

دانشكده مهندسي کامپیوتر

**پرسش و پاسخ تصویری در زبان فارسی**

**گزارش پیشرفت پروژه ی درس یادگیری عمیق**

بهار 1399

مریم سادات هاشمی 98723333

علیرضا اصغری ...

# تعریف مسئله

پروژه‌ی ما درباره پرسش و پاسخ تصویری[[1]](#footnote-1) 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 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 1 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 شکل1) 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.

اتوبوس چه رنگی است؟



VQA System

سفید

شکل 1- مثالی از سیستم پرسش و پاسخ تصویری

# آماده سازی مجموعه داده

مجموعه‌داده‌ای که برای حل مسئله انتخاب کردیم؛ مجموعه داده [VQA v1](https://visualqa.org/vqa_v1_download.html) است. مشخصات کامل مجموعه داده را می‌توانید در جدول 1 مشاهده کنید.

برای ترجمه مجموعه‌داده از دو API زیر استفاده کردیم:

1. ترگمان
2. Google Translate

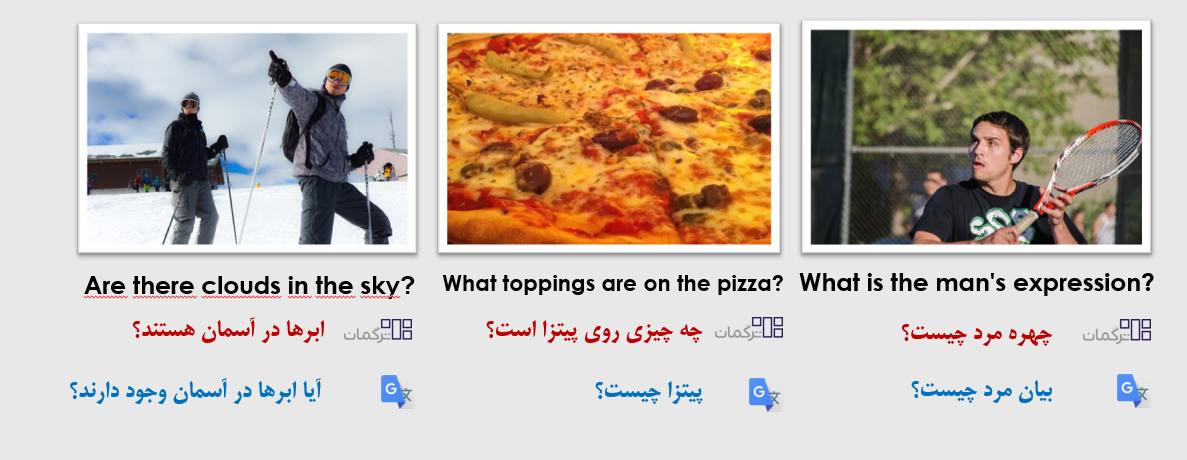
جدول 1- مشخصات مجموعه داده VQA v1

|  |  |  |  |
| --- | --- | --- | --- |
|  | تعداد تصاویر | تعدادسوالات | تعدادپاسخ‌ها |
| آموزشی | 82,783 | 248,349 | 2,483,490 |
| ارزیابی | 40,504 | 121,512 | 1,215,120 |
| تست | 81,434 | 244,302 |  |

به صورت خلاصه چالش های ترجمه به شرح زیر است:

1. ......

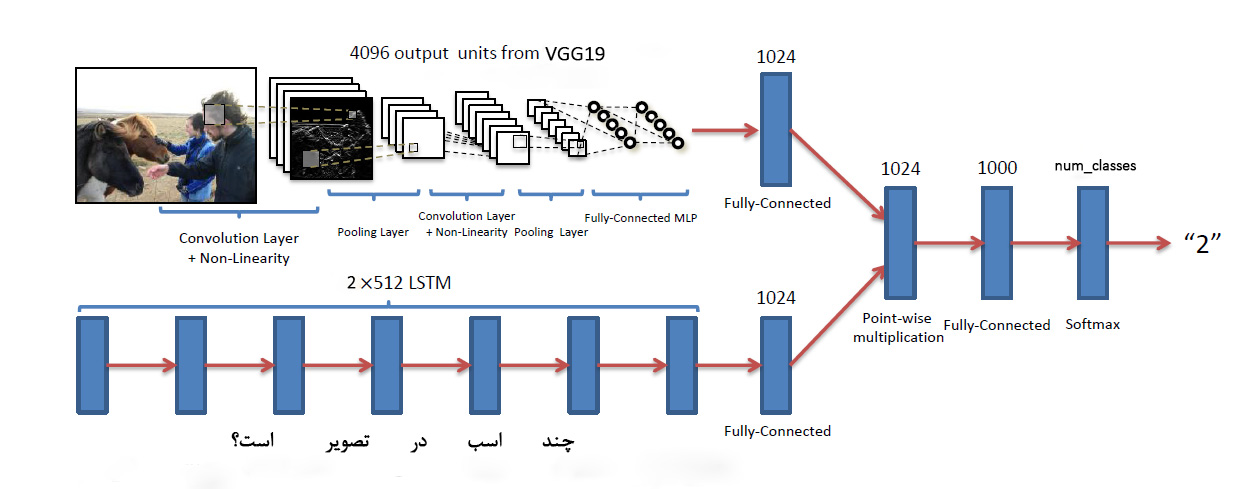
نمونه‌هایی از ترجمه‌ها را در شکل 2 می‌توانید مشاهده کنید.



شکل 2- نمونه‌ای از ترجمه‌ سوال و پاسخ‌ها توسط دو API ترگمان و Google Translate

# شبکه پایه[[2]](#footnote-2)

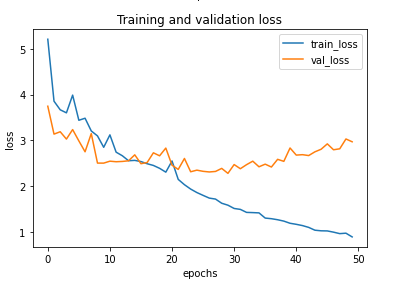
شبکه پایه‌ای که برای حل این مسئله استفاده کردیم، شبکه [1]Vanilla است. جزئیات لایه‌های این شبکه را در شکل 3 می‌توانید مشاهده کنید. در این شبکه، مسئله ما به عنوان یک مسئله طبقه‌بندی در نظر گرفته می‌شود که در آن، 300 تا پاسخ پرتکرار از داده‌های آموزشی را به عنوان کلاس انتخاب می‌کنیم و یک کلاس هم به عنوان Unknown برای سایر در نظر می‌گیریم. تصاویر را از مدل VGG-19 عبور می‌دهیم و بردار 4096 بعدی که در لایه آخر ایجاد می‌کند را به عنوان ویژگی‌های تصویر استخراج می‌کنیم و سپس این ویژگی‌ها را به یک لایه‌ی dense 1024 تایی می‌دهیم. سوال‌ها را در مرحله پیش پردازش به دنباله‌هایی با طول 40 تبدیل می‌کنیم و به یک لایه‌ی embbeding با 1000 کلمه و 100 بعد می‌دهیم و سپس از 2 لایه LSTM و یک لایه‌ی dense عبور می دهیم. با گذراندن این مراحل، به ازای هر تصویر و هر سوال که در ورودی به شبکه داده‌ایم، یک بردار 1024 تایی داریم که ویژگی‌های آن ها در این بردارها کدگذاری[[3]](#footnote-3) شده اند. 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 3- معماری شبکه Vanilla

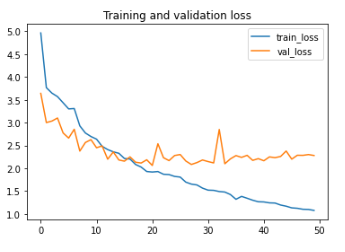
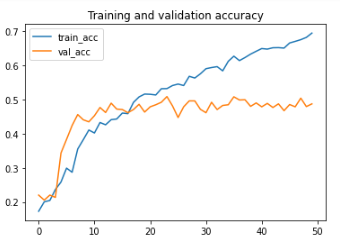
# نتایج شبکه پایه

تمامی پیاده سازی‌ها را می‌توانید از [این لینک](https://github.com/maryamhashemi/Persian_VQA) مشاهده کنید. باتوجه به محدودیتی که در زمان و منابع محاسباتی داشتیم؛ تصمیم گرفتیم که برای تست شبکه پایه‌مان از 3000 داده آموزشی و 1500 داده ارزیابی و 1500 داده تست استفاده کنیم. تعداد epochها را برابر با 50 و مقدار dropout را برابر با 0.5 قرار داده‌ایم. نمودار دقت و خطا بر حسب epoch را می‌توانید در شکل 4 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 4- نتایج مجموعه‌داده تهیه شده بر روی شبکه پایه

در آزمایشی دیگر، مجموعه داده اصلی(زبان انگلیسی) را به مدلمان دادیم اما با این تفاوت که از embedding ای به نام fastText برای استخراج ویژگی از سوالاتمان استفاده کردیم. نتایج مربوط به این آزمایش هم در شکل 5 آورده شده است.



شکل 5- نتایج مجموعه داده اصلی با استفاده از fasttext Embbeding بر روی شبکه پایه

در جدول 2 نتیجه این دو آزمایش را بر روی داده های تست به همراه دقت این مدل که در مقاله [1] معرفی شده است را می توانید مشاهده کنید. با توجه به این که ما مدل را با تعداد کمی از داده ها اجرا کردیم، به نظر می آید که نتیجه بدست آمده نسبت به نتیجه مقاله چندان بد نیست.

جدول 2- دقت بدست آمده بر روی داده‌های تست

|  |  |
| --- | --- |
| روش | دقت |
| LSTM Q + I(baseline paper in English VQA v1) [1] | 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 |

شکل6 و 7 چند نمونه از پیش‌بینی‌های درست و غلط مدل را نمایش می‌دهند.



خروجی پیش‌بینی

مقادیر درست

شکل 6- نمونه های درست پیش‌بینی شده توسط مدل



خروجی پیش‌بینی

مقادیر درست

شکل 7- نمونه‌های اشتباه پیش‌بینی شده توسط مدل

# برنامه‌های آینده

در ادامه می‌خواهیم مدل‌های state of the art پیچیده‌تری (در پروپوزال به آن ها اشاره شده است) را پیاد‌ه‌سازی واجرا کنیم. همچنین مدل پایه‌مان را بر روی کل مجموعه داده اجرا کنیم.

# 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. |

1. Visual Question Answering [↑](#footnote-ref-1)
2. Baseline Method [↑](#footnote-ref-2)
3. Encoding [↑](#footnote-ref-3)