جاسازه‌های [1]رقیب و برتر برایGloVE

چندین روش ایجاد جاسازه رقابتی و اغلب برتر نسبت به GloVE وجود دارد که می‌توان آن‌ها را بر اساس نسل و معماری دسته‌بندی کرد (از قدیمی به جدید). اینها رقبای اصلی هستند:

۱- Word2Vec (گوگل) - رقیب هم‌عصر و مستقیم

در واقع GloVE به عنوان رقیبی مستقیم برای Word2Vec ایجاد شد. این دو ازیک دوره هستند و اغلب با هم مقایسه می‌شوند.
تفاوت کلیدی: Word2Vec از پیش‌بینی بر اساس پنجره محلی متن استفاده می‌کند، در حالی که GloVE از تجزیه ماتریس آماری هم‌رخداد جهانی کلمات بهره می‌برد.
عملکرد ایندو اغلب بسیار مشابه است و "برتر" بودن به مجموعه داده و کاربرد خاص بستگی دارد. این دو را کلاسیک این حوزه می‌دانند.

۲- fastText (فیسبوک)

fastText تکاملی بزرگ از Word2Vec است و رقیبی بسیار قوی برای GloVE محسوب می‌شود.
نوآوری کلیدی: به جای یادگیری بردار برای کل کلمات، بردارها را برایn-gram های کاراکتری (واحدهای زیر-کلمه) یاد می‌گیرد.
مزیت بزرگ: می‌تواند با شکستن کلمات جدید بهn-gram ها، برای کلمات خارج از واژگان (OOV[2]) نیز جاسازه ایجاد کند. این نقطه ضعف بسیار بزرگی برای GloVE و Word2Vec است که در برابر کلمات دیده نشده شکست می‌خورند.
مثال جاسازه کلمه "sunshine" می‌تواند از بردارهای "nshi", "unsh", "sun" و غیره ساخته شود. همچنین اشتباهات املایی را بسیار بهتر مدیریت می‌کند.

۳- جاسازه‌های بافت‌آگاه - (Contextual) انقلاب در حوزه

با ایجاد جاسازه‌های بافت‌آگاه این حوزه یک جهش کوانتومی کرد و جاسازه‌های ایستا [3] مانند GloVe را برای کارهای پیشرفته‌تر NLPمنسوخ کرد.

۳-۱- جاسازه‌های حاصل از مدل‌های زبانی (ELMo)

نوآوری کلیدی: در جاسازه‌های واژه بافت‌آگاه بردار یک کلمه مثل "بانک" بسته به اینکه در عبارت "بانک رودخانه" باشد یا "حساب بانکی" تغییر می‌کند. ELMo از یک مدل LSTM عمیق و دوطرفه که به عنوان یک مدل زبانی آموزش دیده استفاده می‌کند.
نسبت به GloVE معنا پویا و وابسته به محتوی [4]است و چندمعنایی [5]را به خوبی دریافت می‌کند.

۳-۲- BERT [6] و انواع آن (گوگل)

---

[1] embeddings
[2] out-of-vocabulary
[3] static
[4] context
[5] polysemy
[6] Bidirectional Encoder Representations from Transformers

این خانواده از مدل‌ها پس از سال ۲۰۱۸ کاملاً بر این حوزه مسلط شدند. نوآوری کلیدی آن استفاده از معماری ترنسفورمر به طور خاص رمزگذار [۷]استفاده می‌کند که در مقیاس انبوه و با استفاده از هدف "مدل زبانی پوشیده[8] " آموزش داده شده است. این مدل مانندELMo عمیقاً بافت‌آگاه است، اما بسیار قدرتمندتر. می‌توان آن‌ها را دانلود و برای کارهای خاص (مانند تحلیل احساسات، پرسش و پاسخ و غیره) تنظیم دقیق [۹]کرد. از انواع آن می‌توان به BERT، RoBERTa و DistilBERT اشاره کرد.

۳-۳- GPT و انواع آن (OpenAI)

این خانواده از مدل‌ها از معماری ترنسفورمر به طور خاص رمزگشایی [۱۰]استفاده می‌کند و به عنوان یک مدل زبانی مولد آموزش دیده استGPTها خودرگرسیونی[۱۱] هستند و کلمه بعدی را پیش‌بینی می‌کند. و همین آن را برای تولید متن استثنایی می‌کند، در حالی کهBERT مبتنی بر کدگذاری خودکار [۱۲]است و یک کلمه را درون یک جمله درک می‌کند. از آنواع آن می‌توان به GPT-2، GPT-3 و GPT-4 اشاره کرد.

۴- جاسازه‌های مبتنی بر جمله و پاراگراف

تمرکز فعلی از جاسازی کلمات منفرد به جاسازی کل جملات یا اسناد در یک بردار متراکم واحد تغییر کرده است.Sentence-BERT (SBERT) جاسازه‌های معنادار برای جمله ایجاد می‌کند تا بتوان با شباهت کسینوسی [۱۳]مقایسه شان کرد. OpenAI Embeddings API یک نقطه پایانی [۱۴]قدرتمند مانند `text-embedding-ada-002` برای دریافت جاسازه برای رشته‌های متنی ارائه می‌دهد. از مدلهای دیگری که برپایه این تکنولوژی ارایه شده است می‌توان به USE[15]، GTE[16]، E5 اشاره کرد.

از کدام باید استفاده کنیم؟

- برای یادگیری/آموزش/ایده‌پردازی GloVE و Word2Vec به دلیل سادگی و حجم کوچک، هنوز انتخاب‌های بسیار خوبی هستند.
- برای یک سیستم تولید [۱۷]که نیاز به مدیریت کلمات جدید یا نادر دارد fastText مناسب است.
- برای هر task جدی NLP که دقت در آن مهم است (مانند تحلیل احساسات، تشخیص موجودیت‌های نامدار و غیره) یک مدل از پیش آموزش دیدهBERT یا یک نوع سبک‌تر مثل DistilBERT برای سرعت بیشتر انتخاب استاندارد است.
- برای جستجوی معنایی، خوشه‌بندی، یا بازیابی اطلاعات Sentence Transformers مانند SBERT یا OpenAI Embedding API مناسب‌تر هستند.

[7] encoder
[8] masked language model
[9] fine-tune
[10] decoder
[11] autoregressive
[12] autoencoding
[13] cosine similarity
[14] endpoint
[15] universal sentence encoder
[16] general text embeddings
[17] production system