بسمه تعالی

خلاصه پیشنهاد سمینار کارشناسی ارشد

عنوان: پرسش و پاسخ تصویری

۱- شرح مساله (با ارجاع به مراجع)

در سال‌های اخیر پیشرفت‌های زیادی در مسائل هوش مصنوعی و یادگیری عمیق که در تقاطع دو حوزه پردازش زبان طبیعی[1] و بینایی ماشین[2] قرار می‌گیرند؛ رخ داده است. یکی از مسائلی که اخیرا مورد توجه قرارگرفته است؛ پرسش و پاسخ تصویری[3] است. با توجه به یک تصویر و یک سؤال به زبان طبیعی، سیستم سعی می‌کند با استفاده از عناصر بصری تصویر و استنتاج جمع آوری شده از سوال متنی، پاسخ صحیح را پیدا کند.[۱]

پرسش و پاسخ تصویری نسخه گسترش یافته مسئله پرسش و پاسخ متنی[4] است است که اطلاعات بصری به مسئله اضافه شده است. شکل ۱ گویای تفاوت این دو مسئله است.

شکل ۱- مثالی از سیستم پرسش و پاسخ متنی و تصویری

در سیستم پرسش و پاسخ متنی، یک متن و یک سوال متنی به عنوان ورودی به سیستم داده می‌شود و انتظار می‌رود که سیستم با توجه به درک و تفسیری که از متن و سوال بدست می‌آورد؛ یک جواب متنی را خروجی دهد. اما در سیستم پرسش و پاسخ تصویری، یک تصویر و یک سوال متنی به ورودی سیستم داده می‌شود و انتظار می‌رود که سیستم بتواند با استفاده از عناصر بصری تصویر و تفسیری که از سوال بدست می‌آورد؛ یک پاسخ متنی را در خروجی نشان دهد.

مسئله پرسش و پاسخ تصویری پیچیدگی بیشتری نسبت به مسئله پرسش و پاسخ متنی دارد زیرا تصاویر بعد بالاتر و نویز بیشتری نسبت به متن دارند. علاوه بر این، تصاویر فاقد ساختار و قواعد دستوری زبان هستند. در نهایت هم، تصاویر غنای بیشتری از دنیای واقعی را ضبط می‌کنند، در حالی که زبان طبیعی در حال حاضر نشانگر سطح بالاتری از انتزاع دنیای واقعی است. [۲]

---

[1] Natural Language Processing
[2] Computer Vision
[3] Visual Question Answering
[4] Textual Question Answering

۲- مباحث تحت پوشش سمینار(با ارجاع به مراجع)

با توجه به موارد ذکر شده در قسمت شرح مساله، مباحث تحت پوشش این سمینار به ترتیب شامل موارد زیر خواهند بود:

۱) بررسی کلی و تعریف مبحث پرسش و پاسخ تصویری[۲]

۲) کاربرد و اهمیت این مسئله

۳) بررسی چالش‌های موجود در این مسئله

۴) بررسی مجموعه دادگان مطرح و مسابقات مطرح این حوزه

۴.۱) مجموعه داده DAQUAR [۳]

۴.۲) مجموعه داده VQA [۴]

۴.۳) مجموعه داده Visual Madlibs [۵]

۴.۴) مجموعه داده Visual 7w [۶]

۴.۵) مجموعه داده CLEVR [۷]

۴.۶) مجموعه داده Tally-QA [۸]

۴.۷) مجموعه داده KVQA [۹]

۵) بررسی فازهای مختلف مسئله پرسش و پاسخ تصویری[۱۰]:

۵.۱) فاز۱: استخراج ویژگی از تصویر و سوال

۵.۲) فاز۲: درک مشترک تصویر و سوال

۵.۳) فاز۳: تولید جواب

۶) معیارهای ارزیابی مسئله پرسش و پاسخ تصویری[۱۰]

۷) چگونگی ساخت مجموعه داده حاوی پرسش و پاسخ به زبان فارسی

۳- اهمیت موضوع

در طی سال‌های متمادی، محققان به دنبال ساخت ماشین‌هایی بودند که به اندازه‌ی کافی باهوش باشند که از آن به طور موثر همانند انسان‌ها برای تعامل استفاده کنند. مسئله‌ی پرسش و پاسخ تصویری یکی از پله‌های رسیدن به این رویای هوش مصنوعی است و از این جهت حائز اهمیت است.

کاربردهای بسیاری برای پرسش و پاسخ تصویری وجود دارد. یکی از مهم‌ترین موارد دستیار هوشمند برای افراد کم‌بینا و نابینا[5] است[۱۱]. علاوه بر این، در سال های اخیر دستیاران صوتی[6] و عامل‌های گفتگو[7] مانند Siri، Cortana و Alexa در بازار عرضه شدند که می‌توانند با انسان‌ها با استفاده از زبان طبیعی ارتباط برقرار کنند. در حال حاضر این دستیاران با استفاده از صوت و متن این ارتباط را برقرار می‌کنند در نتیجه گفتگوی بین این دستیاران با انسان‌ها مشابه دنیای واقعی نمی‌باشد. این ارتباط را می‌توان با استفاده از داده‌های تصویری و ویدئویی به واقعیت نزدیک‌تر کرد. اینجاست که مسئله‌ی پرسش و پاسخ تصویری برای نزدیک کردن تعامل بین انسان و عامل‌های گفتگو به دنیای واقعی می‌تواند موثر باشد. همین موضوع را می‌توانیم به صورت گسترده‌تری در ربات‌ها مشاهده کنیم. برای این‌که ربات بتواند بهتر با انسان‌ها ارتباط برقرار کند و به سوالات و درخواست‌ها پاسخ دهد؛ نیاز دارد که درک و فهم درستی از اطراف داشته باشد که این مستلزم داشتن تصویری دقیق از پیرامون است. بنابراین این ربات می‌تواند برای پاسخ به پرسش‌ها از دانشی که از طریق تصویر پیرامون خود بدست می‌آورد، جواب درستی را بدهد.

۴- نتیجه ارزیابی در گروه:

قبول ☐ رد ☐ تصحیح ☐ ارسال برای داوری ☐ تاریخ --/--/---- امضاء مدیر گروه:

[5] https://vizwiz.org/
[6] Voice Assistant
[7] Conversational Agents

مراجع

[١] Y. Srivastava, V.Murali, S. Ram Dubey, S. Mukherjee, "Visual Question Answering using Deep Learning: A Survey and Performance Analysis," *arXiv*, vol. abs/1909.01860, 2019.

[٢] Q. Wu, D. Teney, P. Wang. C. Shen, A.R. Dick, A. van den Hengel, "Visual question answering: A survey of methods and datasets," *ArXiv*, vol. abs/1607.05910, 2017.

[٣] M. Malinowski and M. Fritz, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input," *NIPS*, vol. abs/1410.0210, 2014.

[٤] S. Antol, A. Agrawal, J. Lu, M. Mitchell,D. Batra, C. Lawrence Zitnick, D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[٥] L. Yu, E. Park, A. Berg, T. Berg, "Visual madlibs: Fill in the blank image generation and question answering," *IEEE ICCV*, p. 2461–2469, 2015.

[٦] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[٧] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[٨] M. Acharya, K. Kafle, Ch. Kanan, "TallyQA: Answering complex counting questions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[٩] S. Shah, A. Mishra, N. Yadati, P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[١٠] S. Manmadhan, B. C. Kovoor, "Visual question answering: a state-of-the-art review," Springer, 2020, pp. 1-41.

[١١] D.Gurari, Q. Li, A. J Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.