



دانشکده مهندسی کامپیوتر

## توضیح نویسی تصویر با استفاده از ساز و کار توجه در شبکه‌ی عصبی

پروژه برای دریافت درجه کارشناسی در رشته مهندسی کامپیوتر  
گرایش هوش مصنوعی و رباتیک

سروش غلامی زوج

استاد راهنما

دکتر صالح اعتمادی

۱۳۹۸ تیر

سَلَامٌ

آ

## تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: سروش غلامی زوج

عنوان پروژه: توضیح‌نویسی تصویر با استفاده از ساز و کار توجه در شبکه‌ی عصبی

تاریخ دفاع: تیر ۱۳۹۸

رشته: مهندسی کامپیوتر

گرایش: هوش مصنوعی و رباتیک

ب

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضـا
۱	استاد راهنمای	دکتر صالح اعتمادی	استادیار	دانشگاه علم و صنعت ایران	

## تأییدیه‌ی صحت و اصالت نتایج

با اسمه تعالیٰ

این‌جانب سروش غلامی‌زوج به شماره دانشجویی ۹۴۵۲۱۱۷ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پژوهه حاصل کار این‌جانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با این‌جانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احراق حقوق مکتب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی این‌جانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: سروش غلامی‌زوج

تاریخ و امضاء:

ت

## مجوز بهره‌برداری از پایان‌نامه

- بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:
- بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.
  - بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.
  - بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

استاد راهنما: دکتر صالح اعتمادی

تاریخ:

امضا:

## چکیده

توضیح و توصیف یک صحنه یا تولید عنوان برای یک تصویر کاری نسبتاً آسان برای ما انسان‌ها است. با این حال، هنگامی که ما در حوزه هوش مصنوعی صحبت می‌کنیم، تولید چنین عناوینی به صورت خودکار مسئله‌ای پیچیده است. در این پایان نامه، یک مدل شبکه عصبی مبتنی بر یادگیری عمیق ارائه شده است که می‌تواند یک عنوان (توضیح) برای تصویر داده شده به شبکه تولید کند. این مدل از سه بخش اصلی تشکیل شده است: بخش رمزگذاری، بخش رمزگشایی و قسمت مربوط به استفاده از ساز و کار "توجه" در شبکه‌ی عصبی. به منظور تولید عنوان هم به زبان فارسی و هم به زبان انگلیسی، این مدل به طور جداگانه، یک بار بر روی نسخه‌ی اصلی مجموعه دادگان COCO و یک بار هم بر روی نسخه‌ی فارسی‌شده این مجموعه‌ی داده آموزش داده شده است.

واژگان کلیدی: توضیح‌نویسی تصویر، عنوان‌نویسی تصویر، شبکه عصبی، توجه، یادگیری عمیق

# فهرست مطالب

ح

فهرست تصاویر

د

فهرست جداول

۱

فصل ۱: مقدمه

۱ ..... ۱-۱ توضیح نویسی تصویر

۲ ..... ۱-۲ توجه در شبکه‌های عصبی

۳

فصل ۲: مجموعه داده COCO

۳ ..... ۲-۱ محتويات مجموعه داده

۵

فصل ۳: شبکه‌ی عصبی

۵ ..... ۳-۱ شبکه‌ی رمزگذار

۶ ..... ۳-۲ شبکه‌ی رمزگشا

۷ ..... ۳-۳ ساز و کار توجه

۹

فصل ۴: توضیح نویسی تصویر به زبان فارسی

۹ ..... ۴-۱ ترجمه‌ی مجموعه داده به زبان فارسی

۱۰ ..... ۴-۲ پیاده‌سازی مدل

۱۲

فصل ۵: نتایج

۱۳ ..... ۵-۱ نتایج فارسی

۱۳ ..... ۵-۱-۱ نتایج خوب

ج

## فهرست مطالب

ج

۱۶	۵-۱-۲ نتایج بد
۱۹	۵-۲ نتایج انگلیسی
۱۹	۵-۲-۱ نتایج خوب
۲۱	۵-۲-۲ نتایج بد

۲۵

مراجع

۲۶

واژه‌نامه فارسی به انگلیسی

۲۷

واژه‌نامه انگلیسی به فارسی

## فهرست تصاویر

۶ .....	۳-۱	معماری شبکه‌ی VGGNet
۷ .....	۳-۲	معماری کلی شبکه‌ی عصبی استفاده شده
۱۳ .....	۵-۱	گروهی از مردم در ساحل با بادبادک بازی می‌کنند.
۱۴ .....	۵-۲	گروهی از فیل‌ها که در مزرعه ایستاده بودند.
۱۴ .....	۵-۳	یه دختر کوچولو که یه تیکه پیتزا می‌خوره.
۱۴ .....	۵-۴	یک بازیکن تنیس روی زمین تنیس بازی می‌کند.
۱۵ .....	۵-۵	یک اسکی باز در حال اسکی روی یک شیب پوشیده از برف است.
۱۵ .....	۵-۶	یک چراغ راهنمایی در کنار یک ساختمان بزرگ
۱۶ .....	۵-۷	یک هواپیما در آسمان آبی پرواز می‌کند.
۱۶ .....	۵-۸	مردی که در آشپزخانه ایستاده بود.
۱۷ .....	۵-۹	یک میز پر از غذا و یک لیوان شراب.
۱۷ .....	۵-۱۰	یک حمام با یک یخچال سفید و یک ظرفشویی
۱۸ .....	۵-۱۱	مردی که روی کاناپه نشسته بود.
۱۸ .....	۵-۱۲	مردی که سوار بر اسب بود.
۱۹ .....	۵-۱۳	a traffic light on the side of a street.
۱۹ .....	۵-۱۴	a group of elephants standing next to each other.
۲۰ .....	۵-۱۵	a train traveling down the tracks near a train station.
۲۰ .....	۵-۱۶	a little girl eating a hot dog.
۲۰ .....	۵-۱۷	a man swinging a tennis racket on a tennis court.

فهرست تصاویر

خ

- ۲۱ ..... a vase filled with flowers on a table. ۵-۱۸
- ۲۱ ..... a group of people seating on a bench. ۵-۱۹
- ۲۲ ..... a cat sitting on top of a wooden floor. ۵-۲۰
- ۲۲ ..... a woman holding a hot dog in front of a table. ۵-۲۱
- ۲۳ ..... a group of elephants walking across a sandy beach. ۵-۲۲
- ۲۳ ..... a group of people flying kites in the sky. ۵-۲۳
- ۲۴ ..... a close up of a bowl of food. ۵-۲۴

## فهرست جداول

۱۱ ..... مقایسه نتایج مدل انگلیسی و مدل فارسی ۴-۱

۱۲ ..... ارزیابی مدل ۵-۱

# فصل ۱

## مقدمه

### ۱-۱ توضیح نویسی تصویر

یکی از اهداف اصلی در حوزه‌ی بینایی ماشین، درک یک صحنه یا تصویر است. منظور از درک صحنه، توانایی تحلیل یک تصویر ورودی و استخراج داده‌های مورد نیاز از آن است، درست همان‌طور که انسان عمل می‌کند. سیستم‌های درک تصویر از ۳ سطح تشکیل می‌شوند: سطح پایین شامل درک لبه‌ها، بافت عناصر و مناطق در تصویر؛ سطح متوسط شامل درک مرزها، سطوح و حجم در تصویر؛ و سطح بالا شامل درک اشیاء، اعمال و اتفاقات تصویر است. [۱]

توانایی تولید توضیح برای یک تصویر به صورت خودکار، یکی از وظیفه‌های بسیار مرتبط با درک صحنه است. در این نوع وظیفه، دو عمل اصلی می‌بایست انجام شود: درک درست و کامل تصویر، و توضیح مناسب آنچه که از تصویر درک شده است.

در گام نخست، می‌بایست سیستم آنقدر قدرتمند باشد تا بتواند با وجود تمام چالش‌های موجود در حوزه‌ی بینایی ماشین، تصویر ورودی را تحلیل کند. این تحلیل باید به گونه‌ای باشد که اشیاء موجود در تصویر، هم درست و هم کامل، تشخیص داده شوند. این گام، همان سطح بالا در درک صحنه است.

در گام بعد، سیستم باید این توانایی را داشته باشد تا هر آنچه را که در گام نخست از تصویر درک کرده به شیوه‌ای صحیح بیان کند. در حوزه‌ی پردازش زبان و گفتار، به این امر، تولید زبان طبیعی گفته می‌شود. بیان

## فصل ۱. مقدمه

### ۱-۲. توجه در شبکه‌های عصبی

اشیاء موجود در تصویر و روابط میان آنها، همگی با رعایت قوائد دستوری زبان طبیعی، وظیفه‌ی سیستم در این گام است. [۵]

### ۱-۲ توجه در شبکه‌های عصبی

در سال‌های اخیر، به لطف پیشرفت‌های صورت گرفته در زمینه‌ی یادگیری ماشین و تولید مجموعه داده‌های بسیار وسیع، علاقه به حل مسئله‌ی توضیح نویسی تصویر بیش از پیش شده است. این موج جدید از تحقیقات مسبب تولید روش‌های نو و در نتیجه بهبود چشمگیر نتایج بدست آمده در این مسئله نیز شده است. یکی از مهم‌ترین روش‌های ابداع شده در زمینه‌ی یادگیری ماشین، ساز و کار توجه است. این ساز و کار از روی طبیعت و نحوه پردازش محیط توسط انسان الهام گرفته شده است. در این روش، به جای در نظر گرفتن تمام عکس و فشرده‌سازی آن به یک بردار کوچک برای نمایش ویژگی‌های تصویر، به هر قسمت از تصویر به صورت جداگانه و مناسب با زمان توجه می‌شود. این روش کمک می‌کند تا به جزئیات تصویر بیشتر پرداخته شود؛ که همین امر باعث افزایش دقت و کارایی سیستم می‌شود. [۶]

## فصل ۲

# مجموعه داده‌ی COCO

در این فصل به معرفی مجموعه داده‌ی مورد استفاده در این مسئله می‌پردازیم. همچنین به دلیل آنکه یکی از اهداف این پژوهه تولید توضیح برای تصویر به زبان فارسی بوده است، به بررسی چگونگی ترجمه‌ی این مجموعه داده به زبان فارسی نیز می‌پردازیم.

### ۲-۱ محتویات مجموعه داده

یک مجموعه داده‌ی بزرگ مقیاس است که برای مسائل تشخیص شیء، قطعه‌بندی و توضیح نویسی تصویر استفاده می‌شود. این مجموعه داده از ۳۳۰ هزار تصویر در ۸۰ دسته‌بندی متفاوت تشکیل شده است. همچنین در این مجموعه داده برای هر تصویر ۵ توضیح نوشته شده است.<sup>۱</sup>

در این پژوهه، از نسخه‌ی سال ۲۰۱۴ این مجموعه داده استفاده شده است. این نسخه شامل مجموعه‌ی Train و Validation است که هر کدام به ترتیب شامل ۴۱۴۱۱۳ و ۲۰۲۶۵۴ توضیح هستند. همچنین این نسخه در مجموع ۶۰۶۷۶۷ توضیح متفاوت برای تصویرهای موجود در مجموعه داده دارد که در این میان، حدود ۱۶۰۰ جمله‌ی آن بعد از ترجمه به فارسی توسط مترجم هوشمند فرازین دچار مشکل ترجمه‌ی ناقص

<sup>۱</sup> سایت مجموعه داده COCO

## فصل ۲. مجموعه دادهی COCO

### ۲-۱ . محتویات مجموعه داده

شده است و ۵۹۰۰۰ جمله‌ی دیگر بدون مشکل به فارسی ترجمه شده است

## فصل ۳

### شبکه‌ی عصبی

اگر بخواهیم از سطح انتزاع بالا به مدل شبکه عصبی استفاده شده نگاه کنیم، این مدل در واقع از سه بخش کلی تشکیل شده است: شبکه‌ی رمزگذار، شبکه‌ی رمزگشایی و ساز و کار توجه؛ که در این قسمت هر شبکه به تفصیل توضیح داده می‌شود و در آخر نیز نحوه‌ی استفاده از ساز و کار توجه مورد بررسی قرار می‌گیرد.

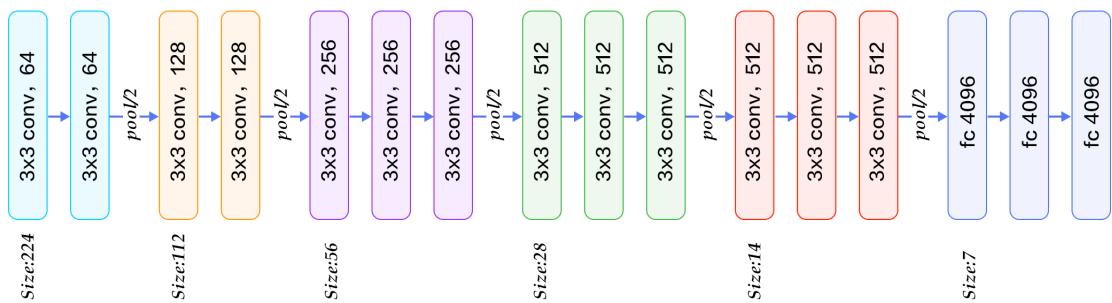
#### ۱- ۳- شبکه‌ی رمزگذار

ورودی این قسمت از مدل تصاویر هستند و خروجی آن ویژگی‌هایی هستند که از تصاویر استخراج شده‌اند. در واقع هدف اصلی این بخش، جدا کردن ویژگی‌های مناسب از تصاویر به منظور استفاده از آن‌ها در شبکه‌ی رمزگشایی است. [۶] برای پیاده‌سازی این بخش از یک مدل از پیش آموزش داده شده به نام VggNet استفاده می‌کنیم که ساختار آن به این شکل است:

در ابتدا دو لایه‌ی پیچشی وجود دارد که هر لایه از ۶۴ فیلتر ۳ در ۳ تشکیل شده است. خروجی این ۲ لایه وارد یک لایه‌ی Pooling می‌شود که از نوع ماکریزمگیر است. سپس دوباره از دو لایه‌ی پیچشی و یک فیلتر Pooling ماکریزمگیر استفاده می‌شود با این تفاوت که این‌بار هر لایه‌ی پیچشی از ۶۴ فیلتر ۳ در ۳ تشکیل شده است.

این بار از سه لایه‌ی پیچشی متتشکل از ۲۵۶ فیلتر ۳ در ۳ و یک فیلتر تجمع‌گیر دیگر استفاده می‌کنیم. در

## ۳-۲. شبکه‌ی رمزگشای



شکل ۱-۳: معماری شبکه‌ی VGGNet

انتها از شش لایه‌ی پیچشی استفاده می‌کنیم که سه لایه‌ی اول با یک لایه‌ی تجمع‌گیر از سه لایه‌ی دیگر جدا شده‌اند و هر شش لایه از ۵۱۲ فیلتر ۳ در ۳ تشکیل شده است. تمام فیلترهای Pooling دارای پنجره‌های ۲ در ۲ هستند.

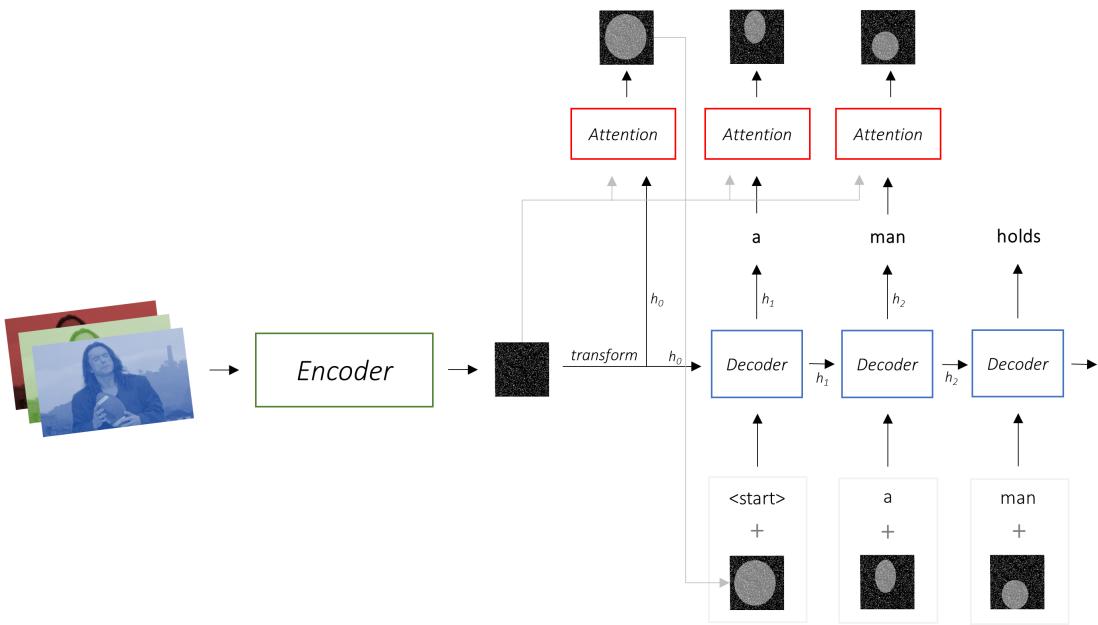
در ضمن به دلیل آنکه از شبکه‌ی VGGNet تنها برای استخراج ویژگی استفاده شده‌است، سه لایه‌ی آخر این مدل در اینجا حذف شده‌اند.

## ۳-۲ شبکه‌ی رمزگشا

در این بخش با استفاده از ویژگی‌های استخراج شده از شبکه‌ی رمزگذار، یک توضیح برای تصویر ورودی تولید می‌شود. در این شبکه، به دلیل متغیر بودن طول توضیح‌ها و همچنین اهمیت ترتیب در کلمات از مدل‌های بازگشتی استفاده می‌کنیم. [۴]

بنابراین، ورودی این بخش از مدل، خروجی قسمت قبل می‌باشد؛ که در واقع همان ویژگی‌های استخراج شده از تصاویر هستند. خروجی بخش رمزگشا، جملات (توضیحات) تولید شده برای هر تصویر هستند. معماری این بخش، یک شبکه‌ی بازگشتی است که در واقع تنها یک سلول LSTM است که طول بردar-Em آن برابر با ۵۱۲ می‌باشد. bedding

با استفاده از همین دو بخش، شبکه‌ی رمزگذار و شبکه‌ی رمزگشا، می‌توان یک مدل کامل برای انجام توضیح نویسی تصویر به کمک شبکه‌ی عصبی ارائه کرد. اما در سال‌های اخیر، پیشرفت‌های چشمگیری در



شکل ۲-۳: معماری کلی شبکه‌ی عصبی استفاده شده

معماری شبکه‌های عصبی برای تولید خروجی دقیق‌تر صورت گرفته است که یکی از تأثیر گذارترین آن‌ها، استفاده از ساز و کار توجه است. [۳]

### ۳-۳ ساز و کار توجه

در شبکه‌ای که از ساز و کار توجه استفاده نمی‌شود، شبکه‌ی رمزگشا به تمام قسمت‌های تصویر با ضریبی (وزنی) یکسان نگاه می‌کند. اما هنگامی که از ساز و کار توجه استفاده می‌شود، می‌توان برای هر پیکسل از تصویر یک وزن جداگانه در نظر گرفت. با انجام این روش، مدل در هر مرحله برای تولید کلمه‌ی بعد می‌داند که باید به کدام بخش از تصویر توجه بیشتری بکند، که بدین منظور از یک ماتریس وزن استفاده می‌شود.

برای به دست آوردن ماتریس وزن، از دو شبکه‌ی متراکم دو لایه استفاده شده است. ورودی این شبکه‌ها از دو بخش تشکیل شده است: ویژگی‌های استخراج شده توسط شبکه رمزگذار و آخرین کلمه‌ی تولید شده توسط شبکه‌ی رمزگشا. در نهایت، خروجی هر دو شبکه، که دو بردار هستند، با یکدیگر جمع شده و وارد یک

### فصل ۳. شبکه‌ی عصبی

#### ۳-۳. ساز و کار توجه

لایه‌ی Softmax می‌شود.

## فصل ۴

### توضیح نویسی تصویر به زبان فارسی

#### ۱- ۴- ترجمه‌ی مجموعه داده به زبان فارسی

از آنجایی که COCO یک مجموعه داده‌ی بزرگ مقیاس است، ترجمه‌ی تمام توضیح‌های این مجموعه داده توسط انسان کاری غیرممکن است. به همین دلیل، برای انجام این کار می‌بایست از یک سیستم خودکار استفاده شود.

در این پروژه، برای نخستین بار، این مجموعه داده‌ی بزرگ به زبان فارسی ترجمه شده است و برای انجام یک مسئله به زبان فارسی استفاده شده است. پیش از این پروژه، هیچ مجموعه داده‌ی فارسی بزرگ مقیاس برای مسئله‌ی توضیح نویسی تصویر وجود نداشت و این اولین بار است که چنین کاری انجام می‌شود. برای استفاده از این مجموعه داده، می‌توانید به لینک‌های زیر مراجعه کنید:

● مجموعه داده‌ی آموزش

<https://drive.google.com/open?id=1rePMKnLd7Vg1JgWbTnNE25N3sXoAYCYb>

● مجموعه داده‌ی سنجش

[https://drive.google.com/open?id=1bl3ibaD2I5nAA77WLb\\_Ikti8y24cFgMc](https://drive.google.com/open?id=1bl3ibaD2I5nAA77WLb_Ikti8y24cFgMc)

به دلیل اعمال محدودیت تعداد کاراکترهای ترجمه شده‌ی روزانه در تمامی سرویس‌های ترجمه‌ی آنلاین و با توجه به حجم بسیار بالای این مجموعه داده، امکان استفاده‌ی رایگان از هیچ‌کدام از این سرویس‌ها وجود ندارد. همچنین در میان سرویس‌های موجود برای ترجمه از زبان انگلیسی به فارسی، دو سرویس Google و فرازین<sup>۱</sup> به دلیل مبتنی بودن بر شبکه‌ی عصبی بهترین خروجی را تولید می‌کنند. در نهایت، به دلیل هزینه‌ی مالی بسیار بالاتر سرویس Google Translate، برای انجام این پروژه‌ی پایانی مقطع کارشناسی از مترجم هوشمند انگلیسی به فارسی فرازین استفاده شده است.

متأسفانه مترجم هوشمند فرازین از ایرادات متعددی رنج می‌برد؛ که همین امر باعث ایجاد مشکلاتی در خروجی بدست آمده از این سرویس می‌شود. یکی از بزرگترین این مشکلات، ترجمه‌ی ناقص یک جمله از انگلیسی به فارسی است. این مشکل زمانی رخ می‌دهد که تمام کلمات یک جمله‌ی انگلیسی به فارسی ترجمه نمی‌شود و معمولاً یک یا دو کلمه به همان صورت انگلیسی در ترجمه آورده می‌شود.

برای مثال، هنگامی که جمله‌ی "A brown horse is grazing grass near a red house." به عنوان ورودی داده شود، مترجم هوشمند فرازین جمله‌ی "یک اسب قهوه‌ای در حال grazing در نزدیکی یک خانه قرمز است." به عنوان ترجمه‌ی آن برمی‌گرداند. این مشکل زمانی عجیب‌تر می‌شود که می‌بینیم با یک تغییر کوچک و نامرتب نسبت به جمله‌ی اصلی، این ترجمه‌ی ناقص از بین می‌رود! برای نمونه، اگر تنها کلمه‌ی "text:" را به ابتدای همان جمله اضافه کیم، ترجمه‌ی جمله به "متن: اسب قهوه‌ای در نزدیکی یک خانه قرمز در حال چرا است." تغییر پیدا می‌کند. و البته باید این نکته را در نظر داشت که ممکن است اضافه کردن کلمه‌ای دیگر به جمله‌ی اصلی، تغییر ناخواسته‌ی دیگری در خروجی اعمال کند.

## ۴-۲ پیاده‌سازی مدل

تمامی مراحل انجام این پروژه به کمک Google Colab انجام شده است. این فایل با صورت کامل-[com](#) گذاری شده است. در این [لينك](#) می‌توانید به این فایل دسترسی داشته باشید. برای نصب ماثوله‌ای مورد نیاز بخش سوم را باید اجرا کنید. برای بارگذاری داده‌ها از بخش چهارم استفاده کنید و در نهایت برای

<sup>۱</sup> [سایت مترجم هوشمند فرازین](#)

استفاده از مدل می‌توانید از بخش آخر در فایل موجود در Google Colab استفاده کنید.

جدول ۱-۴: مقایسه‌ی نتایج مدل انگلیسی و مدل فارسی

تصویر	توضیح انگلیسی	ترجمه‌ی توضیح انگلیسی	توضیح فارسی
	a group of elephants standing next to each other.	گروهی از فیل‌ها کنار هم ایستاده بودند.	گروهی از فیل‌ها که در یک مزرعه ایستاده بودند.
	a train traveling down the tracks near a train station.	قطاری که از ریل قطار به ایستگاه قطار حرکت می‌کرد.	قطاری که روی ریل قطار نشسته بود.
	a little girl eating a hot dog.	یک دختر کوچک که یک هات‌داغ می‌خورد.	یه دختر کوچولو که یه تیکه پیتسا میخوره.
	a man swinging a tennis racket on a tennis court.	مردی که یک راکت تنیس را روی زمین تنیس تاب می‌داد.	یک بازیکن تنیس روی زمین تنیس بازی می‌کند.
	a vase filled with flowers on a table.	گل‌دانی پر از گل روی میز.	یک گل‌دان پر از گل روی یک میز.
	a group of people seating on a bench.	گروهی از مردم روی نیمکت چوبی نشستند.	یک نیمکت چوبی کنار یک نیمکت چوبی نشسته بود.
	a woman holding a hot dog in front of a table.	زنی جلوی یک میز یک سگ داغ را در دست گرفته بود.	یک پسر جوان در حال خوردن غذا در یک رستوران است.
	a group of elephants walking across a sandy beach.	گروهی از فیل‌ها در امتداد ساحل شنی قدم می‌زنند.	گروهی از فیل‌ها در ساحل شنی ایستاده بودند.
	a group of people flying kites in the sky.	گروهی از مردم بادبادک را در آسمان پرواز می‌کرند.	گروهی از مردم در حال پرواز در حال پرواز هستند.

## فصل ۵

### نتایج

در این فصل به بررسی نتایج مدل یادگیری شده با استفاده از داده‌های فارسی و انگلیسی می‌پردازیم. در ابتدا در جدولی تولید توضیح به زبان فارسی با توضیح به زبان انگلیسی مقایسه شده است؛ و در ادامه از هر کدام از زبان‌ها شش مثال خوب و شش مثال بد آورده شده است.

جدول ۱-۵: ارزیابی مدل

معیار مورد استفاده	زبان فارسی	زبان انگلیسی
Bleu-1	0.376	0.658
Bleu-2	0.263	0.483
Bleu-3	0.179	0.346
Bleu-4	0.120	0.246
Meteor	0.215	0.206
Rouge L	0.269	0.484
CIDER	0.375	0.667

## ۱-۵ نتایج فارسی

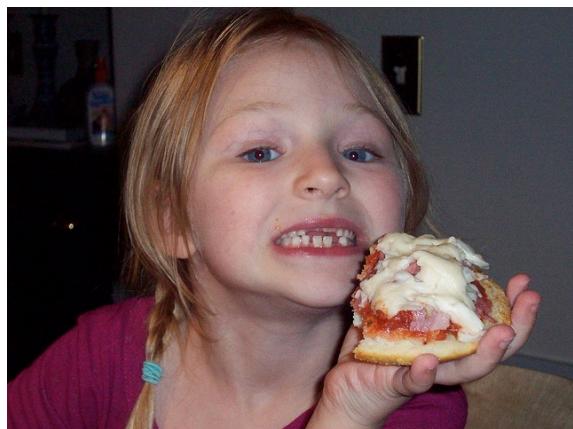
### ۱-۱-۱ نتایج خوب



شکل ۱-۵: گروهی از مردم در ساحل با بادبادک بازی می‌کنند.



شکل ۲-۵: گروهی از فیل‌ها که در مزرعه ایستاده بودند.



شکل ۳-۵: یه دختر کوچولو که یه تیکه پیتزا می‌خوره.



شکل ۴-۵: یک بازیکن تنیس روی زمین تنیس بازی می‌کند.



شکل ۵-۵: یک اسکی باز در حال اسکی روی یک شیب پوشیده از برف است.

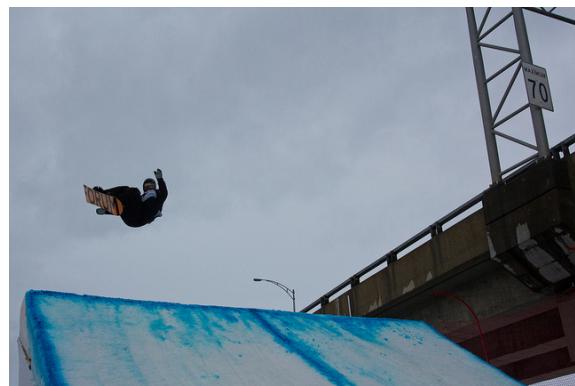


شکل ۵-۶: یک چراغ راهنمایی در کنار یک ساختمان بزرگ

## فصل ۵. نتایج

### ۱-۵. نتایج فارسی

#### **۵-۱-۲ نتایج بد**



شکل ۷-۵: یک هواپیما در آسمان آبی پرواز می‌کند.



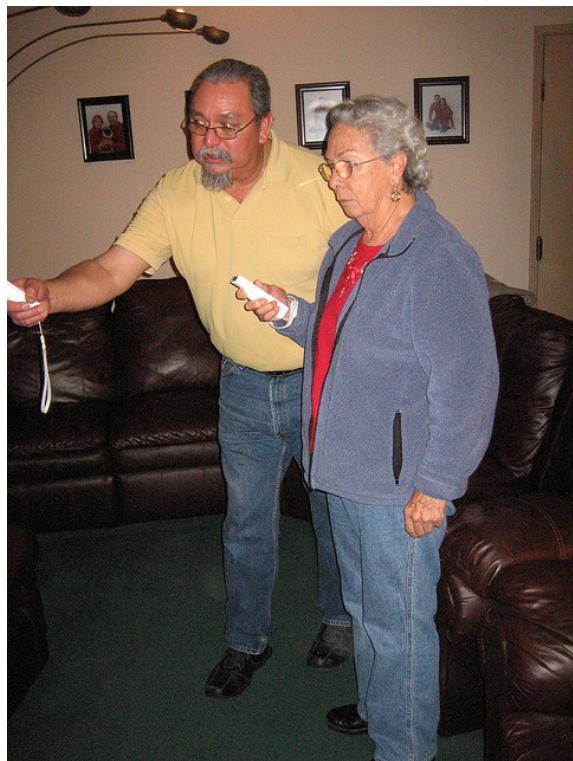
شکل ۸-۵: مردی که در آشپزخانه ایستاده بود.



شکل ۹-۵: یک میز پر از غذا و یک لیوان شراب.



شکل ۱۰-۵: یک حمام با یک یخچال سفید و یک ظرفشویی



شکل ۱۱-۵: مردی که روی کاناپه نشسته بود.



شکل ۱۲-۵: مردی که سوار بر اسب بود.

فصل ٥. نتایج

٥-٢. نتایج انگلیسی

**٥-٢ نتایج انگلیسی**

**٥-٢-١ نتایج خوب**



شكل ١٣ : a traffic light on the side of a street.



شكل ١٤ : a group of elephants standing next to each other.



شكل ١٥ : a train traveling down the tracks near a train station.



شكل ١٦ : a little girl eating a hot dog.



شكل ١٧ : a man swinging a tennis racket on a tennis court.

فصل ٥. نتایج

٥-٢. نتایج انگلیسی



شكل ١٨ : a vase filled with flowers on a table.

٥-٢-٢ نتایج بد



شكل ١٩ : a group of people seating on a bench.

## فصل ٥. نتایج

### ٢-٥. نتایج انگلیسی



شکل ٢٠ : a cat sitting on top of a wooden floor.



شکل ٢١ : a woman holding a hot dog in front of a table.

## فصل ٥. نتایج

### ٥-٢. نتایج انگلیسی



شكل ٥-٢٢ a group of elephants walking across a sandy beach.



شكل ٥-٢٣ a group of people flying kites in the sky.



شكل ٢٤ : a close up of a bowl of food.

## مراجع

- [1] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. Long-term recurrent convolutional networks for visual recognition and description.
- [2] Johnson, J., Karpathy, A., and Fei-Fei, L. DenseCap: Fully convolutional localization networks for dense captioning.
- [3] Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. Beyond short snippets: Deep networks for video classification.
- [4] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-CNN: Towards real-time object detection with region proposal networks.
- [5] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator.
- [6] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention.

# واژه‌نامه فارسی به انگلیسی

Captioning .....	توضیح نویسی .....
Captioning.....	عنوان نویسی .....
Encoder.....	رمزگذار .....
Decoder.....	رمزگشایی .....
Attention Mechanism.....	ساز و کار توجه .....
Abstraction Level.....	سطح انتزاع .....
Convolutional .....	پیچشی .....
Neural Network .....	شبکه‌ی عصبی .....
Deep Learning .....	یادگیری عمیق .....
Dataset .....	مجموعه دادگان .....

# واژه‌نامه انگلیسی به فارسی

Train .....	آموزش .....
Validation .....	اعتبارسنجی .....
Pooling .....	تجمع‌گیر .....
Embedding .....	نهفته سازی .....

**Abstract:**

Describing a scene or generating a caption for an image is a relatively easy task for us. However, when we are speaking in the artificial intelligence domain, generating such a caption automatically is a complex problem. In this thesis, a neural network model is presented which is based on deep learning and can generate a caption for a given image. This model consists of three parts: encoder, decoder, and attention. In order to generate captions both in Farsi and English, the model is separately trained on the original COCO dataset as well as the Persian-translated version of the dataset.

**Keywords:** Image captioning, Neural Network, Attention, Deep Learning



**Iran University of Science and Technology  
Computer Engineering Department**

# **Image captioning using Attention mechanism in neural network**

**A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree  
of Master of Science in Computer Engineering**

**By:**

Soroosh Gholamizoj

**Supervisor:**

**Dr. Sauleh Etemadi**

**July 2019**