دانشگاه صنعتی امیرکبیر
( پلی تکنیک تهران )

# تولید خودکار شرح بر تصاویر با استفاده از شبکه‌های عصبی کانولوشنی عمیق و بازگشتی

Automatic Image Captioning Using Deep Convolutional and Recurrent Neural Networks

**استاد راهنما**

دکتر صفابخش

**پژوهش‌گر**

احمد اسدی

۹۴۱۳۱۰۹۱

اردیبهشت‌ماه ۱۳۹۵

# فهرست مطالب

# ۱ فصل اول

## مقدمه

## ۱.۱ مقدمه

به دنبال پیشرفت تکنولوژی در ساخت دوربین‌های عکاسی و ورود دوربین‌های نیمه‌خودکار و خودکار به بازار، تعداد زیادی از کاربران سیستم‌های رایانه‌ای به استفاده از این تکنولوژی در ثبت تصاویر مورد علاقه خود جذب شده‌اند. دقت و کیفیت مطلوب تصویربرداری از یک سو و سهولت استفاده از دوربین از سوی دیگر، باعث شده‌اند تعداد تصاویر ثبت شده توسط کاربران به طور روزافزون افزایش یابد؛ به‌طوری‌که امروزه اغلب کاربران، تعداد بی‌شماری از این تصاویر را در گوشی‌های تلفن همراه، تبلت‌ها و رایانه‌های شخصی خود نگه‌داری می‌کنند. از جمله مشکلاتی که در اثر ایجاد این حجم وسیع از تصاویر بوجود آمده، مشکل مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه بزرگی از تصاویر موجود، است.

برای دست‌یابی به سامانه‌ای که بتواند تعداد زیادی از تصاویر موجود را مدیریت نماید، ابتدا باید صحنه موجود در تصویر را به درستی درک کرد. درک صحیح از صحنه، عبارت است از بیان تصویر به نحوی که اطلاعات کلی موجود و هدف اصلی تصویر، واضح و مشخص باشد. این بیان می‌تواند شامل اجسام موجود در تصویر، رابطه مکانی بین اجسام، فعالیت به تصویر کشیده شده، شرایط محیطی موثر بر صحنه و مواردی از این دست باشد. از طرفی باید به نحوی محتوای تصاویر را بیان کرد که بتوان عملیات جستجو را بر اساس مدل بیان شده تصاویر انجام داد. در این‌صورت به‌ازای هر تصویر، یک نمونه از مدل مطابق با تصویر ایجاد و ذخیره خواهد شد. پرس‌وجوی[۱] کاربر، به فضای مدل نگاشت شده و تصویر معادل با مدلِ استخراج شده، به عنوان نتیجه جستجو نمایش داده می‌شود. علاوه بر این، مساله مدیریت تصاویر، به مساله مدیریت مدل‌های موجود کاهش داده می‌شود.

تولید شرح کلی بر تصاویر[۲]، بیان مناسبی از صحنه موجود در تصویر را ارائه می‌دهد. شرح تولید شده بر تصاویر، در قالب مجموعه‌ای از جملات زبان طبیعی[۳] ارائه می‌شود که عموما بیان‌گر اجسام موجود در صحنه، ارتباطات مکانی بین اجسام و اطلاعات مشخص دیگر است که در هر پژوهش می‌تواند متفاوت باشد. بنابراین، دست‌یابی به سامانه‌ای که قادر به تولید خودکار شرح کلی بر تصاویر باشد، اساسی‌ترین گام در راستای تولید نرم‌افزارهای مدیریت تصاویر است.

یکی از اولین ایده‌های مطرح شده در این زمینه، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۴] به‌وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

[۱] Query
[۲] Holistic Image Caption
[۳] Natural Language Sentences
[۴] Machine Translation

یک جمله از زبان مبدا[۵]، با روش‌های مختلف تبدیل به یک بردار ویژگی[۶] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد[۷] میگردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند. با توجه به فرایند مذکور، اگر به جای جمله زبان مبدا، یک تصویر را به بردار ویژگی تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی را به جمله زبان مقصد ترجمه نمود، جمله‌ای معادل با تصویر ورودی به‌دست خواهد آمد. که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است.

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده است. علی‌رغم وجود پژوهش‌های فراوان و متفاوت، می‌توان یک بستر کلی برای تمام فعالیت‌های موجود در این زمینه ارائه داد. بر این مبنا، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی گرفته شده‌است، از دو بخش اساسی تشکیل می‌شود.

۱. بازنمایی تصاویر، با استفاده از بردار ویژگی

۲. تبدیل بردار ویژگی به‌دست‌آمده به جملات صحیح زبانی

[۵] Source Language

[۶] Feature Vector

[۷] Destination Language

# ۲ فصل دوم

# درک صحنه

## ۱.۲ شرح بر تصاویر

این یک بخش آزمایشی است

# ۳ فصل سوم

# تولید شرح متناظر صحنه

## ۱.۳ مقدمه

به دنبال پیشرفت تکنولوژی در ساخت دوربین‌های عکاسی و ورود دوربین‌های نیمه‌خودکار و خودکار به بازار، تعداد زیادی از کاربران سیستم‌های رایانه‌ای به استفاده از این تکنولوژی در ثبت تصاویر مورد علاقه خود جذب شده‌اند. دقت و کیفیت مطلوب تصویربرداری از یک سو و سهولت استفاده از دوربین از سوی دیگر، باعث شده‌اند تعداد تصاویر ثبت شده توسط کاربران به طور روزافزون افزایش یابد؛ به‌طوری‌که امروزه اغلب کاربران، تعداد بی‌شماری از این تصاویر را در گوشی‌های تلفن همراه، تبلت‌ها و رایانه‌های شخصی خود نگه‌داری می‌کنند. از جمله مشکلاتی که در اثر ایجاد این حجم وسیع از تصاویر بوجود آمده، مشکل مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه بزرگی از تصاویر موجود، است.

برای دست‌یابی به سامانه‌ای که بتواند تعداد زیادی از تصاویر موجود را مدیریت نماید، ابتدا باید صحنه موجود در تصویر را به درستی درک کرد. درک صحیح از صحنه، عبارت است از بیان تصویر به نحوی که اطلاعات کلی موجود و هدف اصلی تصویر، واضح و مشخص باشد. این بیان می‌تواند شامل اجسام موجود در تصویر، رابطه مکانی بین اجسام، فعالیت به تصویر کشیده شده، شرایط محیطی موثر بر صحنه و مواردی از این دست باشد. از طرفی باید به نحوی محتوای تصاویر را بیان کرد که بتوان عملیات جستجو را بر اساس مدل بیان شده تصاویر انجام داد. در این‌صورت به‌ازای هر تصویر، یک نمونه از مدل مطابق با تصویر ایجاد و ذخیره خواهد شد. پرس‌وجوی[۸] کاربر، به فضای مدل نگاشت شده و تصویر معادل با مدلِ استخراج شده، به عنوان نتیجه جستجو نمایش داده می‌شود. علاوه بر این، مساله مدیریت تصاویر، به مساله مدیریت مدل‌های موجود کاهش داده می‌شود.

تولید شرح کلی بر تصاویر[۹]، بیان مناسبی از صحنه موجود در تصویر را ارائه می‌دهد. شرح تولید شده بر تصاویر، در قالب مجموعه‌ای از جملات زبان طبیعی[۱۰] ارائه می‌شود که عموما بیان‌گر اجسام موجود در صحنه، ارتباطات مکانی بین اجسام و اطلاعات مشخص دیگر است که در هر پژوهش می‌تواند متفاوت باشد. بنابراین، دست‌یابی به سامانه‌ای که قادر به تولید خودکار شرح کلی بر تصاویر باشد، اساسی‌ترین گام در راستای تولید نرم‌افزارهای مدیریت تصاویر است.

یکی از اولین ایده‌های مطرح شده در این زمینه، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۱۱] به‌وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

[۸] Query
[۹] Holistic Image Caption
[۱۰] Natural Language Sentences
[۱۱] Machine Translation

یک جمله از زبان مبدا[۱۲]، با روش‌های مختلف تبدیل به یک بردار ویژگی[۱۳] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد[۱۴] میگردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند. با توجه به فرایند مذکور، اگر به جای جمله زبان مبدا، یک تصویر را به بردار ویژگی تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی را به جمله زبان مقصد ترجمه نمود، جمله‌ای معادل با تصویر ورودی به‌دست خواهد آمد. که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است.

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده است. علی‌رغم وجود پژوهش‌های فراوان و متفاوت، می‌توان یک بستر کلی برای تمام فعالیت‌های موجود در این زمینه ارائه داد. بر این مبنا، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی گرفته شده‌است، از دو بخش اساسی تشکیل می‌شود.

۱. بازنمایی تصاویر، با استفاده از بردار ویژگی

۲. تبدیل بردار ویژگی به‌دست‌آمده به جملات صحیح زبانی

[۱۲] Source Language

[۱۳] Feature Vector

[۱۴] Destination Language

# ۴ فصل چهارم

# آزمون و ارزیابی

## ۱.۴ مقدمه

به دنبال پیشرفت تکنولوژی در ساخت دوربین‌های عکاسی و ورود دوربین‌های نیمه‌خودکار و خودکار به بازار، تعداد زیادی از کاربران سیستم‌های رایانه‌ای به استفاده از این تکنولوژی در ثبت تصاویر مورد علاقه خود جذب شده‌اند. دقت و کیفیت مطلوب تصویربرداری از یک سو و سهولت استفاده از دوربین از سوی دیگر، باعث شده‌اند تعداد تصاویر ثبت شده توسط کاربران به طور روزافزون افزایش یابد؛ به‌طوری‌که امروزه اغلب کاربران، تعداد بی‌شماری از این تصاویر را در گوشی‌های تلفن همراه، تبلت‌ها و رایانه‌های شخصی خود نگه‌داری می‌کنند. از جمله مشکلاتی که در اثر ایجاد این حجم وسیع از تصاویر بوجود آمده، مشکل مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه بزرگی از تصاویر موجود، است.

برای دست‌یابی به سامانه‌ای که بتواند تعداد زیادی از تصاویر موجود را مدیریت نماید، ابتدا باید صحنه موجود در تصویر را به درستی درک کرد. درک صحیح از صحنه، عبارت است از بیان تصویر به نحوی که اطلاعات کلی موجود و هدف اصلی تصویر، واضح و مشخص باشد. این بیان می‌تواند شامل اجسام موجود در تصویر، رابطه مکانی بین اجسام، فعالیت به تصویر کشیده شده، شرایط محیطی موثر بر صحنه و مواردی از این دست باشد. از طرفی باید به نحوی محتوای تصاویر را بیان کرد که بتوان عملیات جستجو را بر اساس مدل بیان شده تصاویر انجام داد. در این‌صورت به‌ازای هر تصویر، یک نمونه از مدل مطابق با تصویر ایجاد و ذخیره خواهد شد. پرس‌وجوی[۱۵] کاربر، به فضای مدل نگاشت شده و تصویر معادل با مدلِ استخراج شده، به عنوان نتیجه جستجو نمایش داده می‌شود. علاوه بر این، مساله مدیریت تصاویر، به مساله مدیریت مدل‌های موجود کاهش داده می‌شود.

تولید شرح کلی بر تصاویر[۱۶]، بیان مناسبی از صحنه موجود در تصویر را ارائه می‌دهد. شرح تولید شده بر تصاویر، در قالب مجموعه‌ای از جملات زبان طبیعی[۱۷] ارائه می‌شود که عموما بیان‌گر اجسام موجود در صحنه، ارتباطات مکانی بین اجسام و اطلاعات مشخص دیگر است که در هر پژوهش می‌تواند متفاوت باشد. بنابراین، دست‌یابی به سامانه‌ای که قادر به تولید خودکار شرح کلی بر تصاویر باشد، اساسی‌ترین گام در راستای تولید نرم‌افزارهای مدیریت تصاویر است.

یکی از اولین ایده‌های مطرح شده در این زمینه، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۱۸] به‌وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

[۱۵] Query
[۱۶] Holistic Image Caption
[۱۷] Natural Language Sentences
[۱۸] Machine Translation

یک جمله از زبان مبدا[۱۹]، با روش‌های مختلف تبدیل به یک بردار ویژگی[۲۰] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد[۲۱] میگردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند. با توجه به فرایند مذکور، اگر به جای جمله زبان مبدا، یک تصویر را به بردار ویژگی تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی را به جمله زبان مقصد ترجمه نمود، جمله‌ای معادل با تصویر ورودی به‌دست خواهد آمد. که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است.

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده است. علی‌رغم وجود پژوهش‌های فراوان و متفاوت، می‌توان یک بستر کلی برای تمام فعالیت‌های موجود در این زمینه ارائه داد. بر این مبنا، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی گرفته شده‌است، از دو بخش اساسی تشکیل می‌شود.

۱. بازنمایی تصاویر، با استفاده از بردار ویژگی

۲. تبدیل بردار ویژگی به‌دست‌آمده به جملات صحیح زبانی

[۱۹] Source Language

[۲۰] Feature Vector

[۲۱] Destination Language

# ۵ فصل پنجم

# جمع‌بندی و نتیجه‌گیری

## ۱.۵ مقدمه

به دنبال پیشرفت تکنولوژی در ساخت دوربین‌های عکاسی و ورود دوربین‌های نیمه‌خودکار و خودکار به بازار، تعداد زیادی از کاربران سیستم‌های رایانه‌ای به استفاده از این تکنولوژی در ثبت تصاویر مورد علاقه خود جذب شده‌اند. دقت و کیفیت مطلوب تصویربرداری از یک سو و سهولت استفاده از دوربین از سوی دیگر، باعث شده‌اند تعداد تصاویر ثبت شده توسط کاربران به طور روزافزون افزایش یابد؛ به‌طوری‌که امروزه اغلب کاربران، تعداد بی‌شماری از این تصاویر را در گوشی‌های تلفن همراه، تبلت‌ها و رایانه‌های شخصی خود نگه‌داری می‌کنند. از جمله مشکلاتی که در اثر ایجاد این حجم وسیع از تصاویر بوجود آمده، مشکل مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه بزرگی از تصاویر موجود، است.

برای دست‌یابی به سامانه‌ای که بتواند تعداد زیادی از تصاویر موجود را مدیریت نماید، ابتدا باید صحنه موجود در تصویر را به درستی درک کرد. درک صحیح از صحنه، عبارت است از بیان تصویر به نحوی که اطلاعات کلی موجود و هدف اصلی تصویر، واضح و مشخص باشد. این بیان می‌تواند شامل اجسام موجود در تصویر، رابطه مکانی بین اجسام، فعالیت به تصویر کشیده شده، شرایط محیطی موثر بر صحنه و مواردی از این دست باشد. از طرفی باید به نحوی محتوای تصاویر را بیان کرد که بتوان عملیات جستجو را بر اساس مدل بیان شده تصاویر انجام داد. در این‌صورت به‌ازای هر تصویر، یک نمونه از مدل مطابق با تصویر ایجاد و ذخیره خواهد شد. پرس‌وجوی[۲۲] کاربر، به فضای مدل نگاشت شده و تصویر معادل با مدلِ استخراج شده، به عنوان نتیجه جستجو نمایش داده می‌شود. علاوه بر این، مساله مدیریت تصاویر، به مساله مدیریت مدل‌های موجود کاهش داده می‌شود.

تولید شرح کلی بر تصاویر[۲۳]، بیان مناسبی از صحنه موجود در تصویر را ارائه می‌دهد. شرح تولید شده بر تصاویر، در قالب مجموعه‌ای از جملات زبان طبیعی[۲۴] ارائه می‌شود که عموما بیان‌گر اجسام موجود در صحنه، ارتباطات مکانی بین اجسام و اطلاعات مشخص دیگر است که در هر پژوهش می‌تواند متفاوت باشد. بنابراین، دست‌یابی به سامانه‌ای که قادر به تولید خودکار شرح کلی بر تصاویر باشد، اساسی‌ترین گام در راستای تولید نرم‌افزارهای مدیریت تصاویر است.

یکی از اولین ایده‌های مطرح شده در این زمینه، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۲۵] به‌وجود آمده است که با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا،

[۲۲] Query
[۲۳] Holistic Image Caption
[۲۴] Natural Language Sentences
[۲۵] Machine Translation

یک جمله از زبان مبدا[۲۶]، با روش‌های مختلف تبدیل به یک بردار ویژگی[۲۷] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد[۲۸] میگردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند. با توجه به فرایند مذکور، اگر به جای جمله زبان مبدا، یک تصویر را به بردار ویژگی تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی را به جمله زبان مقصد ترجمه نمود، جمله‌ای معادل با تصویر ورودی به‌دست خواهد آمد. که بیان‌گر محتوای به تصویر کشیده شده در تصویر ورودی است.

شرح خودکار تصاویر، توجه پژوهش‌گران بسیار زیادی را به خود جلب کرده است و فعالیت‌های متنوع و متعددی در این راستا انجام شده است. علی‌رغم وجود پژوهش‌های فراوان و متفاوت، می‌توان یک بستر کلی برای تمام فعالیت‌های موجود در این زمینه ارائه داد. بر این مبنا، فرایند کلی که در عموم پژوهش‌های انجام‌شده، پی گرفته شده‌است، از دو بخش اساسی تشکیل می‌شود.

۱. بازنمایی تصاویر، با استفاده از بردار ویژگی

۲. تبدیل بردار ویژگی به‌دست‌آمده به جملات صحیح زبانی

[۲۶] Source Language

[۲۷] Feature Vector

[۲۸] Destination Language