**دانشگاه صنعتی امیرکبیر**

**(پلی‌تکنیک تهران)**

**دانشکده مهندسی کامپیوتر و فناوری اطلاعات**

**پروژه کارشناسی**

**گرایش نرم افزار**

# پرسش و پاسخ بصری با استفاده از شبکه های عصبی کانولوشنی عمیق و بازگشتی

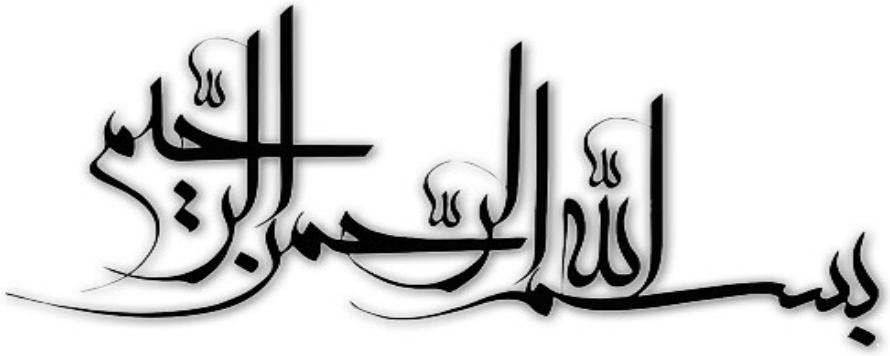
**نگارش**

**علی غلامی**

**استاد راهنما**

**دکتر محمد رحمتی**

**فروردین ۱۳۹۶**


بسم الله الرحمن الرحيم

## صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

**نکات مهم:**

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

**به نام خدا**

# تعهدنامه اصالت اثر

تاریخ: فروردین ۱۳۹۶

اینجانب **علی غلامی** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

**علی غلامی**

امضا

کیستی که من اینگونه به اعتماد
نام خود را
با تو می گویم...
کلید قلبم را
در دستانت می گذارم
نان شادی ام را با تو قسمت می کنم
به کنارت می نشینم
و سر بر شانه ی تو
این چنین آرام
به خواب می روم؟

# سپاس‌گزاری

از پدرم که چون کوهی استوار و مادرم که چون دریای محبت در فراز و نشیب زندگی دلسوزانه همراهم بوده اند؛
از استاد بزرگوار جناب آقای دکتر محمد رحمتی که در کمال سعه صدر، با حسن خلق و فروتنی، رهنمون من شده اند؛
و از سایر عزیزانی که در کنارشان این نتیجه حاصل آمد؛

کمال تشکر و قدردانی را دارم.

علی غلامی
فروردین ۱۳۹۶

# چکیده

در این پروژه، مدلی مبتنی بر شبکه های عصبی کانولوشنی و بازگشتی عمیق، به منظور پرسش و پاسخ بصری ارائه شده است. درک ماشین از تصاویری که به آن ارائه می شود، هیچ گاه با درک انسان از تصاویر قابل قیاس نبوده است. به همین دلیل، پژوهش های انجام گرفته در بینایی ماشین، همواره در صدد بهبود این قابلیت در ماشین ها بوده است. برای این منظور، ویژگی های مهم تصاویر استخراج شده و سپس در مورد وجود یا عدم وجود اشیا مورد نظر در تصاویر تصمیم گیری انجام می گیرد. فراتر از این موضوع، قابلیت درک ماشین از سوالات انسان درباره این تصاویر و پاسخ به این سوالات بر اساس ویژگی ها و دانش مستخرج از تصاویر است که به تازگی مورد توجه پژوهشگران عرصه ی بینایی ماشین قرار گرفته است.

**واژه‌های کلیدی:**

پرسش و پاسخ بصری، شبکه های کانولوشنی عمیق، شبکه های بازگشتی عمیق، پردازش زبان طبیعی، پردازش تصویر)

# فهرست مطالب

# فصل اول

# مقدمه

## ۱-۱ مقدمه

درک انسان از محیط اطراف خود، از ابتدای کودکی شکل می گیرد. اگر از یک کودک، درباره جهانی که برای او قابل رویت است، سوالی پرسیده شود، بلافاصله درباره ی این جهان جملاتی توصیفی با دقت بسیار بالا ارائه می کند. انتقال این قدرت به ماشین ها از این جهات مختلفی حائز اهمیت می باشد. یکی از جهات اصلی آن، کاربرد های بی شماری است که می توان از این قابلیت در امور مختلفی استفاده کرد. دوربین های نظارتی، خودروهای خودران و ... نمونه هایی از این کاربرد ها هستند. در این فصل ابتدا موضوع پژوهش را به طور کامل بیان کرده و اهمیت ارائه راهکار مناسب در این مورد را بررسی می کنیم. سپس رویکرد های مختلف را برای حل این مساله بیان می کنیم.

## ۱-۲ توضیح مساله

درک تصاویر به همراه درک سوالاتی که مرتبط با آن تصاویر پرسیده می شود، می تواند یکی از اساسی ترین قابلیت های یک سامانه مدیریت هوشمند تصاویر، خودروی خودران و یا یک سیستم بازیابی تصویر هوشمند باشد. برای این منظور نیاز است که ابتدا صحنه به نمایش درآمده توسط ماشین هضم گردد. بدین معنی که بدون درک درست از آنچه در تصویر موجود است، ساخت چنین سامانه ای امکان پذیر نخواهد بود.

## ۱-۳ مروری بر روند پژوهش های پیشین

روش هایی که جهت حل مسائل پرسش و پاسخ بصری ارائه شده اند از تنوع بالایی برخوردار هستند. اگرچه، می توان اغلب این روش ها را در راستای حل چهار چالش مهم زیر در نظر گرفت:

۱. چالش استخراج ویژگی از تصاویر
۲. چالش استخراج ویژگی از متن(سوالات پرسیده شده)
۳. چالش ترکیب بردار های ویژگی مستخرج از تصویر و متن
۴. چالش تولید پاسخ متناسب با ویژگی ها ترکیب شده

با تحول عظیمی که در سال ۲۰۱۲ با ارائه مدل کانولوشنی عمیق جهت استخراج ویژگی ارائه شد و نیز افزایش قدرت پردازشی ماشین ها، توجه به سمت شبکه های کانولوشنی جهت استخراج ویژگی از تصاویر بیشتر شد. با پیشرفت این مدل ها و ترکیب آنها با مدل های پردازش زبان طبیعی، تولید شرح بر تصاویر نیز مورد توجه قرار گرفت. از سال ۲۰۱۵ تا به اکنون، یکی از جالب ترین موضوعاتی که توجه پژوهشگران را به سمت خود جلب کرده است، موضوع پرسش و پاسخ بصری می باشد. این موضوع ارتباط تنگاتنگی با موضوع تولید شرح بر تصاویر دارد. به همین منظور، بسیاری از تکنییک های رایج در بحث شرح بر تصاویر، در موضوع پرسش و پاسخ بصری نیز مورد توجه قرار گرفته است.

# فصل دوم

# مروری بر مطالعات گذشته

## ۲-۱ مقدمه

دراین بخش به بررسی راهکار های مختلف ارائه شده در موضوع پرسش و پاسخ بصری می پردازیم. پرسش و پاسخ بصری، یکی از چالش های بزرگ روز در عرصه ی هوش مصنوعی می باشد. بسیاری از ایده های مطرح در حل این چالش ها، از نجوه ی عملکرد ذهن انسان مدل برداری شده است.

### ۲-۱-۱ مغز چگونه تصاویر را درک می کند؟

تصاویر در مغز انسان به طرز خارق العاده ی پردازش می شوند. به گونه ای که در اولین نگاه می توان بیشترین اطلاعات موجود در تصویر را استخراج کرد و آنها را توصیف نمود. بعلاوه، قدرت مغز در تخلیل شنیدار و پاسخ به این تحلیل و نیز صحنه ی دریافتی از طریق بینایی، همواره قابل تامل و مطالعه می باشد.
این ایده که مغز قادر است تا حجم زیادی از اطلاعات را به سرعت پردازش کرده و در مورد آنها تصمیم اتخاذ کند، از جانب پژوهشگران مورد بررسی قرار گرفته است. به عنوان مثال [۲] پژوهشی است که در آن تعدادی از تصاویر به صورت دنباله ای به افرادی نشان داده می شود و نیز توصیفاتی از طرف آنها ارائه می گردد. پژوهشگران با انجام این آزمایش پی بردند که مغز قادر است در کمتر از ۲۰۰ میلی ثانیه به صحنه های دریافتی پاسخ دهد. در پژوهش [۱] آزمایش دیگری انجام شده است که از اهمیت بسیاری بر خوردار است. در پژوهش های قبلی، افرادی که تصاویر را توصیف می کردند، درباره موضوع کلی تصاویر اطلاعاتی داشتند.اما در این آزمایش، تصاویر مختلفی از دنیای واقعی که محدود به شرایط خاصی نبوده اند، بدون ارائه پیش فرض درباره ی موضوع، به افراد نمایش داده شده و از آنها خواسته شده که تصویر را به بهترین شکل توصیف کنند. نتایج بدست آمده به صورت زیر می باشد:
۱. حداکثر زمان لازم برای مغز انسان به منظور درک صحنه، برابر با ۵۰۰ میلی ثانیه می باشد.
۲. این مدت زمان، برای صحنه های ساده و بدون پیچیدگی، به حدود ۱۰۰ میلی ثانیه می رسد.

# فصل سوم

# روش ارائه شده در پژوهش

## ۳-۱ مقدمه

در این فصل باید روشی جهت پیاده سازی یا نحوه ی انجام تحقیقات صورت گرفته ارائه گردد. این روش باید جزییات ویژه ای را پوشش دهد، چالش های مطرح را حل و فصل و نیز نوآوری هایی ارائه دهد. بنابراین، ابتدا به مشکلات اصلی موجود در این روش و سپس به ارائه راه حل های آن می پردازیم.

## ۳-۲ بیان مشکلات موجود

۱. تعداد زیاد پارامتر های شبکه عصبی و نیاز به توان پردازشی بالا تعداد پارامتر های یک شبکه ی عصبی بسیار بالاست. این تعداد در مدل الکسنت به ۱۵ میلیون پارامتر هم می رسد. جالب آن است که بیشتر از ۹۵ درصد این پارامتر ها مربوط به ۳ لایه ی تمام متصل انتهای شبکه می باشد. در صورتی که اندازه و کیفیت تصاویر نیز بالاتر رود، تعداد این پارامتر ها به صورت تصاعدی افزایش می یابد. در بحث استخراج ویژگی از تصاویر جهت دریافت جزییات صحنه، ارائه راهکار جدید برای معماری شبکه عصبی همواره مورد توجه پژوهشگران بوده است.
۲. یکی دیگر از مشکلات موجود، نحوه ی ارزیابی جملات و پاسخ هایی است که ماشین تولید می کند. این پاسخ ها اغلب با توصیفاتی که انسان در مورد تصاویر ارائه می دهد، از لحاظ جزییات بسیار فاصله دارد. به عنوان مثال، در یکی از معیار های مطرح شده برای ارزیابی پاسخ ها، جملات تولید شده توسط ماشین توسط سه داور انسانی بررسی میشود و هریک نمره ای به آن جمله می دهد. بر این اساس دقت جملات تولید شده ارزیابی می گردد. هرچند، این روش از دقت ارزیابی پایینی برخوردار می باشد. بنابراین یکی از چالش ها همواره، ایجاد یک راهکار دقیقتر برای ارزیابی بوده است.

# فصل چهارم
# پیاده سازی، آزمون و ارزیابی

در این فصل از گزارش، ابتدا کلیاتی در رابطه با پیاده سازی پروژه را بیان خواهیم نمود. بعلاوه، مجموعه داده مورد استفاده در آموزش و ارزیابی شبکه را بررسی کرده و همچنین مدل پیشنهادی برای ارزیابی دقت پاسخ های تولیدی را بررسی خواهیم کرد.

## ۴-۱ پیاده سازی

سیستم پرسش و پاسخ بصری مطرح در این گزارش، به زبان پایتون(نسخه ۶.۳) و با استفاده از چارچوب کاری تنسورفلو پیاده سازی شده است. هسته این چارچوب کاری به زبان سی پلاس پلاس و با استفاده از پلتفرم توسعه موازی کودا پیاده سازی شده است. این چارچوب کاری توسط تیم گوگل برین در حال توسعه بوده و پشتیبانی می شود.
ایده ی اصلی مطرح در این چارچوب کاری، بیان محاسبات در قالب گراف است. هر گره ی این گراف، یک واحد محاسباتی را مشخص می کند. با این رویکرد می توان شبکه های پیچیده را به راحتی پیاده سازی نمود.

## ۴-۲ معیار بلیو

این معیار، یکی از معیار های ارزیابی مدل های ترجمه ماشینی است که در حوزه پرسش و پاسخ بصری نیز مورد استفاده قرار می گیرد. در حوزه ترجمه ماشینی، ترجمه های مختلفی از یک جمله در زبان مبدا، می توان در زبان مقصد ارائه داد. برای تشخیص بهترین ترجمه بین ترجمه های کاندید برای یک حمله در زبا ن مبدا، می توان از این معیار استفاده نمود.

## ۴-۳ معیار سایدر

این معیار در بین پژوهشگران حوزه تولید خودکار شرح بر تصاویر ارائه شده است. این معیار در سال ۲۰۱۵ توسط ودانتام و همکارانش ارائه شد. هدف اصلی این معیار این است که توافق شرح های مرجع تولید شده توسط انسان را یافته و سپس میزان انطباق شرح تولید شده خودکار با این توافق را اندازه گیری نماید.

# فصل پنجم

# جمع‌بندی و نتیجه‌گیری و پیشنهادات

تولید و ذخیره سازی روزافزون تصاویر، سهولت در استفاده از دوربین های تصویر برداری و گوشی های موبایل، دسترسی آسان به اینترنت در تمام نقاط شهر و افزایش تعداد شبکه های اجتماعی و نرم افزارهای موبایل، باعث افزایش نیاز کاربران به سامانه های هوشمند مدیریت تصاویر شده است. سامانه هایی که علاوه بر مدیریت ذخیره و بازیابی تصاویر، قدرت دسته بندی خودکار، جستجوی محتوایی، درک و توصیف تصاویر از هر موضوعی باشند. ارائه مدل های هوشمند که بتوانند به طور خودکار برای هر تصویری، توصیف متناظر در قالب جملات زبان طبیعی تولید کنند، از جمله مهم ترین اقدامات در راستای رسیدن به سامانه مدیریت تصاویر به شمار می رود.

در سال های بعد از ۲۰۰۷، می توان گفت توجه پژوهش گران بیشتر به سمت مدل های تولید جمله جلب شد و چالش های موجود دراین حوزه که اغلب بدون راه حل بودند یا با راه حل های ابتدایی حل می شدند، بیش از پیش مورد استقبال پژوهش گران قرار گرفتند. مدل های مختلفی برای تولید جمله به کار گرفته شد. از جمله این مدل ها می توان به روش های موجود در حوزه تولید زبان طبیعی، بازیابی شبیه ترین جمله موجود در مجموعه داده و استفاده از کلیشه زبانی، اشاره کرد. اما هیچ یک از این روش ها، نتوانستند تمام معضلات را حل نمایند.

اما با حل مشکل ناپایداری آموزش شبکه های عصبی بازگشتی در سال ۲۰۱۱ توسط هینتون، فصل جدیدی در حوزه تولید جمله در این مساله شروع شد. شبکه های عصبی بازگشتی، ابزارهای قدرتمندی در کاربرد پیش بینی دنباله های زمانی و تولید جمله به شمار می روند. قابلیت های بالای این مدل ها، پژوهش گران را برآن داشت که تمامی روش های گذشته را کنار گذاشته و تماما از شبکه های عصبی بازگشتی برای تولید جمله استفاده نمایند. استفاده از شبکه های عصبی بازگشتی، ذهن اغلب پژوهش گران را به سمت استفاده از شبکه های کانولوشنی عمیق در مرحله استخراج اطلاعات از تصاویر می کشاند. شبکه های عصبی کانولوشنی عمیق، در استخراج ویژگی های بسیار خوب از تصاویر، قدرت بالایی دارند. از حدود سال ۲۰۱۴ به بعد و با جابجایی مدل های گرافی احتمالی با شبکه های عصبی کانولوشنی عمیق، صفر تا صد فرایند تولید خودکار شرح بر تصاویر، با استفاده از شبکه های عصبی و یادگیری عمیق انجام می شد.

# فصل ششم

# نحوه ی آماده سازی

هدف از ایجاد این فصل، توضیح و ارائه ی نحوه ی پیاده سازی این بخش و بخش های قبلی می باشد. برای کامپایل کردن این پایان نامه، علاوه بر آنکه وقت زیادی صرف شده است، دقت زیادی نیز به کار رفته است. در فصل های گذشته، محتویات هر بخش از فایل مربوط به آن بخش ویرایش شده است و مطالب مناسب اضافه شده اند. علی رغم وجود مطالب اندک، سعی بر آن بوده است جزییات را تا حدی که اینجانب به پرسش و پاسخ بصری تسلط دارم، در بر گیرد.

# منابع و مراجع


[1] Fei-Fei, Li, Iyer, Asha, Koch, Christof, and Perona, Pietro. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10–10, 2007.

[2] Potter, Mary C. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.