ملیکا محمدی فخار - ۹۹۵۲۲۰۸۶

۱.

a.

در one-hot encoding یکی از حالت‌های شناخته شده برای نمایش کلمات، هر کلمه به یک بردار دودویی تبدیل می‌شود که در آن تمامی اجزاء به جز یکی، صفر هستند و این یک مربوط به کلمه مورد نظر است. به عبارت دیگر، هر کلمه به یک بردار با طول برابر تعداد کلمات در vocabulary تبدیل می‌شود که در آن موقعیت متناظر با کلمه مورد نظر به یک می‌شود و بقیه مقادیر صفر هستند.

از معایب آن می‌توان به اینکه رابطه معنایی میان کلمات را در بر نمی‌گیرد و با توجه به sparse بودن بردار، از حافظه به صورت نابهینه استفاده می‌کند اشاره کرد.

از سوی دیگر word embedding یک روش برای نمایش کلمات به شکل بردارهایی با ابعاد کمتر در فضای چند بعدی است. در این روش، هر کلمه به یک بردار از اعداد حقیقی تبدیل می‌شود که نشان‌دهنده ویژگی‌های معنایی و زبانی کلمه است. برای مثال، کلماتی که معانی مشابهی دارند، در فضای برداری در نزدیکی یکدیگر قرار می‌گیرند. این روش باعث می‌شود که مدل‌ها بتوانند معنای کلمات و تشابه معنایی بین آنان را بهتر درک کنند.

b.

۱. ماتریس co-occurrence: ابتدا ماتریس co-occurrence ساخته می‌شود که هر درایه آن نشان می‌دهد چقدر دو کلمه با هم در همان پنجره‌ی متنی درون مجموعه متن تکرار می‌شوند.

۲. احتمال Word-Word Co-occurrence: با استفاده از ماتریس مرحله قبل، GloVe احتمالی را محاسبه می‌کند که یک کلمه در کنار کلمه دیگری ظاهر شود.

۳. تابع هدف: GloVe به دنبال یادگیری تعبیرات کلمات است که این احتمالات رخداد همزمان را حفظ می‌کنند. این روش یک تابع هدف تعریف می‌کند که شباهت بین بردارهای کلمات و احتمالات رخداد همزمان آنها را اندازه‌گیری می‌کند.

۴. آموزش: GloVe از الگوریتم‌های بهینه‌سازی مانند گرادیان کاهشی برای تنظیم بردارهای کلمات به صورت تکراری استفاده می‌کند، با کمینه کردن تفاوت بین احتمالات رخداد همزمان واقعی و احتمالات پیش‌بینی شده توسط الگوریتم.

۵. embeddings: پس از آموزش، embedding را تولید می‌کند که هر کلمه به عنوان یک بردار در یک فضای بعد بالا نمایش داده می‌شود. این embeddingها روابط معنایی و اطلاعات زمینه‌ای را بر اساس الگوهای مشاهده شده در مجموعه متنی به خوبی ثبت یاد می‌گیرند.

به صورت خلاصه، GloVe از توزیع آماری رخداد همزمان کلمات برای یادگیری نمایندگی‌های معنایی کلمات استفاده می‌کند، که باعث می‌شود تشابهات معنایی و روابط بین کلمات در مجموعه متنی معینی را ثبت و درک کند.

منظور از رخداد همزمان، وقوع یک کلمه پس از کلمه‌ی دیگر است. (همان co-occurrence)

c.

کلمات متناظر و همسایگی: Word۲Vec با در نظر گرفتن جفت‌های کلمات در یک مجموعه متنی مشخص، embedding کلمات را یاد می‌گیرد. هر جفت شامل یک کلمه هدف و یک کلمه همسایگی است که در نزدیکی آن در همان جمله یا پنجره متنی قرار می‌گیرد.

مدل Skip-gram: در مدل Word۲Vec، skip-gram کلمات همسایگی را با داشتن یک کلمه هدف پیش‌بینی می‌کند. این روش برای هر کلمه در مجموعه متنی، نمونه‌های آموزشی تولید می‌کند که کلمه هدف با کلمات همسایگی که در داخل یک اندازه پنجره مشخص ظاهر می‌شوند، جفت می‌شوند.

معماری شبکه عصبی: Word۲Vec از یک معماری شبکه عصبی با لایه‌های ورودی، لایه مخفی و لایه خروجی استفاده می‌کند. لایه ورودی کلمه هدف را نمایش می‌دهد و لایه خروجی احتمالات کلمات همسایگی را با توجه به کلمه هدف پیش‌بینی می‌کند.

آموزش: Word۲Vec با استفاده از مدل skip-gram و گرادیان کاهشی، احتمالات را پیش‌بینی می‌کند. در طول آموزش، وزن‌های لایه‌های میانی به گونه‌ای تنظیم می‌شوند که تفاوت بین احتمالات پیش‌بینی شده کلمات همسایگی و کلمات همسایگی واقعی در نمونه‌های آموزشی کمینه شود.

embeddings: پس از آموزش، embedding کلمات بر اساس وزن‌های لایه‌های میانی تولید می‌شوند. هر کلمه به عنوان یک بردار چگال در فضای چندبعدی قرار می‌گیرد که کلمات مشابه نزدیک به هم قرار می‌گیرند. این embeddingها اطلاعات معنایی و همسایگی را در مرحله آموزش از مجموعه متنی یاد گرفته‌اند.

d.

اطلاعات زمینه‌ای: یکی از رویکردهای متداول این است که در نظر گرفتن زمینه‌ای (context) که کلمه در آن ظاهر می‌شود. با تحلیل کلمات یا عبارات محیطی، مدل می‌تواند معنای مقصودی از کلمه مبهم را استنباط کند.

تعبیرات چندمعنایی: برخی مدل‌ها به دنبال تولید تعبیراتی هستند که معانی چندگانه‌ی یک کلمه را به صورت همزمان ثبت می‌کنند. طراحی چنین تعبیراتی نیازمند در نظر گرفتن با دقت چگونگی تعادل معانی مختلف و چگونگی آموزش مؤثر مدل برای یادگیری این معانی است.

چالش‌های مرتبط با تولید تعبیرات برای کلمات با چندین معنا عبارتند از:

ابهام: کلمات با چندین معنا ابهام را به فضای embedding می‌آورند که باعث می‌شود برای مدل امکان یادگیری یک معنای مشخص از کلمه در یک زمینه خاص را سخت کند.

داده‌های پراکنده: داده‌های آموزشی ممکن است تعداد کافی از نمونه‌های هر معنی از کلمه را فراهم نکنند، که منجر به افزایش خطا در مدل‌های پردازش زبان طبیعی می‌شود.

ارزیابی: ارزیابی کیفیت embedding برای کلمات چندمعنایی می‌تواند چالش‌برانگیز باشد، زیرا این امر نیازمند ارزیابی کردن است که چقدر embedding به دست آمده، معانی مختلف کلمه را ثبت می‌کنند.

e.

وقتی با کلماتی که خارج از واژگان مواجه می‌شویم، یکی از رویکردهای متداول استفاده از embedding زیر واژه‌ها یا سطح کاراکتری است. این روش‌ها به مدل اجازه می‌دهند تا embeddingهایی را برای کلماتی که در داده‌های آموزشی وجود نداشته‌اند، با تجزیه آن‌ها به واحدهای کوچک‌تری، مانند زیرواژه‌ها یا کاراکترها، که در طول آموزش مشاهده شده‌اند، تولید کند.

یک روش برای تولید تعبیرات واژه در محیط تولید، استفاده از fastText است که یک گسترش از Word۲Vec است که اطلاعات زیر واژه را دربر می‌گیرد. fastText تعبیراتی را برای زیرواژه‌ها (n-gramهای کاراکتری) و همچنین کلمات کامل یاد می‌گیرد، که به آن امکان می‌دهد تا embeddingهایی برای کلمات خارج از واژگان بر اساس زیرواژه‌های تشکیل‌دهنده آن‌ها تولید کند.

روش دیگر استفاده از تعبیرات سطح کاراکتری است، که هر کاراکتر در یک کلمه توسط embeddingهای خود نماینده می‌شود. سپس embeddingهای کاراکترهای کلمه به طور جداگانه محاسبه می‌شود و سپس ترکیب یا میانگین‌گیری می‌شود تا embeddingی برای کلمه کامل تولید شود.

به طور مثال، اگر کلمه "کامپیوتر" در داده‌های آموزشی وجود نداشته باشد، می‌توانیم از embeddingهای سطح کاراکتری برای نمایش آن استفاده کنیم. embedding کاراکترهای 'ک'، 'ا'، 'م'، 'پ'، 'ی'، 'و'، 'ت' و 'ر' به صورت جداگانه محاسبه شده و سپس ترکیب می‌شوند تا یک embedding برای کلمه "کامپیوتر" تولید شود.

۲.

کافیست هر کلمه را با دو کلمه پیش و پس از خود مقایسه کنیم.

	I	love	computer	science	and	NLP	even	more
I	0	2	1	1	1	1	0	0
love	2	0	1	1	1	1	1	0
computer	1	1	0	1	1	0	0	0
science	1	1	1	0	1	0	0	0
and	1	1	1	1	0	0	0	0
NLP	1	1	0	0	0	0	1	1
even	0	1	0	0	0	1	0	1
more	0	0	0	0	0	1	1	0