^{(n_x)} >$ به‌کار می‌رود [2].

۴ مدل توالی‌به‌توالی

در مدل توالی‌به‌توالی از دو RNN با واحدهای LSTM استفاده شده است. هدف LSTM در اینجا تخمین احتمال شرطی $p(< y^{(1)}, y^{(2)}, \dots, y^{(m)} > | < x^{(1)}, x^{(2)}, \dots, x^{(n)} >)$ است که قبلا هم دیده بودیم (بخش ۲-۳). LSTM این احتمال شرطی را ابتدا با اقتباس بازنمایی بعد ثابت v برای توالی ورودی $< x^{(1)}, x^{(2)}, \dots, x^{(n)} >$ از آخرین مقدار حالت پنهان و در ادامه با محاسبه احتمال $< y^{(1)}, y^{(2)}, \dots, y^{(m)} >$ از رابطه استاندارد مطرح در LM (رابطه (۱)) و درنظر گرفتن v برای حالت پنهان آغازین به‌صورت داده شده در رابطه زیر، حساب می‌کند:

$$p(< y^{(1)}, \dots, y^{(m)} > | < x^{(1)}, \dots, x^{(n)} >) = \prod_{t=1}^{m} p(y^{(t)} | v, y^{(1)}, \dots, y^{(t-1)}) \qquad (۶)$$

در رابطه (۶) هر توزیع احتمالی $p(y^{(t)}|v, y^{(1)}, \dots, y^{(t-1)})$ به‌وسیله یک تابع بیشینه هموار روی همه واژه‌های داخل واژه‌نامه بازنمایی می‌شود. برای LSTM از روابط [11] استفاده شده است. هر جمله در این مدل نیاز است تا با یک علامت خاص مثل <EOS> خاتمه یابد. این امر مدل را قادر می‌سازد تا بتواند توزیع احتمالی را روی توالی با هر طول دلخواهی تعریف کند. شمای کلی مدل در شکل (۱) نشان داده شده است. در این شکل LSTM بازنمایی توالی ورودی $\ll A', B', C', < EOS >$ را حساب و سپس از این بازنمایی برای محاسبه احتمال توالی خروجی $< W', X', Y', Z', < EOS >>$ استفاده می‌کند. در عین حال این مدل را می‌توان ترکیبی از قسمت‌های پ و ت شکل (۳) دانست.

مدل پیاده‌سازی شده در عمل از سه جنبه با مدل معرفی شده در بالا تفاوت دارد. اول، از دو LSTM جداگانه استفاده شده است: یکی برای توالی ورودی و دیگری برای توالی خروجی؛ زیرا، انجام این کار پارامترهای مدل را با هزینه محاسباتی اندکی، به تعداد بسیار زیادی افزایش می‌دهد. دوم اینکه LSTMهای ژرف به‌شکل قابل توجهی LSTMهای سطحی را شکست می‌دهند، به همین دلیل LSTM با ژرفای چهار لایه به‌کار گرفته شده است. سوم اینکه نویسندگان در این مقاله یافته‌اند که وارون کردن توالی ورودی در سرعتِ همگرایی آموزش شبکه و نیز دقت پیش‌بینی آن تأثیر شگرفی ایفا می‌کند. بنابراین به‌جای نگاشت مستقیم توالی $< a, b, c >$ به توالی$< \alpha, \beta, \gamma >$، LSTM برای نگاشت $< c, b, a >$ به $< \alpha, \beta, \gamma >$ آموزش داده می‌شود که در آن $\alpha\ \beta\ \gamma$ ترجمه متناظر با $a\ b\ c$ است. توجیه اثر این پدیده آن است که در نگاشت به روش وارون ابتدای عبارت‌ها که متناظر با یکدیگر هستند به‌هم نزدیک شده و این امر سبب زودتر همگرا شدن الگوریتم SGD و نزدیک شدن به مقادیر بهینه می‌شود [1].

۴-۲- آموزش شبکه

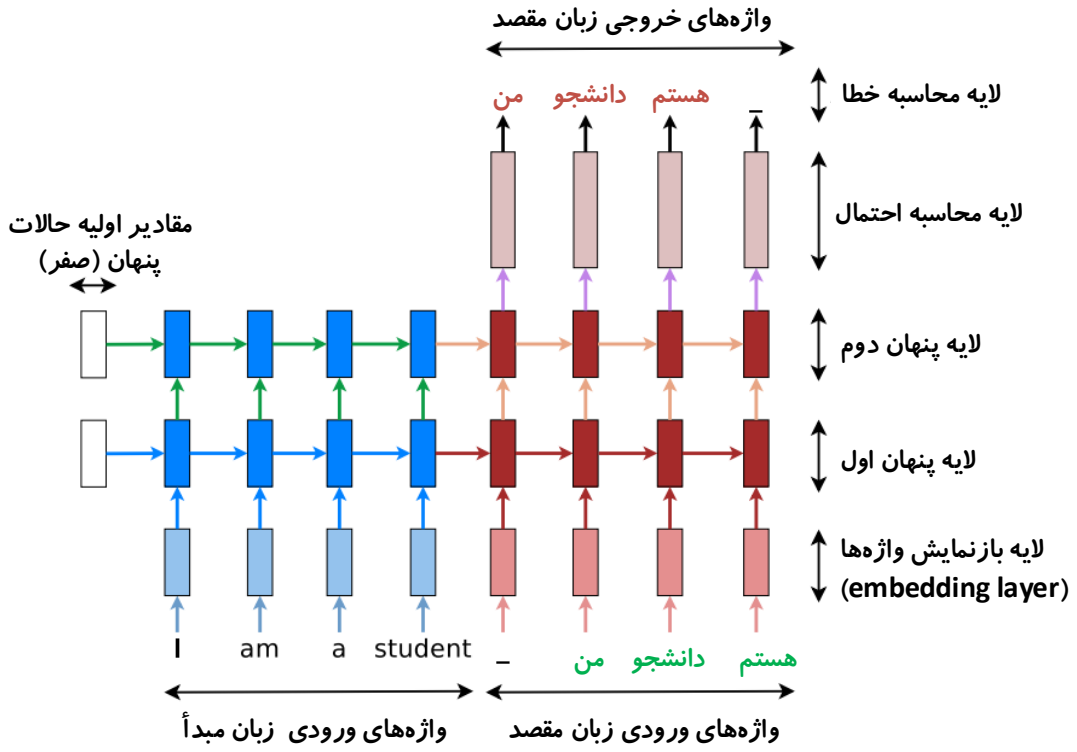
مدل توالی‌به‌توالی پس از معرفی توسط Sutskever و همکاران [1]، بارها و بارها تا به امروز مورد ارجاع دیگران قرار گرفته و تبدیل به یک مدل مرجع در NMT شده است. این مدل در رساله دکتری آقای Luong [5] به‌تفصیل و همراه با برخی اصلاحات توضیح داده شده است. در این

بخش به برخی جزئیات آموزش شبکه مدل توالی‌به‌توالی می‌پردازیم. برای این منظور از توضیحات [5] نیز کمک می‌گیریم.

شکل (۵) یک نمایش دقیق‌تر از مدل ذکر شده در شکل (۱) را نشان می‌دهد. آموزش شبکه بدین نحو است: ابتدا جمله زبان مقصد، سمت راست جمله متناظر خود در زبان مبدأ قرار داده می‌شود. نشان '-' نقش <EOS> را دارد که البته می‌تواند پایان جمله مبدأ یا آغاز جمله مقصد را مشخص کند. بنابراین به هر کدام از دو گروه قابل تعلق است. LSTM سمت چپ یا همان شبکه کدگذار، در هر مرحله‌زمانی یک واژه از جمله زبان مبدأ را خوانده پس از تبدیل به نمایش مناسب حالت داخلی لایه پنهان را بروزرسانی می‌کند. در مرحله پردازش آخرین واژه مقادیر لایه‌های پنهان بردار ثابت v که اکنون نماینده کل جمله ورودی زبان مبدأ است (موسوم به بردار محتوا[۲۵]) را تشکیل می‌دهد. سپس LSTM دوم یا شبکه کدگشا اولین واژه زبان مقصد را به همراه بردار v، به‌عنوان ورودی دریافت می‌کند و پیش‌بینی خود را انجام می‌دهد. برچسب واقعی این داده در واقع واژه بعدی در جمله زبان مقصد است. پس از مقایسه و محاسبه خطا، الگوریتم پس‌انتشار روی هر دو شبکه با شروع از شبکه کدگشا اجرا می‌شود و پارامترها را در خلاف جهت گرادیان تنظیم می‌کند. این روند تا پایان یافتن جمله زبان مقصد ادامه پیدا می‌کند. البته در عمل ممکن است ورودی در قالب یک *دسته*[۲۶] به شبکه داده و گرادیان روی کل آن دسته حساب شود. به بیان دیگر در مجموع، شبکه کدگشا آموزش داده می‌شود تا جمله زبان مقصد را به همان جمله زبان مقصدی تبدیل کند که فقط واژه‌های آن یک واحد نسبت به جمله ورودی به سمت جلو جابه‌جا شده‌اند. این روش اصطلاحا teacher forcing نامیده می‌شود [2] و زمانی مناسب است که جمله زبان مقصد (توالی خروجی) کاملا مشخص باشد. در واقع واژه بعدی به عنوان برچسب در فرایند آموزش بانظارت مورد استفاده قرار می‌گیرد و وزن‌ها بر اساس آن تنظیم می‌گردند.

[۲۵] context vector

[۲۶] batch



شکل (۵) نمایش نحوه آموزش مدل توالی‌به‌توالی روی وظیفه NMT [5].

در مرحله استنتاج[۲۷] یعنی هنگامی که می‌خواهیم جمله ناشناخته زبان مقصد (توالی خروجی) را کدگشایی نماییم، فرایند شرح داده شده در بالا، با اندکی تفاوت و در قالب گام‌های زیر انجام می‌پذیرد:

۱- توالی ورودی با استفاده از شبکه کدگذار به بردار محتوا بدل می‌گردد. در صورتی که از سلول LSTM استفاده شود بردار محتوا برای هر لایه از شبکه حاوی دو متغیر حالت خواهد بود و در صورت استفاده از سلول GRU بردار محتوا برای هر لایه از شبکه دارای یک متغیر است.

۲- یک توالی با اندازه ورودی ۱ که ابتدا حاوی نشانه شروع جمله زبان مقصد است در ورودی شبکه کدگشا قرار داده می‌شود.

۳- بردار محتوای حاصل شده از مرحله ۱ به همراه توالی مرحله ۲ به شبکه کدگشا داده می‌شوند تا نشانه (در اینجا واژه) بعدی جمله زبان مقصد پیش‌بینی شود.

---

[۲۷] inference

۴- از پیش‌بینی مرحله ۴ نمونه برداری شده (به یکی از روش‌های حریصانه یا جست‌وجوی پرتوی محلی که در ادامه توضیح داده خواهد شد) و واژه بعدی انتخاب می‌شود.
۵- واژه انتخاب شده در مرحله ۴ به جمله زبان مقصد (توالی خروجی) الحاق می‌شود.
۶- واژه انتخاب شده در مرحله ۴ به جای نشانه شروع جمله به شبکه کدگشا داده می‌شود و مراحل ۳ و ۴ و ۶ تکرار می‌شوند تا زمانی که نشانه پایان جمله تولید شود یا اینکه طول جمله تولید شده از یک حد از پیش تعیین شده بیشتر شود.

نکته لازم به ذکر دیگر آن است که توالی ورودی انتخاب شده در این مرحله از مجوعه آزمون انتخاب می‌شود. در واقع مرحله استنتاج روی داده‌های آزمون و برای ارزیابی مدل انجام می‌پذیرد.

۴-۲-۲- جزئیات آموزش شبکه

در [1] از LSTMژرف با چهار لایه و ۱۰۰۰ سلول حافظه در هر لایه استفاده شده است. همچنین اندازه واژگان ورودی ۱۶۰هزار و اندازه واژگان خروجی ۸۰هزار کلمه است. حاصل کار یک شبکه LSTM با مجموع ۳۸۰میلیون پارامتر بوده که ۶۴میلیون آن اتصالات برگشتی هستند. دیگر جزئیات پارامترها و آموزش شبکه عبارتند از:

- پارامترها با مقادیر تصادفی از توزیع یکنواخت در بازه [۰٫۰۸+ و ۰٫۰۸-] مقداردهی اولیه شده‌اند.
- برای آموزش از SGD استاندارد با نرخ یادگیری ۰٫۷ استفاده شده است. بعد از گذشت پنج دوره[۲۸]، نرخ یادگیری در هر نیم‌دور، نصف می‌شود. تعداد کل دوره‌های آموزش برابر ۷٫۵ بوده است.
- گرادیان بر روی دسته‌های ۱۲۸تایی از توالی‌ها محاسبه شده و بر اندازه دسته، یعنی ۱۲۸، تقسیم می‌شود.

[۲۸] epoch

- هرچند LSTMها از معضل میرایی گرادیان[۲۹] رنج نمی‌برند، اما ممکن است مشکل انفجار گرادیان[۳۰] را داشته باشند. بنابراین محدودیت سختی بر مقدار نورم گرادیان اعمال می‌شود به‌این نحو که هنگامی که نورم از مقدار آستانه‌ای بیشتر شد، مجددا تنظیم شود. برای هر دسته در مجموعه آموزش مقدار $s = ||g||_2$ محاسبه می‌شود که در آن g مقدار گرادیان پس از تقسیم بر ۱۲۸ است. اگر $s > 5$ شد آنگاه قرار داده می‌شود: $g = \frac{5g}{s}$.
- جملات مختلف طول‌های مختلفی دارند. بیشتر آنها کوتاه هستند (طولی بین ۲۰ تا ۳۰ دارند) اما برخی از آنها طولانی هستند (طولی بیشتر از ۱۰۰ دارند)؛ بنابراین دسته‌های ۱۲۸تایی از جملات که تصادفی انتخاب می‌شوند تعداد کمی جمله طولانی داشته و تعداد زیادی جمله کوتاه و در نتیجه سبب می‌شود تا بیشتر محاسبات داخل هر دسته هدر روند. برای غلبه بر این موضوع سعی شده است همه جملات داخل یک دسته طول تقریبا مساوی داشته باشند. این امر انجام محاسبات را تا ۲برابر تسریع کرده است.

۵ آزمایش‌ها

روش یادگیری توالی‌به‌توالی معرفی شده روی وظیفه ترجمه ماشینی انگلیسی به فرانسوی در دو حالت مختلف آزمایش گردیده است. در حالت اول مدل، برای ترجمه مستقیم جملات انگلیسی به فرانسوی به‌کار گرفته شده و در حالت دوم برای امتیاز دهی مجدد n فهرست بهتر از جملات در وظیفه SMT استفاده شده است. در این قسمت پیاده‌سازی مدل، جزئیات مجموعه داده، آزمایش‌های انجام گرفته و نتایج آنها در قالب امتیازهای ترجمه کسب شده، نمونه جملات ترجمه شده و بلأخره یک نمونه مصورسازی بازنمایی جملات ورودی، بیان شده است.

[۲۹] vanishing gradient
[۳۰] exploding gradient

۵-۱- پیاده‌سازی مدل

پیاده‌سازی مدل اولیه با C++ انجام شده است. این پیاده‌سازی از LSTMژرف با پیکربندی شرح داده شده در بخش ۴-۱-۲ روی یک GPU، تقریبا ۱۷۰۰ واژه بر ثانیه را پردازش می‌کند. این سرعت بسیار پایین است. برای این منظور مدل به صورت موازی شده روی ۸ عدد GPU اجرا می‌شود. هر لایه از LSTM روی یک GPU اجرا شده و فعالیت‌های خود را به محض محاسبه به GPU یا لایه بعدی می‌دهد. چون مدل چهار لایه دارد، چهار GPU دیگر برای موازی‌سازی بیشینه هموار استفاده شده‌اند بنابراین هر GPU مسئول محاسبه یک ضرب ماتریسی (ماتریس با اندازه ۲۰۰۰ × ۱۰۰۰) است. نتیجه حاصل از این موازی‌سازی در سطح GPU، رسیدن به سرعت پردازش ۶۳۰۰ واژه بر ثانیه است. فرایند آموزش در این شیوه پیاده‌سازی، ۱۰ روز به طول انجامید [1].

علاوه بر پیاده‌سازی اولیه، پیاده‌سازی‌های دیگری نیز از این مدل در زبان‌ها و چارچوب‌های مختلف ارایه شده است؛ از جمله دو پیاده‌سازی خوب با زبان پایتون و روی چارچوب‌های کاری Tensorflow و Keras. پیاده‌سازی Tensorflow سازوکارهای جدیدتر مثل سازوکار *attention* را نیز اضافه کرده است [12]. پیاده‌سازی Keras هم به‌جای واژه، در سطح کاراکتر انجام شده است [13].

۵-۲- جزئیات مجموعه داده

همانطور که قبلا گفته شد (بخش ۱-۳-) از مجموعه داده ترجمه انگلیسی به فرانسوی WMT'14 در آزمایش‌ها استفاده شده است [3]. مدل توصیف شده روی یک زیرمجموعه ۱۲میلیون جمله‌ای، شامل ۳۴۸میلیون واژه فرانسوی و ۳۴۰میلیون واژه انگلیسی، آموزش داده شده است. وظیفه ترجمه ماشینی و همچنین این مجموعه داده خاص، به خاطر دردسترس بودن عمومی

یک مجموعه آموزش و یک مجموعه آزمون نشانه‌گذاری شده[۳۱] جهت اهداف آموزش و ارزیابی مدل انتخاب شده است و مدل توالی‌به‌تولی مستقل از یک وظیفه خاص است.

همچنان‌که مدل‌های زبانی عصبی معمولی روی یک بازنمایی برداری در نمایش هر کلمه تکیه می‌کنند، در اینجا نیز یک واژه‌نامه با اندازه ثابت، برای هر دو زبان به‌کار گرفته شده است. برای این منظور، ۱۶۰هزار واژه از پر استفاده‌ترین واژه‌های زبان مبدأ (انگلیسی) و نیز ۸۰هزار واژه از پر استفاده‌ترین واژه‌های زبان مقصد (فرانسوی) برگزیده شده‌اند. هر واژه خارج از این واژه‌نامه‌ها که در جمله‌ها ظاهر شده باشد، با نشانه خاص "UNK" جایگزین شده است.

برای پیاده‌سازی [12] از مجموعه داده ترجمه آلمانی-انگلیسی WMT'16 [14] استفاده شده است و همچنین مدل نمونه پیاده‌سازی شده در [13] از مجموعه داده کوچکتر موجود در [4] استفاده کرده است که قابل جایگزین کردن با مجموعه‌های ذکر شده در بالا نیز هست. ایراد اساسی پیاده‌سازی در سطح کاراکتر [13] این است که معمولا در ترجمه ماشینی واژه‌ها به یکدیگر متناظر می‌شوند نه کاراکترها لذا این مدل از دقت مدل‌های در سطح واژه برخوردار نیست اما ایده خوبی در مورد استفاده در سایر وظایف مبتنی بر نگاشت توالی‌به‌توالی نظیر تولید متن به دست می‌دهد.

۵-۳- کدگشایی و امتیازدهی مجدد

هسته اصلی آزمایش‌های انجام شده در [1]، آموزش یک LSTM ژرف بزرگ روی تعداد زیادی جفت از جمله‌های زبان مبدأ و زبان مقصد است. آموزش با بیشینه کردن احتمال لگاریتمی یک ترجمه صحیح T برای جمله مبدأ داده شده S انجام می‌شود. بنابراین هدف آموزش عبارت است از:

$$1/|\mathcal{S}| \sum_{(T,S) \in \mathcal{S}} \log p(T|S) \qquad (۷)$$

[۳۱] tokenized

که در آن $\mathcal{S}$ مجموعه آموزش است. وقتی آموزش کامل شد، ترجمه‌ها با یافتن درست‌ترین ترجمه از روی LSTM تولید می‌شوند:

$$\hat{T} = \underset{T}{\operatorname{argmax}}\, p(T|S) \qquad (۸)$$

برای یافتن درست‌ترین ترجمه از یک کدگشای ساده با جست‌وجوی پرتوی محلی[۳۲] چپ به راست استفاده شده است که تعداد B فرضیه جزئی[۳۳] را نگه‌داری می‌کند. هر فرضیه جزئی پیشوندی از تعدادی ترجمه است. در هر مرحله زمانی، هر فرضیه جزئی با واژه‌های محتمل از داخل واژه‌نامه گسترش داده می‌شود. این روند تعداد فرایض جزئی را به‌سرعت افزایش می‌دهد. با توجه به مدل احتمال لگاریتمی، تمام این فرضیه‌ها به غیر از B فرضیه محتمل اول کنار گذاشته می‌شوند. به‌مجرد اینکه نشانه "EOS" به یک فرضیه الصاق شد، از جست‌وجوی پرتوی محلی حذف و به مجموعه فرایض کامل افزوده می‌گردد. هرچند این روش کدگشایی تقریبی است؛ اما، برای پیاده‌سازی راحت خواهد بود. سیستم پیشنهادی حتی با اندازه پرتوی ۱ و نیز اندازه پرتوی ۲ بیشترین مزایای این روش جست‌وجو را فراهم می‌آورد. امتیازهای BLEU حاصله از آزمایش‌های انجام شده روی مدل، در جدول (۱) ذکر شده است.

۵-۴- وارون‌سازی جملات مبدأ

درحالی‌که LSTM قابلیت حل مسائل با وابستگی‌های طولانی مدت را دارد، در طول آزمایش‌های انجام شده در [1] پژوهشگران یافته‌اند که وقتی جمله‌های مبدأ وارون شده و به‌عنوان ورودی به شبکه کدگذار داده می‌شوند، LSTM بهتر آموزش می‌بیند. توجه شود که جملات مقصد وارون نمی‌شوند. با انجام این عمل ساده، مقدار سرگشتگی[۳۴] مدل از ۵,۸ به ۴,۷ کاهش یافته‌است و مقدار امتیاز BLEU کسب شده از ترجمه‌های کدگشایی شده مدل نیز از ۲۵,۹ به ۳۰,۶ افزایش داشته است.

---

[۳۲] beam search

[۳۳] partial hypothesis

[۳۴] perplexity

نویسندگان [1] توضیح کاملی برای توجیه اثر این پدیده نداشته‌اند. توجیه اولیه آنها بدین ترتیب است که عمل وارون‌سازی جملات زبان مبدأ باعث معرفی بسیاری از وابستگی‌های کوتاه مدت به مجموعه داده می‌شود. وقتی جمله‌های زبان مبدأ را با جمله‌های زبان مقصد الحاق می‌کنیم، هر واژه در جمله مبدأ از واژه نظیرش در جمله مقصد دور می‌افتد. در نتیجه، مسئله یک دارای یک *تأخیر زمانی کمینه*[۳۵] خیلی بزرگ می‌شود [1]. با وارون‌سازی واژه‌ها در جمله مبدأ فاصله میانگین بین واژه‌های نظیر به نظیر در جمله مبدأ با جمله مقصد تغییر نمی‌کند. هرچند تعداد کمی از واژه‌های آغازین جمله مبدأ در این حالت به واژه‌های آغازین جمله مقصد بسیار نزدیک می‌شوند؛ بنابراین تأخیر زمانی کمینه مسئله تا حد زیادی کاهش می‌یابد و الگوریتم پس‌انتشار زمان کمتری را برای استقرار ارتباط میان واژه‌های جمله‌های مبدأ و جمله‌های مقصد سپری خواهد نمود. این امر درنهایت منجربه بهبود قابل توجه کارآمدی کلی مدل می‌گردد.

جدول (۱) کارآمدی LSTM روی مجموعه آزمون ترجمه انگلیسی به فرانسوی WMT'14 (ntst14). توجه شود که یک مجموعه متشکل از ۵ LSTM با اندازه پرتو ۲، ارزان‌تر (سبک‌تر) از یک LSTM تک با اندازه پرتوی ۱۲ است [1].

روش	امتیاز BLEU (ntst14)
Bahdanau و همکاران [10]	۲۸,۴۵
یک LSTM روبه‌جلو، اندازه پرتوی ۱۲	۲۶,۱۷
یک LSTM با ورودی وارون، اندازه پرتوی ۱۲	۳۰,۵۹
پنج LSTM با ورودی وارون، اندازه پرتوی ۱	۳۳,۰۰
دو LSTM با ورودی وارون، اندازه پرتوی ۱۲	۳۳,۲۷
پنج LSTM با ورودی وارون، اندازه پرتوی ۲۱	۳۴,۵۰
پنج LSTM با ورودی وارون، اندازه پرتوی ۱۲	**۳۴,۸۱**

[۳۵] minimal time lag

جدول (۲) روش‌های مشابه که شبکه‌های عصبی را در کنار ترجمه ماشینی سنتی روی مجموعه داده WMT'14 در ترجمه انگلیسی به فرانسوی استفاده کرده‌اند [1].

روش	امتیاز BLEU (ntst14)
لـبه پژوهش [15]	**۳۷,۰۰**
چـــو و همکاران [9]	۳۴,۵۴
امتیازدهی مجدد ۱۰۰۰فهرست بهتر با یک LSTM روبه‌جلو	۳۵,۶۱
امتیازدهی مجدد ۱۰۰۰فهرست بهتر با یک LSTM وارون	۳۵,۸۵
امتیازدهی مجدد ۱۰۰۰فهرست بهتر با پنج LSTM وارون	**۳۶,۵۰**
پیش‌گویی امتیازدهی مجدد ۱۰۰۰فهرست بهتر	~۴۵

ایده وارون‌سازی جمله‌های ورودی از این مهم نشئت گرفته است که در ابتدا تصور شده وارون‌سازی فقط به پیش‌بینی با اطمینان‌تر واژه‌های آغازین در زبان مقصد کمک می‌کند و منجربه پیش‌بینی کم اطمینان‌تر واژه‌های پایانی می‌شود. هرچند LSTMای که روی جملات مبدأ وارون شده آموزش دیده، در مقایسه با LSTM معمولی، روی جمله‌های طولانی عملکرد بهتری از خود نشان داده است (رجوع شود به بخش ۵-۶-).

۵-۵- ارزیابی نتایج

به‌منظور ارزیابی کیفیت ترجمه‌های صورت گرفته توسط مدل از امتیاز BLEU [16] استفاده شده است. برای محاسبه امتیاز BLEU، اسکریپت آماده multi-bleu.pl[۳۶] به‌کار رفته است. این امتیاز دهی در کارهای قبلی نیز استفاده شده است [9] و [10]، بنابراین قابل اطمینان خواهد بود. به‌عنوان نمونه، این اسکریپت برای [10] امتیاز ۲۸,۴۵ را تولید کرده است. نتایج در جدول‌های (۱) و (۲) ارایه شده‌اند. بهترین نتیجه از مجموعه LSTMهایی که در مقداردهی اولیه تصادفی و ترتیب تصادفی ریزدسته‌ها تفاوت داشته‌اند، حاصل شده است. هرچند سازوکار

[۳۶] چندین نوع محاسبه از امتیاز BLEU وجود دارد کـــه هر نوع با یک اسکریپت زبان perl تعریف شده است.

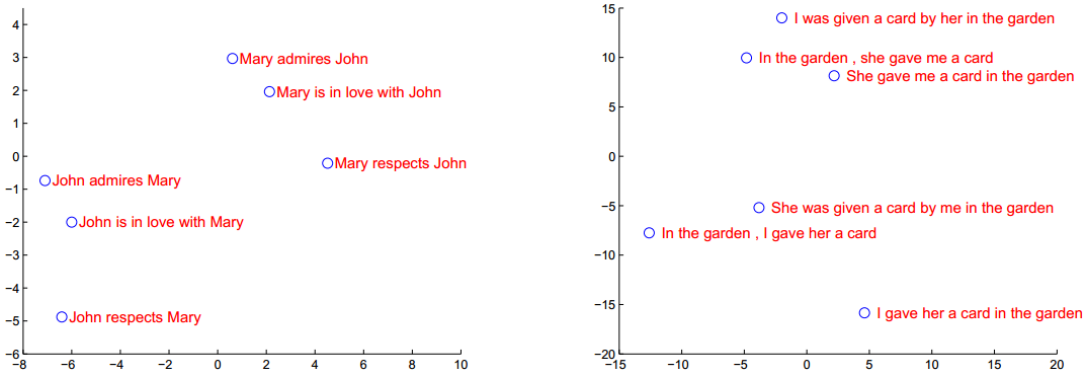
کدگشایی ترجمه به‌کار برده شده در اینجا (جست‌وجوی پرتوی محلی)، سازوکار ساده و ضعیفی است؛ با این حال نخستین بار است که یک سیستم ترجمه ماشینی عصبی خالص، سیستم ترجمه ماشینی مبتنی بر عبارات را با اختلاف قابل توجهی شکست می‌دهد. این سیستم همچنین فاقد قابلیت کنترل واژه‌های خارج از واژه‌نامه است و همان‌طور که قبلا هم بیان شد کلیه واژه‌های بیرون از واژه‌نامه با واژه "UNK" جایگزین شده‌اند. بنابراین در صورتی که سازوکاری برای کنترل این واژه‌ها نیز به مدل اضافه شود یا اندازه واژه‌نامه افزایش یابد، عملکرد این سیستم باز هم جای بهبود خواهد داشت.

۵-۶- کارآمدی روی جملات طولانی

خروجی مدل روی جمله‌های طولانی (از منظر تعداد واژه) کارآمدی بسیار خوب LSTM را در این زمینه تأیید می‌کند. یک مقایسه کمی از نتایج حاصل شده در شکل (۷) نشان داده شده است. همچنین جدول (۳) چندین جمله طولانی و ترجمه‌های تولید شده توسط مدل برای آنها را ارایه می‌کند.

۵-۷- تحلیل مدل

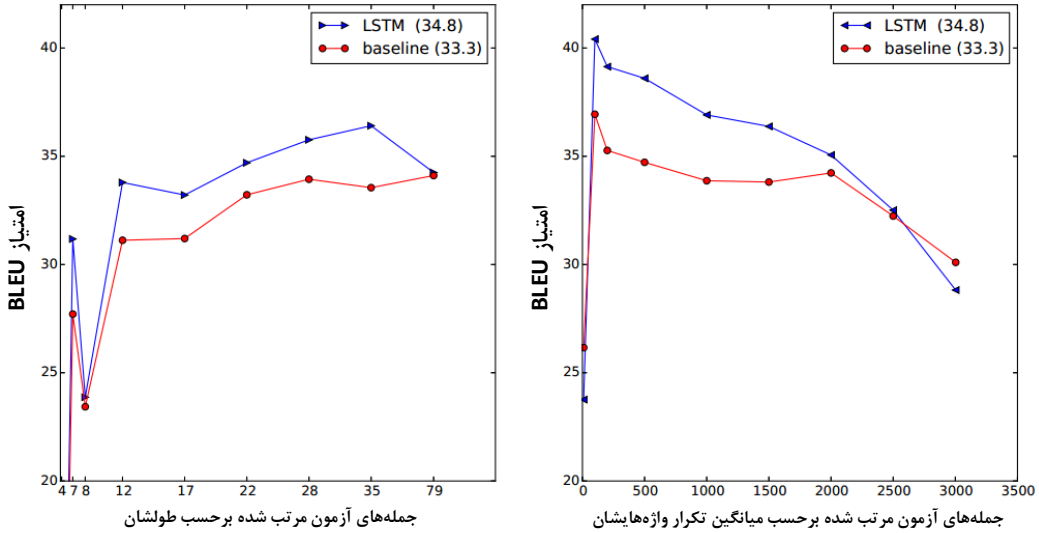
یکی از ویژگی‌های جذاب مدل توالی‌به‌توالی ارایه شده در [1]، توانایی تبدیل یک توالی از واژه‌ها به یک بردار با ابعاد ثابت است. شکل (۶) تعدادی از بازنمایی‌های یادگرفته شده در روند آموزش را مصورسازی کرده است. این تصویر به وضوح نشان می‌دهد که بازنمایی‌های ایجاد شده به ترتیب واژه‌ها حساس هستند؛ زیرا از جمله‌هایی با واژه‌های یکسان و ترتیب متفاوت در تصویر استفاده شده است. بازنمایی واقعی مدل در ابعاد بالاتری بود و برای نگاشت روی دو بعد روش PCA به‌کار برده شده است.



شکل (۶) این شکل یک تصویر PCA دوبعدی از حالت‌های پنهان LSTM را نشان می‌دهد که پس از پردازش جمله‌های نشان داده شده در شکل، گرفته شده است. عبارات با توجه به معنایشان خوشه‌بندی شده‌اند که معنا در این مثال به طور عمده تابعی از ترتیب ظاهر شدن واژه‌ها در عبارت است. رسیدن به چنین خوشه‌بندی با روش‌های سنتی موجود، سخت است. توجه شود که در همه جملات واژه‌های یکسانی استفاده شده است و تنها ترتیب، موجب تفاوت آنها شده است [1].

جدول (۳) تعدادی مثال از ترجمه‌های طولانی تولید شده توسط مدل توالی‌به‌توالی در مقایسه با ترجمه صحیح. خواننده می‌تواند صحت نتایج را با استفاده از مترجم گوگل تا حد خوبی درک کند [1].

نوع	جملـــه
مدل	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
ترجمه صحیح	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
مدل	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
ترجمه صحیح	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
مدل	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
ترجمه صحیح	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .



شکل (۷) نمودار سمت چپ کارآمدی سیستم را به‌عنوان تابعی از طول جمله‌ها نشان می‌دهد که محور افقی در آن طول واقعی جمله‌ها بر حسب تعداد واژه‌های آنها است. کاهش امتیازی در جملاتی با طول کمتر از ۳۵ واژه وجود ندارد. تنها یک کاهش جزئی در جمله‌های خیلی طولانی مشاهده می‌شود. نمودار سمت راست کارآمدی LSTM را روی جمله‌هایی با واژه‌های کمتر به‌کار رفته نشان می‌دهد که محور افقی در آن جمله‌های آزمایش شده برحسب میانگین تکرار واژه‌هایشان است [1].

۶ نتیجه‌گیری و کارهای آتی

در این گزارش یک مدل یادگیری ژرف جدید برای یادگیری و نگاشت توالی از ورودی‌ها به توالی از خروجی‌ها مطرح و بحث گردید. نشان داده شد که یک شبکه LSTM ژرف با واژگان محدود روی وظیفه ترجمه ماشینی، قادر به شکست سیستم‌های ترجمه ماشینی استاندارد مبتنی بر عبارات با واژگان نامحدود است. موفقیت این رویکرد نسبتا ساده روی وظیفه ترجمه ماشینی نشان دهنده این است که این مدل باید روی دیگر وظیفه‌های مبتنی بر توالی نیز در صورت فراهم بودن مجموعه داده‌های آموزش کافی، بسیار خوب عمل کند.

در طی فرایند آموزش این اصل نیز کشف شده که وارون سازی توالی مبدأ سبب افزایش دقت و بهبود کارآمدی مدل می‌شود. می‌توان نتیجه گرفت پیدا کردن روشی که وابستگی‌های کوتاه مدت را زودتر معرفی کند در هر صورت آموزش مدل را خیلی ساده‌تر می‌کند. لذا به نظر می‌رسد

که حتی آموزش یک RNN استاندارد (مدل غیر توالی‌به‌توالی) نیز با این روش بهتر باشد. البته این مورد در عمل مورد آزمایش قرار نگرفته است و بنابراین به صورت یک فرضیه باقی است.

نتیجه قابل ذکر دیگر، قابلیت LSTM در یادگیری صحیح ترجمه توالی‌های طولانی است. در ابتدا تصور می‌شد که LSTM به دلیل حافظه محدود خود در یادگیری جمله‌های طولانی شکست بخورد؛ همچنان‌که پژوهشگران دیگر در کارهای مشابه عملکرد ضعیفی را برای LSTM گزارش کرده بودند. با این حال اما روی جمله‌های خیلی طولانی در حالت وارون همچنان مشکل تضعیف حافظه پابرجاست و احتمالا قابلیت بهبود داشته باشد.

در نهایت نتایج رضایت بخش این مدل یادگیری نشان دهنده این است که یک مدل ساده از شبکه‌های عصبی ژرف، که هنوز جای بهبود و بهینه‌سازی‌های زیادی در خود دارد، قادر به شکست بالغ‌ترین سیستم‌های ترجمه ماشینی سنتی است. کارهای آتی می‌تواند بر روی افزایش دقت مدل توالی‌به‌توالی و پیچیده‌تر کردن آن در راستای یادگیری بهتر توالی‌های طولانی باشد. در آینده نزدیک این مدل‌ها روش‌های سنتی را کاملا منسوخ می‌کنند. نتایج همچنین نشان می‌دهد این رویکرد روی دیگر وظیفه‌های مبتنی بر نگاشت توالی‌به‌توالی می‌تواند موفقیت آمیز ظاهر شود. این مهم، زمینه را برای حل مسائل مختلفی در دیگر حوزه‌های علوم آماده می‌سازد.

می‌توان از این مدل برای ترجمه ماشینی متون طولانی انگلیسی به فارسی و بالعکس استفاده کرد در این وظیفه اثر وارون سازی جمله زبان مقصد باید بررسی شود؛ زیرا، به نظر می‌رسد در زبان‌های از راست به چپ با این کار تأخیر زمانی کمینه افزایش پیدا کند و نتیجه بدتری حاصل شود.

در وظایف دیگر مثل سیستم پرسش و پاسخ نیز می‌توان از این مدل استفاده کرد. در تولید محتوا و برای کامل کردن متون تاریخی و اشعاری که بخش‌هایی از آنها وجود ندارد یا از بین رفته است استفاده از این مدل جالب و ارزشمند به نظر می‌رسد.

علاوه بر استفاده در وظایف جدید، تغییر معماری خود مدل نیز، جهت افزایش دقت وظایف نام برده پیشنهاد می‌شود. برای مثال استفاده از RNN دوسویه، ترکیبی و نیز دارای حالت در شبکه کدگذار و کدگشا، استفاده از ژرفای بیشتر لایه‌ها، تغییر دیگر ابرپارامترهای شبکه نظیر

نرخ آموزش و افزودن سازوکار توجه می‌تواند از جمله پیشنهادهایی باشد که در ساختن مدل‌های با دقت بیشتر قابل استفاده هستند. همچنین برای مواردی که داده‌های برچسب‌دار به اندازه کافی موجود نیستند یا تمامی توالی خروجی یکجا دردسترس نیست (مثل یادگیری برخط یا یادگیری تقویتی)، استفاده از روش بیان شده در مرحله استنتاج به هنگام آموزش، به جای teacher forcing راهکار مناسبی به نظر می‌رسد.

مراجع

[1] Q. V. Le Ilya Sutskever, Oriol Vinyals, I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Nips*, pp. 1–9, 2014.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[3] "ACL 2014 ninth workshop on statistical machine translation." [Online]. Available: http://www.statmt.org/wmt14/medical-task/index.html. [Accessed: 13-Nov-2017].

[4] "Tab-delimited bilingual sentence pairs from the Tatoeba project (Good for anki and similar flashcard applications)." [Online]. Available: http://www.manythings.org/anki/. [Accessed: 13-Nov-2017].

[5] M. T. Luong, "Neural machine translation," Stanford university, 2016.

[6] A. Karpathy, "Connecting images and natural language," Stanford University, 2016.

[7] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks.," *Emnlp*, no. October, pp. 1044–1054, 2013.

[8] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," *Emnlp*, no. October, pp. 1700–1709, 2013.

[9] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," pp. 1–15, 2014.

[11] A. Graves, "Generating sequences with recurrent neural networks," pp. 1–43, 2013.

[12] M.-T. Luong, E. Brevdo, and R. Zhao, "Neural machine translation (seq2seq) tutorial," *https://github.com/tensorflow/nmt*, 2017.

[13] "Sequence to sequence example in Keras (character-level)," 2017. [Online]. Available: https://github.com/fchollet/keras/blob/master/examples/lstm_seq2seq.py. [Accessed: 13-Nov-2017].

[14] "Index of /wmt16/translation-task." [Online]. Available: http://data.statmt.org/wmt16/translation-task/. [Accessed: 04-Dec-2017].

[15] N. Durrani, B. Haddow, P. Koehn, and K. Heafield, "Edinburgh's phrase-based machine translation systems for WMT-14," *Proc. Ninth Work. Stat. Mach. Transl.*, pp. 97–104, 2014.

[16] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for

automatic evaluation of machine translation," ... *40Th Annu. Meet. ...*, no. July, pp. 311–318, 2002.

واژه‌نامه

واژه‌نامه فارسی به انگلیسی

واژه‌ی فارسی	معادل انگلیسی
انفجار گرادیان	Exploding Gradient
بانظارت	Supervised
پردازش زبان طبیعی	Natural Language Processing
پس‌انتشار	Backpropagation
تابع بیشینه هموار	Softmax Function
تأخیر زمانی کمینه	Minimal Time Lag
ترجمه ماشینی	Machine Translation
ترجمه ماشینی آماری	Statistical Machine Translation
ترجمه ماشینی عصبی	Neural Machine Translation
تشخیص گفتار	Speech Recognition
توالی	Sequence
جست‌وجوی پرتوی محلی	Beam Search
حافظه کوتاه مدت بلند	Long-Short Term Memory
دسته	Batch
دوره	Epoch
سرگشتگی	Perplexity
شبکه عصبی پیچشی	Convolutional Neural Network
شبکه عصبی رو به جلو ژرف	Deep Feed-forward Neural Network
شبکه عصبی ژرف	Deep Neural Network
شبکه عصبی مکرر	Recurrent Neural Network
فرضیه جزئی	Partial Hypothesis
کدگذار	Encoder
کدگشا	Decoder
گذر جلو	Forward Pass
مدل زبانی	Language Model
مدل زبانی عصبی	Neural Language Model
میرایی گرادیان	Vanishing Gradient
نشانه‌گذاری شده	Tokenized

واژه‌نامه انگلیسی به فارسی

معادل فارسی	واژه‌ی انگلیسی
پس‌انتشار	Backpropagation
دسته	Batch
جست‌وجوی پرتوی محلی	Beam Search
شبکه عصبی پیچشی	Convolutional Neural Network
کدگشا	Decoder
شبکه عصبی رو به جلو ژرف	Deep Feed-forward Neural Network
شبکه عصبی ژرف	Deep Neural Network
کدگذار	Encoder
دوره	Epoch
انفجار گرادیان	Exploding Gradient
گذر جلو	Forward Pass
مدل زبانی	Language Model
حافظه کوتاه مدت بلند	Long-Short Term Memory
ترجمه ماشینی	Machine Translation
تأخیر زمانی کمینه	Minimal Time Lag
پردازش زبان طبیعی	Natural Language Processing
مدل زبانی عصبی	Neural Language Model
ترجمه ماشینی عصبی	Neural Machine Translation
فرضیه جزئی	Partial Hypothesis
سرگشتگی	Perplexity
شبکه عصبی مکرر	Recurrent Neural Network
توالی	Sequence
تابع بیشینه هموار	Softmax Function
تشخیص گفتار	Speech Recognition
ترجمه ماشینی آماری	Statistical Machine Translation
بانظارت	Supervised
نشانه‌گذاری شده	Tokenized
میرایی گرادیان	Vanishing Gradient

❊❊❊

Iran University of Science and Technology
School of Computer Engineering

A Survey of Sequence-to-Sequence Architectures with Neural Networks

A Project Submitted in IUST NLP Course
Phase 3 (final phase)

By:
Morteza Zakeri Nasrabadi

Instructor:
Dr. Behrouz Minaei

February 2018