



دانشکده مهندسی کامپیوتر

پرسش و پاسخ تصویری در زبان فارسی

گزارش پیشرفت پروژه ی درس یادگیری عمیق

استاد درس :

جناب آقای دکتر پيله ور

دانشجویان :

مریم سادات هاشمی ۹۸۷۲۳۳۳۳

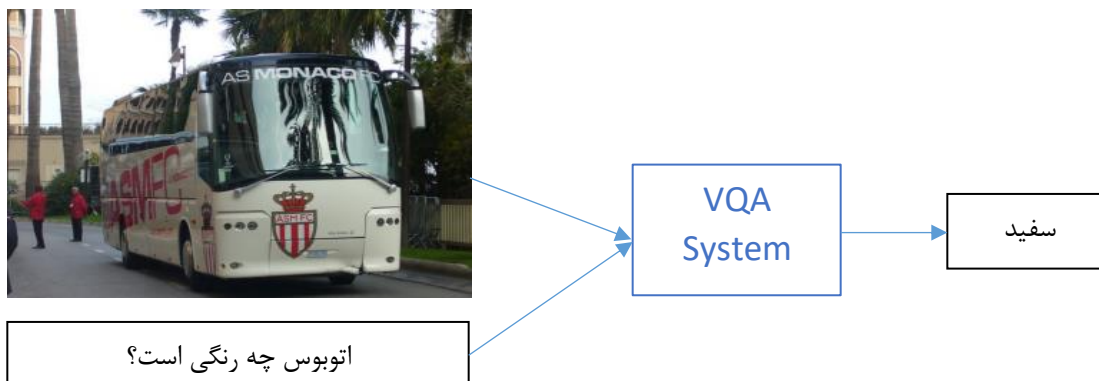
علیرضا اصغری ۹۷۷۲۲۰۱۴

بهار ۱۳۹۹

۱ تعریف مسئله

پروژه‌ی ما درباره پرسش و پاسخ تصویری^۱ در زبان فارسی است. تاکنون کارهای تحقیقاتی زیادی در این باره انجام شده است که سوال و پاسخ به زبان انگلیسی است. با توجه به جست‌وجوهایی که در این زمینه داشته‌ایم؛ تا به حال مقاله‌ی رسمی ارائه نشده است که این مسئله را برای زبان فارسی حل کرده باشند و به تبع آن هیچ مجموعه‌داده‌ی مناسبی هم برای این مسئله در زبان فارسی وجود ندارد. به همین دلیل ما تصمیم گرفتیم که مسئله‌ی پرسش و پاسخ تصویری را در زبان فارسی انجام دهیم و برای آن یک مجموعه‌داده فارسی تهیه کنیم. بنابراین بخشی از چالش کار ما تهیه‌ی مجموعه داده‌ای به زبان فارسی است. با توجه به زمان و منابعی که در اختیار داریم، تصمیم گرفتیم که یکی از مجموعه‌داده‌های پرکاربرد در این زمینه را از زبان انگلیسی به زبان فارسی ترجمه کنیم.

در شکل ۱ مثالی از مسئله‌ی پرسش و پاسخ تصویری آمده است که درک مسئله را شفاف‌تر می‌کند. انتظار ما از سیستم این است که اگر یک تصویر (مشابه شکل ۱) و یک سوال متنی (مانند «اتوبوس چه رنگی است؟») به عنوان ورودی به سیستم داده شد؛ سیستم یک پاسخ متنی (مانند «سفید») را بدست آورد که این پاسخ حاصل تفسیر و درکی است که سیستم از ورودی‌ها (تصویر + سوال متنی) داشته است.



شکل ۱- مثالی از سیستم پرسش و پاسخ تصویری

۲ آماده سازی مجموعه داده

مجموعه‌داده‌ای که برای حل مسئله انتخاب کردیم؛ مجموعه داده [VQA v1](#) است. مشخصات کامل مجموعه داده را می‌توانید در جدول ۱ مشاهده کنید.




برای ترجمه مجموعه‌داده از دو ابزار زیر استفاده کردیم:

۱. Targoman API
۲. Google Translate

تعداد تصاویر	تعداد سوالات	تعداد پاسخها
82,783	248,349	2,483,490
40,504	121,512	1,215,120
81,434	244,302	
تست		

به صورت خلاصه چالش های ترجمه به شرح زیر است:

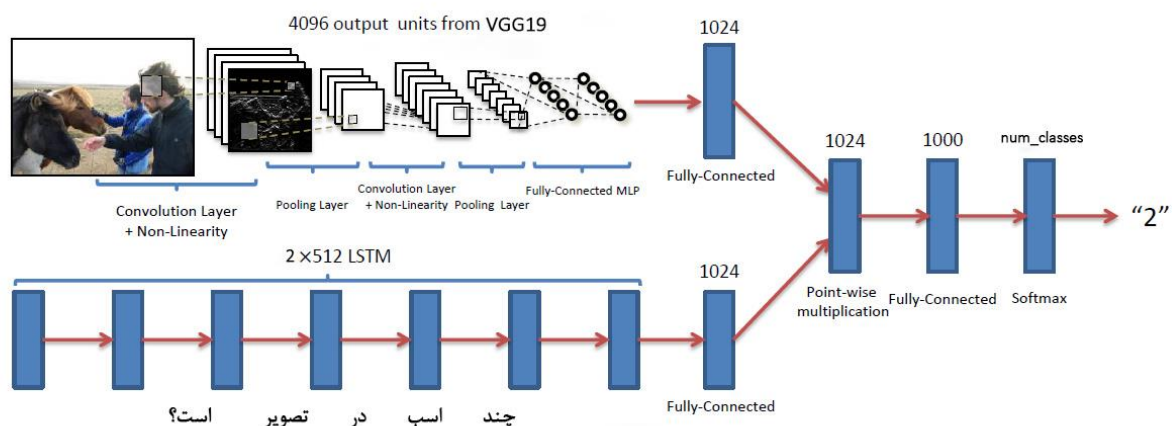
۱. **تفاوت کیفیت و عملکرد مترجم های ماشینی** : همانطور که در شکل ۲ مشاهده می نمایید در برخی موارد هر دو مترجم به خوبی عمل می کنند و معنای جمله انگلیسی در جمله فارسی ترجمه شده به وضوح دیده می شود. در برخی مواقع یکی از مترجم ها بهتر از دیگری عمل می کند. در آخر، حالتی وجود دارد که هر دو مترجم بد عمل می کنند و ترجمه ی غلطی را از جمله انگلیسی تولید می کنند.
۲. **ارزیابی کیفیت ترجمه ها** : استفاده از معیاری برای ارزیابی کیفیت متون ترجمه شده، امری مهم در ترجمه ماشینی است. این ارزیابی می تواند توسط عامل انسانی انجام گیرد و یا این که توسط ماشین انجام شود. طبعاً کیفیت روش های انسانی بالاتر است. اما در صورت وجود حجم بالایی از داده، این امر به زمان و هزینه بالایی احتیاج دارد.
۳. **تصحیح ترجمه های غلط** : برای هرچه بهتر شدن نتایج می توان ترجمه های غلط را تشخیص و تصحیح نمود. این کار می تواند به روش های جمع سپاری^۲ و توسط عوامل انسانی انجام شود، همان گونه که دیتاست VQA تشکیل شده است. اما این مساله با توجه به حجم بالای داده های ما، نیاز به هزینه و زمان بالایی دارد.
۴. **زبان رسمی و غیر رسمی ترجمه** : در نمونه های ترجمه شده توسط هر دو ابزار، مشاهده می شود که برخی ترجمه ها به زبان فارسی غیر رسمی است و ماشین مترجم تصمیم می گیرد که به ما ترجمه ی رسمی برگرداند یا غیررسمی. نمونه هایی از ترجمه ها را در شکل ۲ می توانید مشاهده کنید.

		
Are there clouds in the sky?	What toppings are on the pizza?	What is the man's expression?
ابرها در آسمان هستند؟	چه چیزی روی پیتزا است؟	چهره مرد چیست؟
آیا ابرها در آسمان وجود دارند؟	پیتزا چیست؟	بیان مرد چیست؟

شکل ۲- نمونه ای از ترجمه سوال و پاسخها توسط Google Translate و Targoman API

۳ شبکه پایه ۳

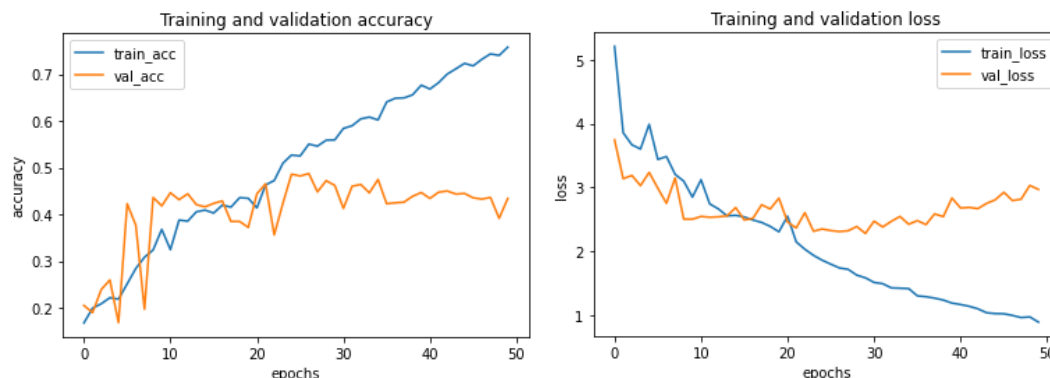
شبکه پایه‌ای که برای حل این مسئله استفاده کردیم، شبکه Vanilla [1] است. جزئیات لایه‌های این شبکه را در شکل ۳ می‌توانید مشاهده کنید. در این شبکه، مسئله ما به عنوان یک مسئله طبقه‌بندی در نظر گرفته می‌شود که در آن، ۳۰۰ پاسخ پرتکرار از داده‌های آموزشی را به عنوان کلاس انتخاب می‌کنیم و یک کلاس هم به عنوان Unknown برای سایر در نظر می‌گیریم. تصاویر را از مدل VGG-19 عبور می‌دهیم و بردار ۴۰۹۶ بعدی که در لایه آخر ایجاد می‌کند را به عنوان ویژگی‌های تصویر استخراج می‌کنیم و سپس این ویژگی‌ها را به یک لایه dense ۱۰۲۴ تایی می‌دهیم. سوال‌ها را در مرحله پیش پردازش به دنباله‌هایی با طول ۴۰ تبدیل می‌کنیم و به یک لایه embedding با ۱۰۰۰ کلمه و ۱۰۰ بعد می‌دهیم و سپس از ۲ لایه LSTM و یک لایه dense عبور می‌دهیم. با گذراندن این مراحل، به ازای هر تصویر و هر سوال که در ورودی به شبکه داده‌ایم، یک بردار ۱۰۲۴ تایی داریم که ویژگی‌های آن‌ها در این بردارها کدگذاری^۴ شده‌اند. در نهایت این دو بردار را با استفاده از elementwise multiplication ترکیب می‌کنیم. بردار حاصل را به عنوان ورودی به لایه‌های dense بعدی می‌دهیم که این لایه‌ها در واقع کار طبقه‌بندی را برای ما انجام خواهند داد.



شکل ۳- معماری شبکه Vanilla

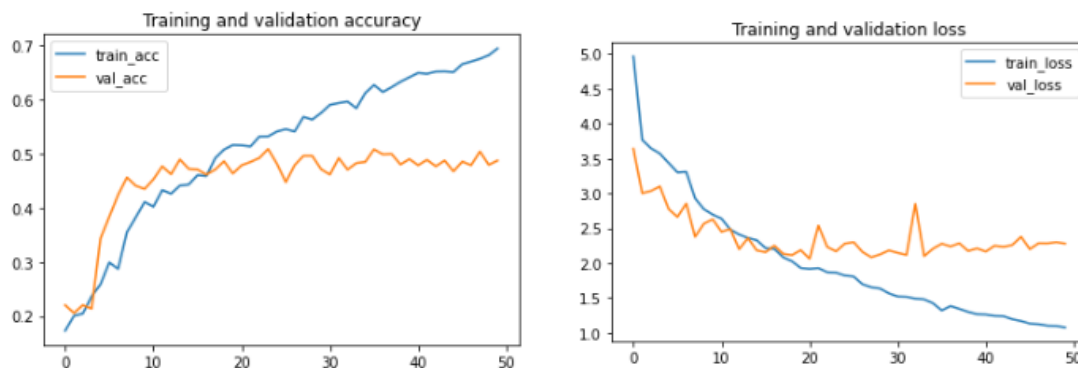
۴ نتایج شبکه پایه

تمامی پیاده سازی‌ها را می‌توانید از [این لینک](#) مشاهده کنید. با توجه به محدودیتی که در زمان و منابع محاسباتی داشتیم؛ تصمیم گرفتیم که برای تست شبکه پایه‌مان از ۳۰۰۰ داده آموزشی و ۱۵۰۰ داده ارزیابی و ۱۵۰۰ داده تست استفاده کنیم. تعداد epochها را برابر با ۵۰ و مقدار dropout را برابر با ۰.۵ قرار داده‌ایم. نمودار دقت و خطا بر حسب epoch را می‌توانید در شکل ۴ مشاهده کنید. همانطور که مشخص است مدل ما overfit شده است که علت آن هم حجم کم داده‌هایی است که در این مرحله برای اجرای مدل‌مان استفاده کرده‌ایم.



شکل ۴- نتایج مجموعه داده تهیه شده بر روی شبکه پایه

در آزمایشی دیگر، مجموعه داده اصلی (زبان انگلیسی) را به مدل‌مان دادیم اما با این تفاوت که از fastText به عنوان تعبیه معنایی کلمات استفاده کردیم. نتایج مربوط به این آزمایش هم در شکل ۵ آورده شده است.



شکل ۵- نتایج مجموعه داده اصلی با استفاده از fasttext Embedding بر روی شبکه پایه

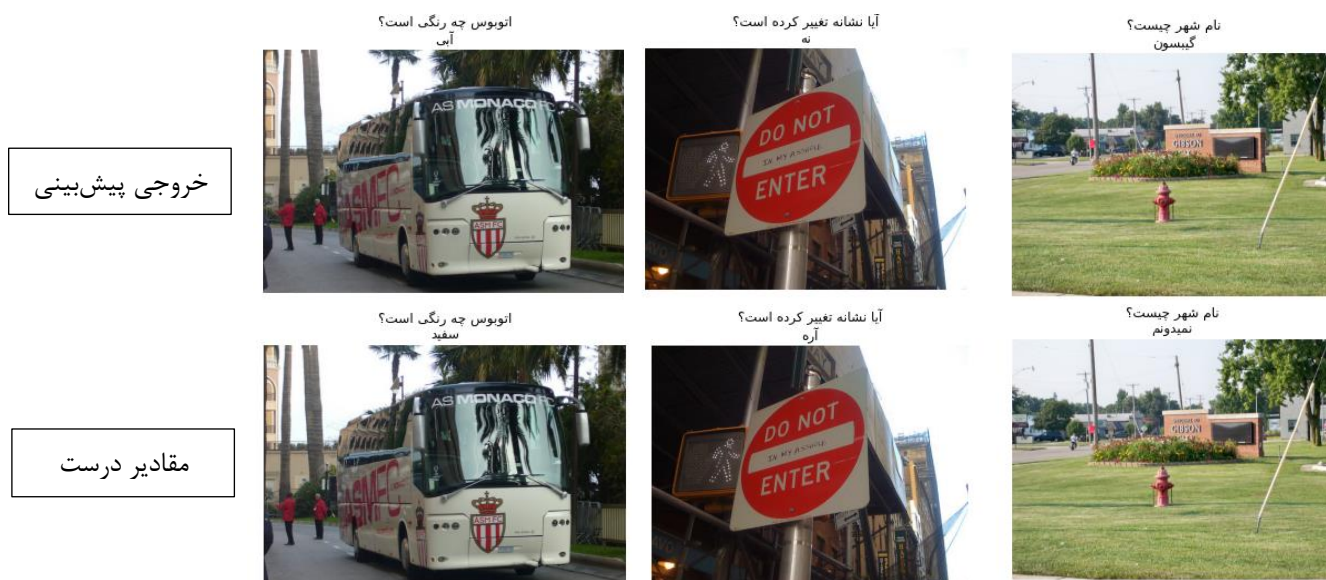
در جدول ۲ نتیجه این دو آزمایش را بر روی داده های تست به همراه دقت این مدل که در مقاله [1] معرفی شده است را می‌توانید مشاهده کنید. به نظر می‌رسد، نتیجه بدست آمده با توجه به اجرای مدل با مقدار بسیار کمی از داده ها به نسبت داده های بسیار زیاد مدل مقاله، مناسب است.

روش	دقت
[1] LSTM Q + I(baseline paper in English VQA v1)	53.74
LSTM Q + I(Our Implementation in English VQA v1+fasttext with 3000 questions)	48.7
LSTM Q + I(Our Implementation in Persian VQA v1 with 3000 questions)	43.39

شکل ۶ و ۷ چند نمونه از پیش‌بینی‌های درست و غلط مدل را نمایش می‌دهند.



شکل ۶- نمونه‌های درست پیش‌بینی شده توسط مدل



شکل ۷- نمونه‌های اشتباه پیش‌بینی شده توسط مدل

۵ برنامه‌های آینده

در ادامه می‌خواهیم مدل‌های لبه علمی و پیچیده‌تری (در پروپوزال به آن‌ها اشاره شده است) را پیاده‌سازی و اجرا کنیم و بخش‌های استخراج ویژگی تصاویر در شبکه تصویر و همچنین بخش‌های مربوط به شبکه سوال را بهبود بخشیم. علاوه بر آن قصد داریم که از تعبیه‌های معنایی متناسب با ساختار مساله استفاده کنیم و همچنین مدل پایه‌مان را به شکل مجزا بر روی کل مجموعه داده ترجمه شده توسط ترگمان و گوگل اجرا نموده و نتایج را با یکدیگر مقایسه کنیم.

۶ مراجع

- [1] Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra, Dhruv and Lawrence Zitnick, C and Parikh, Devi, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015.