<https://medium.com/analytics-vidhya/cnn-lstm-architecture-and-image-captioning-2351fc18e8d7>

معماری CNN-LSTM و شرح تصاویر

یادگیری عمیق یکی از زمینه های مطالعاتی است که به سرعت در حال پیشرفت و تحقیق است که راه خود را به همه زندگی روزمره ما باز می کند. این به سادگی استفاده از شبکه های عصبی مصنوعی با استفاده از سخت افزار مدرن سنگین و پیشرفته است. این امکان توسعه، آموزش و استفاده از شبکه‌های عصبی را می‌دهد که بسیار بزرگ‌تر (لایه‌های بیشتر) از آنچه قبلاً تصور می‌شد ممکن است. هزاران نوع شبکه عصبی خاص وجود دارد که توسط محققان به عنوان اصلاحات یا تغییراتی در مدل های موجود پیشنهاد شده است. برخی از برجسته‌ترین آنها مانند CNN و RNN.

شبکه های عصبی کانولوشن برای نگاشت داده های تصویر به یک متغیر خروجی طراحی شده اند. آنها به قدری مؤثر بوده اند که برای هر نوع مشکل پیش بینی که شامل داده های تصویری به عنوان ورودی است، بهترین روش هستند.

شبکه‌های عصبی مکرر یا RNN برای کار با مشکلات پیش‌بینی توالی طراحی شده‌اند. برخی از این مشکلات پیش‌بینی توالی شامل یک به چند، چند به یک و چند به چند است.

شبکه‌های LSTM شاید موفق‌ترین شبکه‌های RNN باشند، زیرا به ما اجازه می‌دهند توالی وسیع‌تری از کلمات یا جملات را برای پیش‌بینی در خود محصور کنیم.

مدل CNN-LSTM

یکی از جالب‌ترین و کاربردی‌ترین مدل‌های عصبی از اختلاط انواع مختلف شبکه‌ها با هم در مدل‌های ترکیبی حاصل می‌شود.

مثال

وظیفه ایجاد زیرنویس برای تصاویر را در نظر بگیرید. در این حالت یک تصویر ورودی و یک دنباله خروجی داریم که عنوان تصویر ورودی است.

آیا می توانیم این را به عنوان یک کار پیش بینی توالی یک به چند مدل کنیم؟

بله، اما چگونه LSTM یا هر مدل پیش‌بینی توالی دیگری تصویر ورودی را درک می‌کند. ما نمی‌توانیم مستقیماً تانسور تصویر RGB را وارد کنیم، زیرا آنها برای کار با چنین ورودی‌هایی مجهز نیستند. ورودی با ساختار فضایی، مانند تصاویر، نمی تواند به راحتی با استاندارد Vanilla LSTM مدل شود.

آیا می توانیم برخی از ویژگی ها را از تصویر ورودی استخراج کنیم؟

بله، این دقیقاً همان کاری است که ما باید انجام دهیم تا بتوانیم از معماری LSTM برای هدف خود استفاده کنیم. ما می‌توانیم از معماری عمیق CNN برای استخراج ویژگی‌هایی از تصویر استفاده کنیم که سپس به معماری LSTM داده می‌شود تا کپشن را تولید کند.

این مدل CNN LSTM نامیده می شود که به طور خاص برای مشکلات پیش بینی توالی با ورودی های فضایی، مانند تصاویر یا فیلم ها طراحی شده است. این معماری شامل استفاده از لایه‌های شبکه عصبی کانولوشن (CNN) برای استخراج ویژگی روی داده‌های ورودی همراه با LSTMs برای انجام پیش‌بینی توالی بر روی بردارهای ویژگی است. به طور خلاصه، CNN LSTM ها دسته ای از مدل ها هستند که هم از نظر مکانی و هم از لحاظ زمانی عمیق هستند و در مرز بینایی کامپیوتری و پردازش زبان طبیعی قرار دارند. این مدل ها پتانسیل بسیار زیادی دارند و به طور فزاینده ای برای بسیاری از کارهای پیچیده مانند طبقه بندی متن، تبدیل ویدیو و غیره استفاده می شوند. در اینجا یک معماری عمومی از یک مدل CNN LSTM آورده شده است.

زیرنویس تصویر

توصیف یک تصویر مشکل ایجاد یک توصیف متنی قابل خواندن توسط انسان از یک تصویر است، مانند عکسی از یک شی یا صحنه. این بینایی کامپیوتری و پردازش زبان طبیعی را با هم ترکیب می کند.

مدل‌های شبکه عصبی برای عنوان‌نویسی شامل دو عنصر اصلی است:

استخراج ویژگی.

مدل زبان

بقیه مقاله افکار و مشاهدات ما را در حین اجرای یک مدل CNN-LSTM برای شرح تصاویر روشن می کند. توجه داشته باشید که این پست آموزشی در مورد اجرای زیرنویس تصویر نیست، بلکه هدف آن بررسی معماری CNN-LSTM و کاربرد عملی آن است. این کد با زبان python3 نوشته شده و در Keras پیاده سازی شده است. در اینجا الزامات و پیش نیازهای لازم برای اینکه بتوانید پیاده سازی را به طور کامل درک کنید، آورده شده است. اگر علاقه مند به آموزش پیاده سازی هستید، می توانید به https://bit.ly/2XFCEmN مراجعه کنید.

مجموعه داده دارای یک مجموعه داده آموزشی از پیش تعریف شده (6000 تصویر)، یک مجموعه داده توسعه (1000 تصویر) و مجموعه داده آزمایشی (1000 تصویر) است.

اطلاعات مجموعه داده و همچنین آماده سازی داده برای مدل را می توان در همان لینک بالا مشاهده کرد.

در اینجا، ما فقط تکه های مهم کدی که برای ایجاد و اجرای مدل استفاده شده است را نشان خواهیم داد. به شما توصیه می شود از مجموعه داده های متفاوتی استفاده کنید و مجموعه داده های خود را بر اساس آن آماده کنید.

استخراج ویژگی

مدل استخراج ویژگی یک شبکه عصبی است که یک تصویر می‌تواند ویژگی‌های برجسته را استخراج کند، اغلب به شکل یک بردار با طول ثابت. یک شبکه عصبی کانولوشنال عمیق یا CNN به عنوان زیرمدل استخراج ویژگی استفاده می شود. این شبکه را می توان مستقیماً بر روی تصاویر موجود در مجموعه داده شما آموزش داد. از طرف دیگر، می توانید از یک مدل کانولوشن از قبل آموزش دیده همانطور که نشان داده شده است استفاده کنید.

مدل زبان

برای شرح تصاویر، ما در حال ایجاد یک مدل مبتنی بر LSTM هستیم که برای پیش‌بینی توالی کلمات، به نام عنوان، از بردارهای ویژگی به‌دست‌آمده از شبکه VGG استفاده می‌شود.

مدل زبان

برای شرح تصاویر، ما در حال ایجاد یک مدل مبتنی بر LSTM هستیم که برای پیش‌بینی توالی کلمات، به نام عنوان، از بردارهای ویژگی به‌دست‌آمده از شبکه VGG استفاده می‌شود.

مدل زبان برای 20 دوره آموزش داده شده است. شما می توانید با پارامترهای دیگر بازی کنید و آنها را تا جایی که می خواهید تنظیم کنید. ما در اینجا یک نتیجه را نشان می دهیم که پس از آموزش شبکه خود به دست آوردیم.

نتیجه

معماری CNN-LSTM دارای کاربردهای گسترده ای است زیرا در رأس بخش بینایی کامپیوتری و پردازش زبان طبیعی قرار دارد. این به ما امکان می دهد از مدل های عصبی پیشرفته برای کارهای NLP مانند ترانسفورماتور برای داده های تصویری و ویدیویی متوالی استفاده کنیم. در عین حال، شبکه های CNN بسیار قدرتمند می توانند برای داده های متوالی مانند زبان طبیعی استفاده شوند. از این رو، به ما این امکان را می‌دهد تا از جنبه‌های مفید مدل‌های قدرتمند در کارهایی که قبلاً برای آنها استفاده نشده‌اند، استفاده کنیم. این پست فقط برای معرفی مفهوم مدل های عصبی ترکیبی و تشویق مردم به استفاده روزافزون از معماری های مختلف مدل های CNN-LSTM بود.