بهبود شبکه‌های رقابتی مولد برای تولید خودکار تصویر از روی متن

الهام پژهان [۱و۲] و محمد قاسم‌زاده [۱*]

[۱] دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

[۲] دانشکده علوم کامپیوتر، دانشگاه کپنهاگ، کپنهاگ، دانمارک

چکیده

این پژوهش در رابطه با به‌کارگیری ابزارهای یادگیری عمیق و فناوری پردازش تصویر در تولید خودکار تصویر از روی متن است. تولید تصویر از روی متن یک مساله پیچیده در حوزه بینایی ماشین و پردازش زبان طبیعی به‌حساب می‌آید. در این رابطه دو هدف اصلی دنبال می‌شود. نخست این‌که تصویر تولیدشده بایستی تا حدممکن واقعی به نظر برسد؛ و دوم اینکه تصویر تولیدشده، توصیفی معنادار از متن ورودی باشد. پژوهش‌های پیشین از یک جمله برای تولید تصاویر بهره می‌برند. در این پژوهش یک مدل سلسله‌مراتبی مبتنی بر حافظه ارائه شده‌است که از سه توصیف مختلف که در قالب جمله ارائه می‌شوند، برای تولید و بهبود تصویر استفاده می‌کند. طرح پیشنهادی با بهره‌گیری از شبکه‌های رقابتی مولد، بر به‌کارگیری اطلاعات بیشتر جهت تولید تصاویر با وضوح بالا تمرکز دارد. هدف این مقاله تمرکز بر به‌کارگیری اطلاعات بیشتر برای تولید تصاویری با وضوح بالاتر است. ساختار شبکه ارائه‌شده به این صورت است که ابتدا با استفاده از نخستین جمله، یک تصویر اولیه با وضوح پایین تولید می‌شود؛ سپس این تصویر بر اساس دو جمله بعدی، بهبود یافته و تصاویری با وضوح بالاتر تولید می‌شود. ساختار حافظه موجود، در هر گام شرایطی را فراهم می‌کند که هر بار اطلاعاتی که اهمیت بیشتری برای بهبود تصویر دارند، بر اساس سازوکار توجه، بازیابی شوند. پیاده‌سازی و اجرای برنامه‌های مربوط به این حوزه نیاز به منابع پردازشی بالا دارند؛ لذا طرح پیشنهادی با بهره‌گیری از بستره سخت‌افزاری دانشگاه کپنهاگ بر روی یک کلاس‌تر با ۲۵ واحد پردازش گرافیکی پیاده‌سازی و تحت آزمون قرار گرفت. آزمایش‌ها روی مجموعه‌دادگان CUB-200 و ids-ade انجام شدند.. نتایج آزمایش‌ها بر اساس دو معیار Inseption score و R-precision نشان می‌دهند که مدل ارائه شده می‌تواند تصاویر با کیفیت بالاتری نسبت به دو مدل پایه StackGAN و AttGAN تولید کند.

واژگان کلیدی: شبکه رقابتی مولد، یادگیری عمیق، مدل سلسله‌مراتبی، پردازش زبان طبیعی.

Improvement of generative adversarial networks for automatic text-to-image generation

Elham Pejhan [1,2] and Mohammad Ghasemzadeh [1*]

[1] Computer Engineering Department, Yazd University, Yazd, Iran.

[2] Computer Science Department, University of Copenhagen, Copenhagen, Denmark.

Abstract:

This research is related to the development of technology in the field of automatic text to image transformation; also known as image generation from text. In this regard, two main goals are pursued; first, the produced image must look as real as possible; and second, the produced image should be a meaningful description of the input text. In recent years, generative adversarial networks that are capable of producing a wide range of content such as images, text and audio, have been emerged. The problem of producing images from text is a complex task in the field of machine vision and natural language processing. With the advancement of new technologies, automatic image production from text has become especially important due to its application in various fields, such as automated content production. The basic methods for producing images from text actually used a combination of search

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۵/۳۱ • تاریخ پذیرش: ۱۴۰۰/۳/۳ • تاریخ انتشار: ۱۴۰۱/۱۲/۲۹ • نوع مطالعه: پژوهشی

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]

and supervised learning. These methods use the correlation between the image regions and the words in the text that can be depicted. These selected words are then used to retrieve the images that are related to them. The problem with this solution is that it cannot generate images with new content. For this reason, in recent years, some studies have been introduced for text to image generation based on GANs and deep convolutional neural networks. In 2016, Reed et al. represented a model for text-to-image generation using GANs for the first time. They were able to generate images of flowers and birds with a resolution of 64 x 64. However, their proposed method usually lacked precise details of objects such as bird eyes and they were not capable of producing higher resolution images, such as 128 x 128. Zheng et al. proposed StackGAN model that divides the problem into two smaller subproblems. In the first step, they generate a low-resolution image with the initial design and color of the objects based on the text. In the second step, the output of the previous step and the text are given as input to the system to improve the original image and produce a high-quality image. In 2019, Zhou and colleagues introduced the DM-GAN network. Their proposed model relies on dynamic key-value memory and focuses on improving the quality of the image produced in the first step. Primary image properties are used as the search key in the memory module. In memory units at each step, the words associated with the generated image are dynamically selected and written. The methods presented so far have used only one sentence to produce the initial image and also to improve it in the next steps. While the datasets used in this field contains at least five descriptions for each image. The proposed method is a Multi Sentences Hierarchical GAN (MSH-GAN) for text to image generation. In this paper, we have looked at two key options: 1) produce a higher quality image in the first step, and 2) use two additional descriptions to improve the original image in the next steps. Our goal is to focus on using more information to produce higher resolution images. The structure of the network is in such a way that in the first stage, one sentence is used to generate an initial low-resolution image. Then in the next steps, the initial image improves based on the next two sentences, and the model generates higher-resolution images. The structure of the existing memory retrieves more important text information in each step to improve image quality based on the attention mechanism. Implementing programs related to this field require massive processing resources. Therefore, the proposed method was implemented and tested on a cluster with 25 GPUs using the hardware platform of the University of Copenhagen. The experiments were performed on CUB-200 and ids-ade datasets. The experimental results based on Inception score and R-precision evaluation metrics show that the proposed model can produce higher quality images than the two basic models StackGAN and AttGAN.

Keywords: Generative Adversarial Network, Deep Learning, Hierarchical Model, Natural Language Processing

۵- مقدمه

در حوزه پردازش تصاویر معروف است که «ارزش یک تصویر بیش از هزار کلمه است». از آنجایی که تصاویر می‌توانند وقایع را بهتر به نمایش بگذارند و تأثیر عمیق‌تر و ماندگارتری ایجاد کنند، از دیرباز برای تشریح مفاهیم و نمایش اطلاعات مورد توجه بوده‌اند. با پیشرفت فناوری‌های نوین، مسئله تولید خودکار تصویر از روی متن، به‌جهت کاربرد آن در حوزه‌های مختلف، مانند تولید خودکار محتوا، اهمیت ویژه‌ای را به خود اختصاص داده است. این مسئله فصل مشترک حوزه‌های مختلف دانش و فناوری، از جمله بینایی ماشین و پردازش زبان طبیعی است [19].

یکی از رایج‌ترین و چالش برانگیزترین مسائل در حوزه پردازش زبان طبیعی و بینایی ماشین، توصیف‌گذاری خودکار تصاویر است. در این مسئله، هدف این است که از روی تصویر داده‌شده، یک توصیف متنی، به‌صورت خودکار تولید شود. از یک دیدگاه سطح بالا، مسئله توصیف‌گذاری و تولید تصویر از متن را می‌توان نمونه‌هایی از مسئله ترجمه زبانی به‌حساب آورد؛ بدین صورت که مفاهیم و اطلاعات مختلف را می‌توان به دو زبان مختلف متن و تصویر بیان و هر یک را به دیگری ترجمه کرد. با این وجود، این دو مسئله به‌طور کامل متفاوت با یک ترجمه زبانی هستند؛ در واقع تبدیل متن به تصویر و تبدیل تصویر به متن مسائل چندبعدی تلقی می‌شوند. برای مثال فرض کنید بخواهیم جمله ساده «این یک گل زیبا است» را به برای مثال زبان فرانسه ترجمه کنیم. در این حالت، تعداد محدودی جمله معتبر را می‌توان به‌عنوان ترجمه قابل قبول ارائه داد، در حالی که اگر بخواهیم یک تصویرِ متناسب با این جمله را تولید کنیم، ممکن است تعداد بی‌شماری تصویر با آن مطابقت کنند.

اگرچه این رفتار چندبعدی در مسئله توصیف‌گذاری تصویر نیز وجود دارد، اما به‌دلیل وجود پیوستگی در زبان، این مسئله ساده‌تر از مسئله تبدیل متن به تصویر است، چون برای تولید هر کلمه، از کلمات قبلی نیز می‌توان بهره برد؛ درحالی‌که چنین وضعیتی برای مسئله تبدیل متن به تصویر برقرار نیست.

در مسئله تولید تصویر از روی متن دو هدف اصلی

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

وجود دارد: ۱) تصویر تولیدشده بایستی درحدامکان واقعی به نظر برسد؛ ۲) تصویر تولیدشده، توصیفی معنادار از متن ورودی باشد. در سال‌های اخیر شبکه‌های رقابتی مولد[1] [4] ارائه‌شده‌اند که قادر به تولید دامنه‌هایی گسترده از محتوا مانند تصاویر، متن و صوت هستند. در این ساختار، یک شبکه عصبی عمیق، داده ساختگی تولید می‌کند و شبکه عمیق دیگری وظیفه تشخیص واقعی یا ساختگی‌بودن داده ورودی را به عهده دارد. دامنه عملکرد این شبکه‌ها می‌تواند مواردی چون خلق یک نقاشی، شعر و یا یک قطعه موسیقی قابل‌قبول باشد [1]. درواقع، اغلب مدل‌های ارائه‌شده برای مسئله تولید متن از روی تصویر، بر اساس شبکه‌های رقابتی مولد طراحی‌شده‌اند [18، 17، 10]. یکی از مدل‌های محوری، StackGAN نام دارد که از روشی سلسله‌مراتبی برای تولید و بهبود تصویر بهره می‌برد [17]. پس از آن مدل‌های دیگری نیز بر پایه روش‌های سلسله‌مراتبی ارائه‌شده‌اند [21 ، 14 ، 7] که سازوکار توجه[2] را نیز به آن اضافه کرده‌اند که قادر هستند تصاویری یک‌نواخت و با وضوح بیشتر را تولید کنند.

هرچند روش‌های سلسله‌مراتبی توانسته‌اند پیشرفت قابل‌توجهی را در این حوزه ایجاد کنند، اما همچنان با چالش‌هایی روبه‌رو هستند که در ادامه دو مورد اصلی بیان‌شده‌اند. نخست این‌که تصاویر تولیدشده نهایی، وابسته به تصویر تولیدشده در گام نخست هستند و روش سلسله‌مراتبی در صورتی که تصویر اولیه مناسب نباشد، موفقیت چندانی نخواهد داشت. دومین چالش این است که هر واژه در توصیف ورودی، سطح متفاوتی از اطلاعات را برای تصویر نهایی در بر دارد. اطلاعات بصری باید از میزان اهمیت هر واژه در هر گام برای بهبود تصویر آگاه باشند [10].

روش‌هایی که تاکنون ارائه‌شده‌اند تنها از یک جمله برای تولید تصویر اولیه و همچنین بهبود آن در گام‌های بعدی استفاده کرده‌اند. این موضوع در حالی است که مجموعه دادگان مورد استفاده در این حوزه برای هر تصویر حاوی دست کم پنج توصیف برای هر تصویر هستند.

در این مقاله یک روش سلسله‌مراتبی مبتنی بر شبکه رقابتی مولد بر پایه چند جمله به نام MSH-GAN[3] ارائه شده‌است. هدف این مقاله تمرکز بر به‌کارگیری اطلاعات بیشتر برای تولید تصاویری با وضوح بالاتر است. ساختار شبکه ارائه‌شده به این صورت است که ابتدا با استفاده از نخستین جمله، یک تصویر اولیه با وضوح پایین تولید می‌شود، سپس این تصویر بر اساس دو جمله بعدی، بهبود یافته و تصاویری با وضوح بالاتر تولید می‌شود. ساختار حافظه موجود، در هر گام شرایطی را فراهم می‌کند که هر بار اطلاعاتی که اهمیت بیشتری برای بهبود تصویر دارند، بر اساس سازوکار توجه، بازیابی شوند.

برای ارزیابی عملکرد مدل پیشنهادی MSH-GAN چند مجموعه آزمایش روی مجموعه‌دادگان CUB-200 [13] و ids-ade [6] انجام شدند و کیفیت تصاویر تولیدشده بر اساس معیارهای IS[4] [11] و R-precision [10] مورد بررسی قرار گرفته‌اند. نتایج آزمایش‌ها نشان‌می‌دهد که مدل ارائه‌شده می‌تواند تصاویر با کیفیت بالاتری نسبت به دو مدل پایه StackGAN [17] **و** AttGAN [14] تولید کند.

سهم اصلی این پژوهش سه مورد زیر است:

- یک شبکه رقابتی مولد سلسله‌مراتبی ترکیب‌شده با حافظه و بر مبنای استفاده از سه جمله برای تولید تصاویر با کیفیت بهتر ارائه کردیم.
- یک تابع هزینه متناسب با مدل پیشنهادی ارائه کردیم که بتواند به‌صورت دقیق‌تری، مرتبط‌بودن تصویر با متن را ارزیابی، و تصویر با کیفیت بالاتری تولید کند.
- بر روی مجموعه‌دادگان جدیدی با بیش از یک شیء و با پیچیدگی بیشتر تمرکز داشتیم.

برای تشریح مدل ارائه‌شده، در بخش ۲، شبکه‌های رقابتی مولد و ساختار مبتنی بر حافظه را معرفی می‌کنیم. در بخش ۳ به‌اختصار برخی پژوهش‌های انجام‌شده در این زمینه را توضیح خواهیم داد. روش پیشنهادی و جزئیات مربوط به آن در بخش ۴ و نحوه پیاده‌سازی و اجرای آزمایش‌ها در بخش ۵ آمده‌اند. درنهایت در بخش ۶ نتایج آزمایش‌ها و بحث روی آنها ارائه شده‌است.

۲- مبانی نظری پژوهش

در این بخش دانش پایه مورد نیاز برای درک بهتر مدل پیشنهادی، ارائه شده‌است.

۲-۱- شبکه‌های رقابتی مولد

شبکه‌های GAN، یک دسته بسیار مهم از شبکه‌های مولد هستند که در سال ۲۰۱۴ توسط *یان گودفلو* [4] مطرح

[1] Generative Adversarial networks (GANs)

[2] Attention

[3] Multi Sentences Hierarchical GAN (MSH-GAN)

[4] Inception Score

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]

شدند. این شبکه‌ها بر اساس رویکرد تئوری بازی‌ها بناگذاری شده‌اند که در آن یک شبکه یادگیری عمیق که مولد[1] D نامیده می‌شود با یک‌روند تخاصم با شبکه‌ای دیگر رقابت می‌کند. شبکه عمیق دیگر که متمایزکننده[2] G نامیده می‌شود، سعی دارد نمونه‌های تولیدشده از شبکه مولد را از داده‌های اصلی تشخیص دهد. رقابت بین این دو شبکه، در نهایت باعث یادگیری بهتر و بهبود عملکرد هر دو شبکه می‌شود. رابطه (۱) رقابت بین D و G را که گونه‌ای از بازی کمینه – بیشینه است، نشان‌می‌دهد.

$$V(D, G)=E_{x\sim p_{data}(x)}[\log D(x)]+E_{z\sim p_z(z)}[\log(1-D(G(z))]$$

در این رابطه، عبارت نخست آنتروپی داده واقعی است که از D عبور می‌کند و شبکه D سعی دارد آن را به بیشینه مقدار، یعنی یک برساند. در مقابل، عبارت دوم، آنتروپی داده تصادفی است که از G عبور می‌کند و شبکه D سعی دارد آن را به مقدار صفر نزدیک کند. وظیفه شبکه G به‌طور کامل عکس رفتار شبکه D بوده و تلاش می‌کند عبارت را کمینه کند.

شبکه‌های GAN را می‌توان به مدل شرطی نیز گسترش داد؛ در صورتی که هر دو شبکه مولد و شبکه متمایزکننده شامل اطلاعات اضافی دیگری باشند. شرط می‌تواند هر نوع اطلاعات کمکی مانند برچسب‌های کلاس یا هر داده دیگری باشد. این شرط را می‌توان به‌عنوان لایه ورودی اضافی به هر دو شبکه مولد و شبکه متمایزکننده اعمال کرد. هر دو شبکه برای تنظیم و یادگیری پارامترهای خود، از این ورودی‌های اضافی استفاده می‌کنند. در مسئله تولید تصویر از روی متن درواقع شرط شبکه مولد، متن ورودی است که تصویر از روی آن تولید می‌شود.

۲-۲- مدل شباهت چندوجهی مبتنی بر توجه عمیق[3]

مدل DAMSM [14] با یک رمزگذار متن و یک رمزگذار تصویر، نواحی مختلف تصویر و لغات موجود در جمله را به یک فضای مشترک نگاشت می‌کند و قادر است شباهت متن-تصویر را محاسبه و میزان کیفیت تصویر را ارزیابی کند. رمزگذار متن، یک واحد حافظه طولانی کوتاه‌مدت دوجهته[4][20] است که برای ذخیره‌سازی و دسترسی بهتر به اطلاعات، مورد استفاده قرار می‌گیرد و می‌تواند بردارهای معنادار را از توصیف ورودی استخراج کند. Bi-LSTM نوع خاصی از شبکه‌های عصبی بازگشتی[5] است که توانایی یادگیری وابستگی‌های بلندمدت را دارد. در این مدل هر لغت، به دو واحد پنهان که در دو جهت هستند، مرتبط است و با پیوست این دو بردار، نمایشی معنادار از واژه به‌دست می‌آید. هر واژه موجود در توصیف، بردار $e \in R^{D\times T}$ است که T تعداد واژه‌های توصیف و D طول بردار ویژگی هر لغت است. بردار ویژگی جمله نیز از الحاق آخرین خروجی لایه پنهان به‌دست می‌آید.

رمزگذار تصویر[6]، یک لایه شبکه عصبی کانولوشن به نام Inception-v3 [12] است که ماتریس حاصل از آن، حاوی ویژگی‌های مربوط به نواحی تصویر است. هر ناحیه از تصویر، بردار $f \in R^{N\times R}$ است که R تعداد نواحی تصویر و N طول بردار ویژگی هر ناحیه است. بردار ویژگی مربوط به‌کل تصویر نیز، از آخرین لایه pooling در مدل به‌دست می‌آید. درنهایت، برای محاسبه میزان تشابه نواحی تصویر و واژگان جمله، بردار ویژگی‌های متن و تصویر با استفاده از یک لایه پرسپترون به فضایی با ابعاد یکسان نگاشت می‌شوند.

۲-۳ شبکه‌های مبتنی بر حافظه پویا[7]

شبکه مبتنی بر حافظه، ابتدا اطلاعات را در یک حافظه خارجی ذخیره و سپس در گام‌های بعدی، از این اطلاعات استفاده می‌کند. در سال‌های اخیر نوعی از این شبکه‌ها با بهره‌گیری از حافظه کلید-مقدار[8] [9] در ساختار شبکه‌های رقابتی مولد استفاده شده‌است. در این مدل، هر مقدار از ماژول حافظه، دارای یک وزن است که کلید حافظه نام دارد و در زمان محاسبه نتیجه خروجی، مورد استفاده قرار می‌گیرد. شبکه مبتنی بر حافظه پویا [5]، شبکه‌ای است که در پژوهش‌های اخیر در مسئله تولید تصویر از روی متن به کار گرفته شده‌است.

در این مدل، در هر گام به‌صورت پویا، لغاتی که بیشترین ارتباط با تصویر تولیدشده دارند در حافظه نوشته می‌شوند؛ لذا باعث تولید تصاویری می‌شود که ارتباط بیشتری با توصیف ورودی دارند. در بخش خواندن از حافظه، ویژگی‌های تصویر تولید شده اولیه، به فرم یک پرس‌وجو برای بازیابی اطلاعات از حافظه به کار می‌رود. این اطلاعات در گام‌های بعدی برای بهبود تصویر اولیه به کار می‌روند.

[1] Generator
[2] Discriminator
[3] Deep Attentional Multimodal Similarity Model (DAMSM)
[4] Bidirectional Long Short Term Memory (Bi-LSTM)
[5] Recurrent Neural Networks (RNNs)
[6] Image encoder
[7] Dynamic Memory Networks
[8] key-value

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

۳- سابقه پژوهش

روش‌های اولیه [22 ، 15] برای تولید تصویر از روی متن درواقع از ترکیب فرآیند جستجو و یادگیری با نظارت استفاده کرده‌اند. روش کار بدین صورت است که همبستگی واژه‌های موجود در متن و نواحی تصویر محاسبه و واژه‌هایی که قابل به تصویر در آمدن هستند، انتخاب می‌شوند. این واحدها در ادامه برای بازیابی تصاویری که به آنها مربوط هستند، مورد استفاده قرار می‌گیرند. مشکل این راهکار این است که توانایی تولید تصاویری با محتوای جدید را ندارد؛ به همین دلیل در سال‌های اخیر پژوهش‌هایی برای تولید تصویر از روی متن مبتنی بر شبکه‌های رقابتی مولد و شبکه‌های عمیق رمزگذار کانولوشن ارائه‌شده‌اند [21 ، 18 ،17 ، 14] . این روش‌ها به‌عنوان روش‌های چندوجهی شناخته می‌شوند که ویژگی‌هایی از مدل‌ها، الگوریتم‌ها و ایده‌های مختلف برای بهبود راه‌حل مسائل با هم ترکیب می‌شوند [7 ،3]. *رید* و همکارانش [10] در سال ۲۰۱۶ برای نخستین‌بار با استفاده از شبکه‌های رقابتی مولد، معماری جدیدی برای مدل‌سازی تصاویر ارائه دادند که می‌تواند نویسه‌ها را به پیکسل‌ها انتقال دهد. سامانه پیشنهادی آن‌ها قادر است یک جمله را در قالب یک تصویر نمایش دهد. آنها توانستند تصاویر قابل‌قبولی از گل‌ها و پرندگان با وضوح ۶۴×۶۴ تولید کنند. با این حال، روش پیشنهادی آن‌ها معمولاً فاقد جزئیات دقیقی از اشیا تصویر مانند چشم پرندگان بود و توانایی تولید تصاویری با وضوح بالاتر برای مثال ۱۲۸×۱۲۸ را نداشت.

ژنگ و همکارانش [17] با الهام از تصویرسازی انسانی، مدلی را پیشنهاد دادند که برای مدیریت بهتر، مسئله را به دو زیر مسئله کوچک‌تر تقسیم‌بندی کرده و با به‌کارگیری شبکه دسته‌ای GAN به حل آن‌ها می‌پردازد. در گام نخست، طرح و رنگ اولیه اشیای موجود در متن مشخص می‌شوند که خروجی این مرحله، تصویری با وضوح پایین خواهد بود. در گام دوم خروجی مرحله قبل و متن به‌عنوان ورودی به سامانه داده می‌شوند تا تصویر اولیه بهبود یابد و تصویری با کیفیت بالا تولید شود. آن‌ها همچنین برای ارتقای وضوح تصاویر تولیدشده، در شبکه GAN به‌جای استفاده از یک شبکه مولد و یک شبکه تفکیک‌کننده، از چند شبکه که در ساختاری درختی قرار گرفته‌اند، استفاده کردند.

در این روش، تصاویر یک صحنه با مقیاس‌های متفاوت از شاخه‌های مختلف درخت تولید می‌شوند. در این حوزه پژوهش‌هایی که از شبکه رقابتی مولد به‌صورت سلسله‌مراتبی استفاده کرده‌اند قادر به تولید تصاویری با کیفیت بیشتر هستند [21 ، 14]. در این روند ابتدا تصویری با کیفیت پایین تولید می‌شود و در گام‌های بعدی کیفیت این تصویر بهبود می‌یابد. StackGAN از شبکه GAN، در دو گام استفاده می‌کند و پس از تولید تصویر اولیه، در گام دوم دوباره با به‌کارگیری شبکه GAN، تصویری با وضوح ۲۵۶×۲۵۶ تولید می‌کند. نسخه دوم با نام StackGAN++ارائه شد که برای بهبود تصویر، به‌جای استفاده از شبکه GAN، از ساختاری درختی استفاده کرده و آن را در چند مرحله به کار می‌گیرد تا تصویری یکنواخت‌تر تولید کند.

شبکه AttGAN [14] که بعد از آن ارائه شد مشابه دو شبکه یادشده است؛ به‌علاوه اینکه از سازوکار توجه نیز استفاده می‌کند. شیوه این شبکه به این صورت است که در گام نخست، بر اساس تعبیه متن[1]مربوط به‌کل جمله، تصویری با وضوح پایین تولید می‌شود و سپس در گام‌های بعدی با سازوکار توجه، لغاتی را که حائز اهمیت بیشتری برای بهبود تصویر هستند، مورد استفاده قرار می‌گیرند. با این روند، ابتدا تصویری کلی خواهیم داشت و به‌مرور جزئیات به تصویر اضافه می‌شوند.

در شبکه‌هایی که تاکنون بیان شدند، کیفیت تصویر نهایی وابسته به کیفیت تصویری است که در گام نخست تولید می‌شود. در صورتی که تصویر تولیدشده کیفیت قابل قبولی نداشته‌باشد، گام‌های بعدی نیز قادر نیستند آن را به‌خوبی بهبود دهند. برای حل این مشکل، در سال ۲۰۱۹، *ژو* و همکارانش شبکه DM-GAN [21] را ارائه دادند. این شبکه با تکیه‌بر حافظه پویای کلید-مقدار، بر روی بهبود کیفیت تصویر تولیدشده در گام نخست تمرکز دارد. ویژگی‌های تصویر اولیه به‌عنوان کلید جستجو در ماژول حافظه استفاده می‌شود. در واحدهای حافظه در هر گام، کلمات مرتبط با تصویر تولیدشده به‌طور پویا انتخاب و نوشته می‌شوند.

با توجه به پیشرفت‌های اخیر در استفاده از ساختار سلسله‌مراتبی شبکه‌های GAN و حافظه پویا در حل مسئله تولید متن از روی تصویر، ما روشی بر این مبنا ارائه کرده‌ایم. روش پیشنهادی ما برخلاف سایر روش‌ها، به‌جای یک جمله، از سه جمله برای تولید و بهبود تصویر بهره می‌برد.

۴- روش پیشنهادی

روش پیشنهادی، شبکه رقابتی مولد سلسله‌مراتبی مبتنی بر چند جمله برای تولید تصویر از روی متن است. ما در این پژوهش، دو گزینه کلیدی را مورد بررسی قرار خواهیم

[1] Text embedding

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

داد: ۱) تولید تصویری با کیفیت بالاتر در گام نخست، و ۲) استفاده از دو توصیف اضافه برای بهبود تصویر اولیه در گام‌های بعدی. برای نیل به مورد نخست، از پیوست سه جمله برای تولید تصویر اولیه استفاده می‌کنیم.

در مجموعه‌دادگان موجود در این حوزه، برای هر تصویر دست‌کم پنج توصیف وجود دارد. برای رسیدن به دومین هدف، برخلاف کارهای پیشین که تنها از یک جمله برای تولید و بهبود تصویر استفاده کرده‌اند، در این پژوهش سه جمله را به کار برده‌ایم. به‌دلیل حجم بالای محاسبات و نیاز به سخت‌افزار با قدرت بیشتر، امکان انجام آزمایش با پنج جمله وجود نداشت. در ادامه، ابتدا معماری مدل پیشنهادی را شرح می‌دهم و پس از آن تابع هزینه ارائه‌شده برای مدل پیشنهادی معرفی می‌شود.

جزئیات معماری مدل

مدل پیشنهادی، یک مدل سلسله‌مراتبی است که تصویر را از وضوح کم به زیاد تولید می‌کند؛ به‌صورتی که ابتدا تصویری شامل کلیات مربوط به توصیف، تولید و طی مراحل بعدی جزئیات به آن اضافه می‌شود. معماری سلسله‌مراتبی در سال‌های اخیر پیشرفت به سزایی در این زمینه داشته است [2، 8، 16]. در ابتدا، رمزگذار متن که یک مدل Bi-LSTM است، تعبیه متن جمله اولیه را تولید می‌کند. بردار حاصل، بردار ویژگی مربوط به جمله ورودی محسوب‌می‌شود.

از آنجایی که تعداد متون تعبیه‌شده در مجموعه‌داده‌های آموزشی کم هستند، برای افزایش عمومیت مدل خروجی مرحله قبل به ماژول تقویت شرطی[1] [17] ارسال می‌شود تا متغیر شرطی اضافه‌ای تولید شود. عملکرد این ماژول به این صورت است که متغیر شرطی از فضای توزیع گوسی $N(\mu(\varphi_t), \Sigma(\varphi_t))$ گرفته‌می‌شود که φ_t، تعبیه متن مربوط به جمله ورودی، $\mu(\varphi_t)$ و $\Sigma(\varphi_t)$ به‌ترتیب میانگین و کوواریانس بردار تعبیه متن هستند. با این روش، مشکل بیش‌برازش کاهش می‌یابد و مدلی قدرتمندتر[2]خواهیم داشت؛ سپس این متغیر شرطی به بردار نویز الحاق و به ورودی شبکه مولد داده‌می‌شود تا تصویری با وضوح ۶۴×۶۴ تولید کند. تصویر اولیه تولید شده به‌همراه داده‌های آموزشی و بردار تعبیه متن، به مدل متمایزکننده داده می‌شوند. متمایزکننده، کیفیت تصویر تولیدشده و اینکه چقدر به توصیف ورودی مربوط است، بررسی می‌کند.

در گام بعد، جمله دوم به‌عنوان ورودی به کار می‌رود. در این مرحله به‌جای تعبیه متن کل جمله، بردار ویژگی لغات آن را به کار می‌بریم؛ سپس بردار ویژگی تصویر تولیدشده در گام قبل و بردار ویژگی جمله قبلی، به حافظه پویا داده‌می‌شود تا با استفاده از سازوکار توجه، لغاتی که اهمیت بیشتری برای بهبود تصویر اولیه دارند، بازیابی و برای تولید تصویر با وضوح ۱۲۸×۱۲۸ به شبکه مولد دوم داده شوند. این روند در گام بعد، با استفاده از سومین جمله برای تولید تصویری با وضوح ۲۵۶×۲۵۶ تکرار می‌شود.

تابع هزینه

تابع هزینه برای شبکه مولد و شبکه متمایزکننده به‌صورت جدا تعریف می‌شود که در ادامه شرح آنها آمده است.

۴-۱ -۲ تابع هزینه شبکه مولد

تابع هزینه شبکه مولد از سه بخش تشکیل شده که در رابطه (۲) آمده است:

$$L=L_G+\lambda_1 L_{CA}+\lambda_2 L_{DAMSM}$$

در این رابطه، L_G مجموع توابع هزینه سه شبکه مولد موجود در شبکه پیشنهادی است که هر یک از آنها از رابطه (۳) به‌دست می‌آید:

$$L_{G_i}=-\frac{1}{2}E_{\tilde{I}_i\sim p_{G_i}}D_i^{uc}(\tilde{I}_i)-\frac{1}{2}E_{\tilde{I}_i\sim p_{G_i}}D_i^{c}(\tilde{I}_i,\varphi_t)$$

عبارت نخست در این رابطه، تابع هزینه شرطی و عبارت دوم تابع هزینه غیرشرطی را نمایش می‌دهد. تابع هزینه غیر شرطی، تلاش می‌کند، تصویر تولیدشده تا حد امکان واقعی به نظر برسد و تابع شرطی نیز تلاش می‌کند میزان تطابق تصویر تولیدشده با جمله ورودی را افزایش دهد.

عبارت دوم در رابطه (۲)، تابع هزینه مربوط به ماژول شرطی است که در بخش ۴-۱ اشاره شد و در رابطه (۴) آمده است. این رابطه، معیار واگرایی کولبک-لیبلر[3] است و میزان تشابه رفتار توزیع ماژول شرطی نسبت به توزیع نرمال را نشان‌می‌دهد.

$$L_{CA}=D_{KL}(N(\mu(\varphi_t),\Sigma(\varphi_t))\parallel N(0,I)$$

عبارت سوم، تابع هزینه DAMSM [14] است که معیاری از ارتباط تصویر تولید شده و توصیف ورودی را نشان‌می‌دهد. این عبارت ارتباط توصیف ورودی و تصویر

[1] Conditional Augmentation (CA)
[2] Robust
[3] Kullback-Leibler divergence

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

تولیدشده را در سطح جمله و در سطح واژه ارزیابی می‌کند.

۴-۲ -۲ تابع هزینه شبکه متمایزکننده

تابع هزینه برای هر شبکه متمایزکننده به‌صورت مستقل و با توجه به رابطه (۵) به‌دست می‌آید:

$$\begin{aligned} L_{D_i} = & \frac{1}{2}[E_{I_i \sim p_{data_i}}\{\max(0,1-D_i^{uc}(I_i)\} \\ & + E_{I_i \sim p_{data_i}}\{\max(0,1-D_i^{c}(I_i,\varphi_t)\}] + \\ & \frac{1}{3}[E_{\tilde{I}_i \sim p_{G_i}}\{\max(0,1+D_i^{uc}(\tilde{I}_i)\} \\ & + E_{\tilde{I}_i \sim p_{G_i}}\{\max(0,1+D_i^{c}(\tilde{I}_i,\varphi_t)\} \\ & + E_{\tilde{I}_i \sim p_{data_i}}\{\max(0,1+D_i^{c}(I_i,\overline{\varphi_t}))\}] \end{aligned} \quad (۵)$$

در رابطه بالا، φ_t توصیف ورودی، I_i تصویر اصلی مربوط به توصیف در داده آموزشی، $\tilde{I}_i$ تصویر تولیدشده در گام iام است. گفتنی است که عبارت آخر بیان‌گر ارتباط یک جمله اشتباه (جمله‌ای غیر از توصیف ورودی از داده آموزشی) با تصویر ورودی است که ارتباط منفی خوانده می‌شود و قدرت یادگیری را بهبود می‌دهد.

۵- پیاده‌سازی و اجرا

پیاده‌سازی روش پیشنهادی به زبان پایتون و بر بستر سخت‌افزاری کلاستری با ۲۵ واحد پردازشی در دانشگاه کپنهاگ انجام شد. چهار واحد پردازش گرافیکی[1] مدل GTX-1080 با ۳۲ گیگابایت حافظه دسترسی تصادفی[2] برای این آزمایش مورد استفاده قرار گرفت.

۵-۱- مجموعه‌دادگان

برای ارزیابی مدل پیشنهادی، آزمایش‌ها بر روی دو مجموعه‌داده CUB200 [13] و ids-ad [6] انجام شدند. چون یکی از اهداف مقاله، تولید تصاویر با پیچیدگی بیشتر و دارای بیش از یک شیء و با جزئیات است، دادگان ids-ade مورد استفاده قرار گرفتند که دارای ۳۵۲۸ داده آموزشی و ۴۴۱ داده آزمایشی است. گفتنی است که این مجموعه، برای هر تصویر، پنج توصیف وابسته دارد که جمله نخست توصیف کلی از تصویر است که به‌طورمعمول به دسته‌ای که تصویر به آن تعلق دارد، اشاره می‌کند. نمونه‌ای از تصویر به‌همراه توصیف‌های آن از این مجموعه‌دادگان در شکل(۱) نشان داده‌شده‌است.

همان‌گونه که پیشتر اشاره شد، پژوهش حاضر نخستین مطالعه‌ای است که مجموعه‌دادگان ids-ade را در این حوزه به کار گرفته است، از این رو مقایسه روش پیشنهادی با روش‌های پایه روی مجموعه‌دادگان پرندگان با ۸۸۵۵ داده آموزشی و ۲۹۳۳ داده آزمایشی انجام شده‌است.

1. It's a bedroom with jade green walls.
2. The bedroom suite is all light colored wood. There is a bed, dresser, night stand and an armoire.
3. The bed is a Full bed with a wooden headboard and footboard inset with scrolled metal.
4. An off-white oriental carpet is in the middle of the white floor.

The armoire is elevated on a platform.

5. It's a bedroom with jade green walls.
6. The bedroom suite is all light colored wood. There is a bed, dresser, night stand and an armoire.
7. The bed is a Full bed with a wooden headboard and footboard inset with scrolled metal.
8. An off-white oriental carpet is in the middle of the white floor.
9. The armoire is elevated on a platform.

شکل- ۱: نمونه‌ای از مجموعه‌داده ids-ade [6]

Figure-1: An example of ids-ade dataset.

جزئیات پیاده‌سازی

برای تعبیه متن، مدل آموزش‌دیده[3] Bi-LSTM با اندازه ۲۵۶ به کار گرفته شد. در مجموعه‌دادگان پرندگان برای رمزگذار تصویر، نیز مدل پیش‌آموزش‌دیده Inception-v3 را به کار گرفتیم. ازآنجایی‌که داده‌های مجموعه ids-ade، با داده‌های ImageNet که رمزگذار روی آن آموزش داده شده، متفاوت است، در این پژوهش از

[1] Graphics Processing Unit (GPU)
[2] Random-Access Memory (RAM)
[3] Pre-trained model

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

رمزگذار آموزش داده‌شده روی این مجموعه‌داده استفاده کردیم. در فرآیند یادگیری بهینه‌ساز ADAM با اندازه دسته[1] ۱۰ به کار رفته‌است. تعداد تکرار برای داده‌های CUB-200، ۶۰۰ و همچنین برای دادگان ids-ade، ۲۰۰۰ در نظر گرفته شدند.

معیارهای ارزیابی

برای ارزیابی کارایی مدل پیشنهادی، به‌وسیله آن سی‌هزار تصویر از روی توصیف‌های موجود در مجموعه‌دادگان آزمایشی تولید شد. در شبکه‌های رقابتی مولد، جهت بررسی کیفیت و تنوع تصاویر تولیدشده از معیار IS بهره می‌برند. این معیار به‌وسیله رابطه (۶) محاسبه می‌شود:

$$IS(G) = exp\left(E_{x\sim p_g} D_{KL}(p(y|x) \,||\, p(y))\right) \qquad (۶)$$

در رابطه (۶)، عبارت $p(y|x)$ به‌معنای توزیع شرطی y نسبت به x است که در آن y برچسب پیش‌بینی شده به‌وسیله مدل Inception-v3 است. با توجه به رابطه (۶)، هر چه توزیع تصاویر تولیدشده به توزیع داده‌های آموزشی نزدیک‌تر و همچنین تنوع تصاویر بیشتر باشد، IS مقدار بیشتری خواهد داشت و در نتیجه مدل بهتری داریم.

چون این معیار قادر به ارزیابی میزان ارتباط تصویر به متن ورودی نیست، معیار R-precision به کار می‌رود. روش کار به این صورت است که ابتدا برای تصویر تولیدشده، صد توصیف انتخاب می‌شود. R توصیف از این توصیف انتخابی، مربوط و وابسته به تصویر هستند.

تعداد R-۱۰۰ توصیف به‌صورت تصادفی از مجموعه توصیف‌های نامرتبط با تصویر انتخاب می‌شوند. سپس با استفاده از معیار شباهت کسینوسی[2]، میزان شباهت صد توصیف با تصویر تولیدشده محاسبه و R توصیف از شبیه‌ترین توصیف‌ها انتخاب می‌شوند. اگر تعداد r توصیف از مجموعهR، مرتبط با تصویر باشد، آنگاه مقدار R-precision برابر با r/R است. در آزمایش‌های انجام‌شده، مقدار R=1 در نظر گرفته شد.

۶- نتایج آزمایشی و تحلیل

برای ارزیابی کیفیت تصاویر تولیدشده از مدل پیشنهادی در این پژوهش، معیار IS محاسبه شد. در این معیار، هر چه مقدار IS بیشتر باشد، تصویر با کیفیت بالاتری داریم. نتایج به‌دست‌آمده برای مجموعه‌دادگان CUB-200، که در جدول- ۱ آمده است. با خط مبنای مرزهای دانش[3] مقایسه شد؛ چنانکه در جدول مشهود است، مقدار IS برای مدل پیشنهادی MSH-GAN (که به‌صورت پررنگ نشان داده‌شده‌است) عملکرد بهتری نسبت به روش‌های دیگر دارد؛ درواقع، مقدار IS در مدل پیشنهادی در مقایسه با روش StackGAN و روش AttGAN که روش‌های به‌حساب می‌آیند، به‌ترتیب از ۳/۷۰ (۲۹/۷۲٪ بهبود) و ۴/۳۶ (۱۰/۱٪ بهبود) به ۴/۸۰ افزایش یافته است که بهبود قابل‌توجهی است.

(جدول- ۱): مقدار IS بر روی مجموعه دادگان CUB-200

Table-1: The IS on the CUB-200 dataset.

Model	IS (↑)
GAN-INT-CLS [10]	2.88
StackGAN [17]	3.70
AttGAN [14]	4.36
DM-GAN [21]	4.69
MSH-GAN	4.80

برای ارزیابی اینکه تصاویر تولیدشده از شبکه مبتنی بر GAN چقدر به متن ورودی به مدل ارتباط دارد، از معیار R-precision استفاده کردیم. با توجه به اینکه مدت زمان زیادی از ارائه این معیار در پژوهش‌های این حوزه نمی‌گذرد، مدل پیشنهادی تنها با دو روش AttGAN و DM-GAN مقایسه و نتایج در جدول-۲ آورده شده‌است. در این معیار، هرچه مقدار به‌دست‌آمده بیشتر باشد، مدل پیشنهادی بهتر بوده و قادر است تصاویری تولید کند که بیشتر به جمله ورودی مرتبط است.

همان‌طور که در جدول-۲ نشان داده‌شده‌است، مقدار R-precision در روش پیشنهادی بر روی مجموعه‌دادگان پرندگان مقدار بیشتری را دارد و توانسته‌است، نتایج را ۱۶/۸۸٪ بهبود دهد.

(جدول- ۲): مقدار R-precision بر روی مجموعه‌دادگان CUB-200.

Table-2: The R-precision on the CUB-200 dataset.

Model	R-Precision (↑)
AttGAN	67.82
DM-GAN	72.31
MSH-GAN	79.27

[1] Batch size
[2] Cosine similarity
[3] State-of-the-art

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]
[DOI: 10.61186/jsdp.19.4.33]

Ground truth	Examples of Descriptions	Generated images	
	1. The bird has a white and gray speckled belly and breast with a short orange bill. 2. A bird with a grey head and white throat, the bill is short and pointed, with grey covering the rest of the body. 3. A large bird with a grey coloring.		
	1. This bird has a red breast and a white belly and has a red head. 2. A small predominantly red colored bird with a small rounded beak, and a speckled white and red belly 3. A small red bird with light brown sides and a small brown beak.		
	1. A medium sized bird that has tones of grey and a large sized bill 2. This small bird has a thick beak, with brown feathers on it. 3. A bird that has a thick yellow beak, and a thick gray neck.		

(شکل- ۲): نمونه‌ای از نتایج تصاویر تولید شده از مدل پیشنهادی MSH-GAN

Figure-2: Example results of generated images from proposed model MSH-GAN

نمونه‌ای از تصاویر تولیدشده از مدل پیشنهادی بر روی مجموعه‌دادگان پرندگان در شکل(۲) آورده شده‌است. در شکل(۲)، ستون نخست تصویر واقعی در داده آزمایشی است. ستون دوم، سه جمله از ده جمله مربوط به تصویر است که مدل به‌صورت تصادفی انتخاب می‌کند. ما برای هر داده موجود در مجموعه آزمایشی، پنج تصویر تولید کردیم که در ستون سوم شکل(۲) ، دو تصویر تولیدشده به‌وسلهٔ مدل پیشنهادی نشان داده‌شده‌است.

با توجه به آنچه که در بخش‌های قبل بیان شد، یکی از اهداف این مقاله تمرکز بر روی مجموعه‌دادگان ids-ade است که حاوی تصاویر پیچیده‌تری نسبت به مجموعه‌دادگان پرندگان است.

در مدل پیشنهادی، از پنج توصیفی که برای هر تصویر در داده‌های آموزشی وجود دارد، ما سه توصیف را انتخاب می‌کنیم. برای انتخاب این سه جمله، ما سه دیدگاه مختلف را اجرا کردیم که در ادامه به‌همراه نام انتخابی برای آن‌ها، شرح داده‌شده‌اند:

- CS123: استفاده از سه جمله نخست به همان ترتیبی که در داده‌های آموزشی آمده‌اند.
- CS1RR: استفاده از نخستین توصیف در داده آموزشی، به‌عنوان جمله ورودی به گام نخست در مدل پیشنهادی و انتخاب دو جمله تصادفی از چهار جمله باقی مانده برای گام‌های دوم و سوم.
- CSRRR: انتخاب تصادفی سه جمله از پنج توصیف موجود در داده‌های آموزشی.

مدل MSH-GAN با تکیه‌بر راه‌کارهای ارائه‌شده در بالا بر روی مجموعه‌دادگان ids-ade مورد ارزیابی قرار گرفته است و نتایج در جدول-۳ درج‌شده‌اند. بهترین نتیجه به‌صورت پررنگ نشان داده‌شده‌است.

(جدول- ۳): ارزیابی مدل MSH-GAN بر روی دادگان ids-ade

Table-3: Evaluation of MASH-GAN model on ids-ade dataset.

Method	IS (↑)	R-Precision (↑)
CS123	4.98	69.95
CS1RR	**5.19**	**80.59**
CSRRR	4.87	69.57

با توجه به نتایج به‌دست‌آمده در جدول (۳)، مشخص است که راه‌کار CS1RR بهترین نتیجه را دارد. نتیجه به‌دست‌آمده مطابق با آن چیزی است که ما از مدل پیشنهادی انتظار داشتیم. همان‌طورکه پیش‌تر ذکر شد، در مجموعه‌دادگان ids-ade نخستین جمله حاوی اطلاعات کلی مثل اشیای موجود در تصویر یا برچسب کلاس تصویر است و چهار جمله بعدی به توصیف دقیق‌تری از تصویر می‌پردازد؛ بنابراین با انتخاب نخستین جمله از داده آموزشی، مدل پیشنهادی یک نمای کلی را به تصویر می‌کشد و با استفاده از دو جمله بعدی که تصادفی انتخاب می‌شوند، جزئیات بیشتری اضافه می‌شود و تصویر بهبود می‌یابد.

برای اینکه بتوانیم نتایج مدل پیشنهادی بر روی دادگان ids-ade را با روشی از آخرین پیشرفت‌های علمی در این حوزه مقایسه کنیم، بهترین روش را از میان روش‌هایی که در جدول (۱) آمده است، یعنی DM-GAN را انتخاب کردیم. کد مربوط به این مدل را از بستر github

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]

دریافت کرده و آن را بر روی مجموعه‌دادگان ids-ade کردیم. از آن‌جایی که برای مدل DM-GAN تنها یک توصیف لازم داریم، این روش را با دو راه‌کار زیر آزمودیم:

- C-R: انتخاب تصادفی یک توصیف از پنج توصیف.
- C-1: انتخاب نخستین جمله از پنج جمله که حاوی اطلاعات کلی از تصویر است.

نتایج اجرای مدل DM-GAN بر اساس دو راه‌کار بالا بر روی مجموعه‌دادگان ids-ade در جدول(۴) آمده است. با مقایسه جدول‌های ۳ و ۴، می‌بینیم که اجرای مدل پیشنهادی، بر روی مجموعه‌دادگان ids-ade نتایج بهتری را تولید می‌کند.

(جدول- ۴): ارزیابی مدل MSH-GAN بر روی دادگان ids-ade

Table-4: Evaluation of MASH-GAN model on ids-ade dataset.

Model	IS (↑)	R-Precision (↑)
C-R	4.61	69.73
C-1	4.35	65.3

در شکل (۳) نمونه‌هایی از تصاویر تولیدشده به‌وسیله مدل پیشنهادی بر روی مجموعه‌دادگان ids-ade آمده است. در هر ردیف، سه توصیف از آن نمونه به‌همراه تصویر تولیدشده نشان داده‌شده‌است. گفتنی است که در هر سطر، جمله نخست، نخستین جمله در نمونه آزمایشی است و دو جمله دیگر به‌صورت تصادفی از بین چهار جمله دیگر انتخاب‌شده‌اند.

(شکل- ۳): نمونه‌هایی تصاویر تولید شده در مجموعه دادگان ids-ade

Figure-3: Examples of generated images on ids-ade dataset

همان‌طور که در شکل (۳) نیز مشخص است، در برخی موارد مدل قادر به تولید تصویری قابل قبول از نظر واقعی‌بودن و مرتبط‌بودن به متن، نیست.

از آنجا که در مجموعه‌دادگان ids-ade، جملات دوم تا پنجم در واقع اشیا موجود در تصویر را به‌طورمعمول به‌صورت جدا توصیف می‌کنند و دو جمله‌ای که برای بهبود تصویر به کار می‌روند به‌صورت تصادفی انتخاب می‌شوند؛ بنابراین تصویر تولیدشده از مدل تا حدی وابسته به انتخاب جملات است. همچنین بررسی نتایج به‌دست‌آمده نشان‌می‌دهد، در حالاتی که جمله نخست حاوی اطلاعات مناسبی نباشد، تصویر تولیدشده از مدل، کیفیت پایین‌تری خواهد داشت.

۷- نتیجه‌گیری

در این مقاله یک روش سلسله‌مراتبی مبتنی بر شبکه‌های رقابتی مولد بر پایه چند جمله برای مسئله تولید تصویر از روی متن ارائه دادیم. عملکرد روش پیشنهادی بر روی مجموعه‌دادگان مربوط به پرندگان و طراحی داخلی بررسی شد.. نتایج آزمایش‌ها نشان‌می‌دهد که روش پیشنهادی نسبت به روش‌های پیشین عملکرد بهتری دارد. با وجود بهبود حاصل، مدل پیشنهادی همچنان با چالش‌هایی روبه‌رو است. نتایج به‌دست‌آمده به‌خصوص برای مجموعه‌دادگانids-ade که جملات به‌هم وابسته هستند، حساس به نوع انتخاب جملات است. مجموعه‌دادگان ids-ade که در این پژوهش برای نخستین‌بار در این حوزه مورد استفاده قرار گرفته، مجموعه‌ای است که تصاویر دارای جزئیات بیشتری هستند. از طرف دیگر این مجموعه نسبت به مجموعه‌دادگانی که در قبل در این حوزه مورد استفاده قرار گرفته‌اند، تعداد نمونه‌های کمتری برای آموزش دارد که برای یادگیری شبکه‌های GAN یک چالش محسوب‌می‌شود. ازاین‌رو مدل پیشنهادی در بعضی موارد قادر به تولید تصویر مناسب برای جملات ورودی نیست. در ادامه این پژوهش می‌توان با روش‌های مبتنی بر اشتراک پارامترهای شبکه در گام‌های مختلف، حجم محاسباتی را کاهش داد تا بتوان مدل را با هر پنج جمله آموزش داد.

اقرار و تقدیر

این پژوهش ضمن بهره‌وری از بورس فرصت مطالعاتی، اعطایی از طرف «وزارت علوم، تحقیقات و فناوری ایران»، در «دانشکده علوم کامپیوتر دانشگاه کپنهاگ» اجرا شد.

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]

۸- مراجع 8-Refrence

۱. حاجی اسمعیلی، محمد مهدی و غلامعلی، منتظر، "رنگ‌آمیزی خودکار تصاویر خاکستری به‌کمک شبکه‌های زایای رقابتی"، مجله پردازش علائم و داده‌ها، دوره ۱۶، شماره ۱، صفحات ۷۴-۵۷، ۱۳۹۸.

[۱] M. M. Haji-Esmaeili, and G. Montazer, "Automatic Coloring of Grayscale Images Using Generative Adversarial Networks, ", Journal of Signal and Data Processing (JSDP), vol. 16 (1), pp. 57-74, 2019.

[۲] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy, " in IEEE Transactions on Pattern Analysis, 2017 .

[۳] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "Tac-gan-text conditioned auxiliary classifier generative adversarial network, " arXiv preprint arXiv:1703.06412, 2017.

[۴] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets, " in Advances in neural information processing systems, 2014.

[۵] C. Gulcehre, S. Chandar, K. Cho, and Y. Bengio, "Dynamic neural turing machine with continuous and discrete addressing schemes, " Neural computation, vol. 30, no. 4, pp. 857-884, 2018 .

[۶] N. Ilinykh, S. Zarrieß, and D. Schlangen, "Tell Me More: A Dataset of Visual Scene Description Sequences, " in Proceedings of the 12th International Conference on Natural Language Generation, 2019.

[۷] K. J. Joseph, A. Pal, S. Rajanala, and V. N. Balasubramanian, "C4synth: Cross-caption cycle-consistent text-to-image synthesis, " in IEEE Winter Conference on Applications of Computer Vision (WACV), 2019 .

[۸] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, and J. Gao, "Object-driven text-to-image synthesis via adversarial training, " in Proc. of the IEEE Conf.e on Computer Vision and Pattern Recognition, 2019 .

[۹] A. Miller, A. Fisch, J. Dodge, A. H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents, " in Proceeding of Empirical Methods in Natural Language Processing (EMNLP), 2016 .

[۱۰] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis, " arXiv preprint arXiv:1605.05396, 2016.

[۱۱] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans, " in Advances in neural information processing systems (NIPS), 2016 .

[۱۲] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision, " in Proc. of the IEEE conf. on computer vision and pattern recognition, 2016 .

[۱۳] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The caltech-ucsd birds-200-2011 dataset, 2011 .

[۱۴] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks, " in Proc. of the IEEE conf. on computer vision and pattern recognition, 2018 .

[۱۵] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes, " in European Conf. on Computer Vision, 2016 .

[۱۶] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation, " in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019 .

[۱۷] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, " in Proc.of the IEEE int. conference on computer vision, 2017.

[۱۸] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks, " in IEEE transactions on pattern analysis and machine intelligence, 2017.

[۱۹] Z. Zhang, Y. Xie, and L. Yang, " Photo-graphic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network" in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018.

[۲۰] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification, " in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2016 .

[۲۱] M. Zhu, P. Pan, W. Chen, and Y. Yang, "dm-gan: Dynamic memory generative adversarial net. for text-to-image synthesis, " in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2019 .

[۲۲] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, "A text-to-picture synthesis system for augmenting communication, " in proceeding of Association for the Advanced of Artificial Intelligence (AAAI) , 2007.

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]

الهام پژهان: دانشجوی مقطع دکتری در رشته مهندسی کامپیوتر، گرایش هوش‌مصنوعی و رباتیک، در دانشگاه خود را در سال ۱۳۹۲ از دانشگاه شیراز در رشته مهندسی کامپیوتر دریافت کرد. ایشان از مهر ۱۳۹۸ تاکنون، در دانشکده علوم کامپیوتر دانشگاه کپنهاگ دانمارک به‌عنوان پژوهش‌گر مهمان حضور دارند. زمینه پژوهشی مورد علاقه ایشان پردازش زبان طبیعی، پردازش تصویر و یادگیری عمیق است.

نشانی رایانامه ایشان عبارت است از:
e.pejhan@stu.yazd.ac.ir

محمد قاسم‌زاده: دانشیار دانشکده مهندسی کامپیوتر در دانشگاه یزد، که در سال ۱۳۶۸ کارشناسی خود را در رشته علوم و مهندسی کامپیوتر از دانشگاه شیراز، و کارشناسی‌ارشد مهندسی کامپیوتر، گرایش هوش ماشین و رباتیک را در سال ۱۳۷۴ از دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) دریافت کرد. برای مقطع دکتری از بهمن ۱۳۸۰ تا بهمن ۱۳۸۴ در دانشگاه‌های تِریِر و پتسدام آلمان، به پژوهش مشغول بود. وی از تز دکترای خود در «علوم نظری کامپیوتر» در دی‌ماه ۱۳۸۴ دفاع کرد. همچنین در سال ۱۳۹۵ فرصت مطالعاتی خود را به‌عنوان پژوهش‌گر مهمان و پژوهش‌گر پسادکترا در HPI آلمان گذراند. حوزه‌های پژوهشی ایشان شامل طراحی و تحلیل الگوریتم، روش‌های هوش‌مصنوعی، پردازش زبان طبیعی، داده‌های حجیم و امنیت نرم‌افزار است.

نشانی رایانامه ایشان عبارت است از:
m.ghasemzadeh@ yazd.ac.ir

[Downloaded from jsdp.rcisp.ac.ir on 2025-08-21]

[DOI: 10.61186/jsdp.19.4.33]