پاسخ‌گویی به پرسش‌های مطرح شده از تصاویر
به کمک روش‌های یادگیری عمیق

مجید رفیعی، دکتر میرحسین دزفولیان

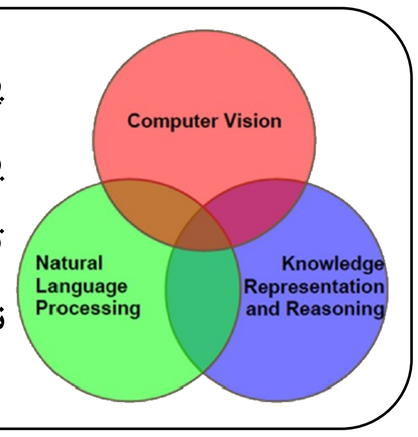
گروه آموزشی کامپیوتر، دانشکده فنی مهندسی، دانشگاه بوعلی سینا همدان

ستاد ملی هفته پژوهش و فناوری

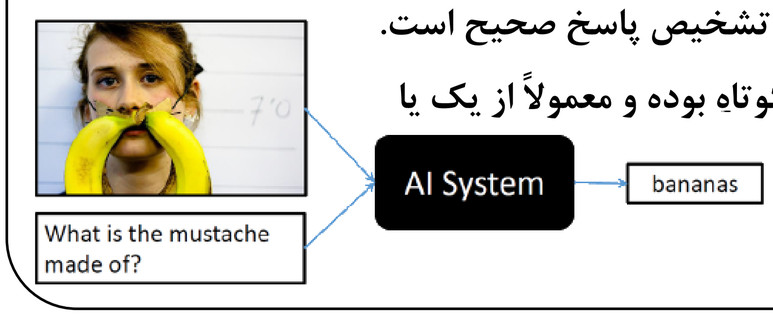
دانشگاه بوعلی سینا

VQA Visual Question Answering

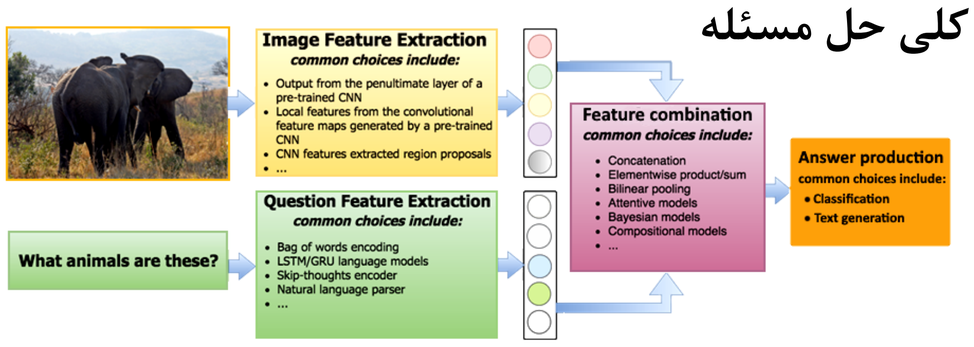
پاسخ‌گویی به پرسش‌های مطرح شده از تصاویر (VQA) یک مسئله‌ی جدید و میان رشته‌ای در حوزه‌های تحقیقاتی بینایی ماشین، پردازش زبان‌های طبیعی و نمایش دانش و استدلال منطقی است.



در عمومی‌ترین شکل مسئله، یک تصویر و یک پرسش متنی به رایانه داده می‌شود. وظیفه رایانه تشخیص پاسخ صحیح است. پاسخ‌ها اغلب بسیار کوتاهِ بوده و معمولاً از یک یا چند کلمه تشکیل شده‌اند. [۱]



روش کلی حل مسئله



روش‌های موجود به چهار بخش تقسیم می‌شوند: [۲]

۱- تعبیه سازی توام

۲- به کارگیری مکانیزم توجه

۳- روش‌های ترکیبی

۴- بهره از پایگاه‌های دانش خارجی

اصول روش پیشنهادی:

۱- در نظر گرفتن مفهوم پرسش h

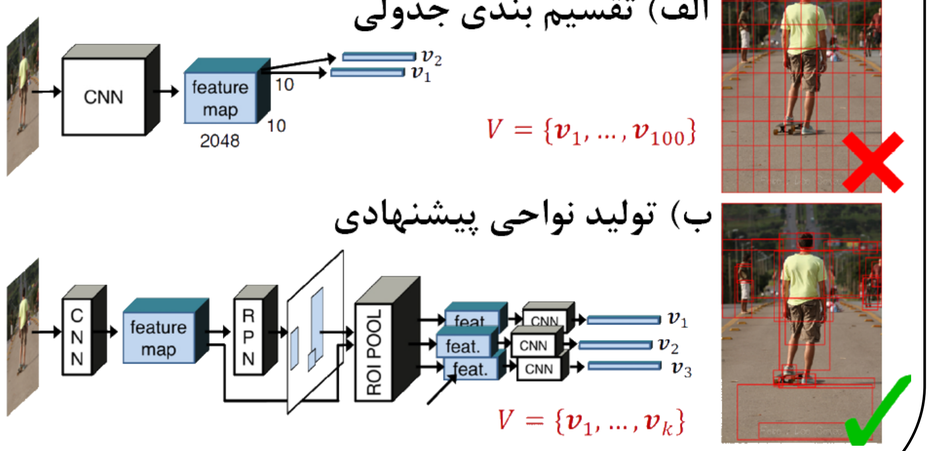
۲- یافتن نواحی مرتبط در تصویر V

۳- یادگیری تابع توجه f

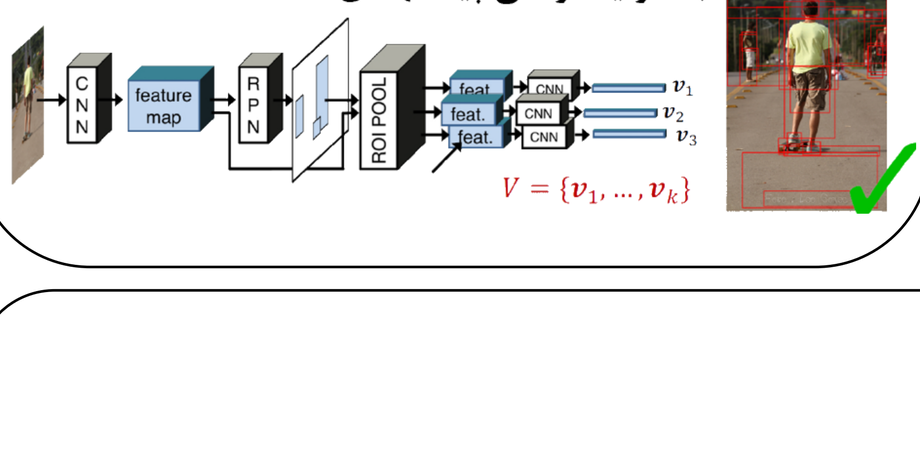
$$\hat{v} = f(h, V)$$

روش های ناحیه بندی تصویر [۳]

الف) تقسیم بندی جدولی



ب) تولید نواحی پیشنهادی



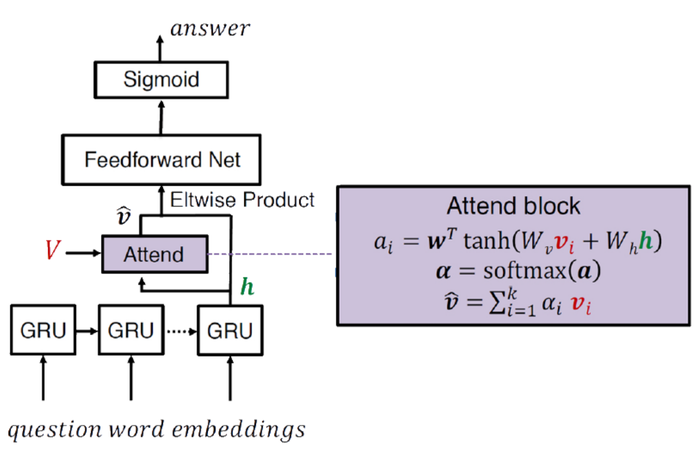
ویژگی‌های روش پیشنهادی:

۱- به کارگیری مکانیزم توجه دوگانه

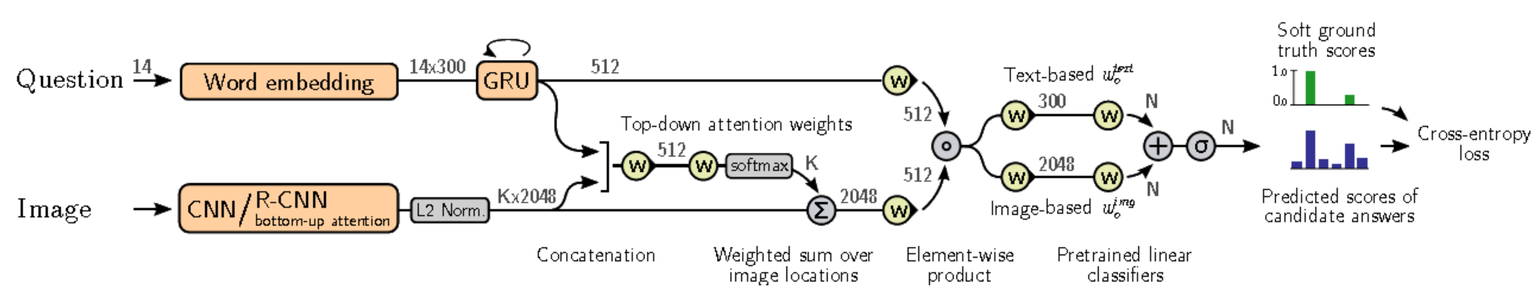
۲- معماری ساده و سبک

۳- امکان بهره گیری از مزایای انتقال آموزش

مکانیزم توجه پیشنهادی



معماری شبکه‌ی پیشنهادی



تحلیل نتایج

به لطف مکانیزم توجه پیشنهادی و نیز برخی نوآوری‌های فنی در پیاده‌سازی شبکه‌ی عصبی مورد نظر، روش پیشنهادی موفق به کسب نتایجی در محدوده‌ی روش‌های state of the art گردیده و همچنین نسبت به مقاله‌ی مرجع بیش از یک درصد بهبود عملکرد را به ارمغان آورده است

	همه	سایر	شمارشی	دودویی
HieCoAtt	56.3	46.8	36.5	76.0
MCB-Att	57.9	49.5	36.2	76.6
Bottom-up	62.4	54.5	42.7	78.6
روش پیشنهادی	**63.2**	**55.8**	**42.8**	**80.3**

+1.2%

کلمات کلیدی

Visual Question Answering (VQA), Computer Vision (CV), Natural Language Processing (NLP), joint embedding, attention mechanisms, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Region-based Convolutional Neural Network (R-CNN), Fast R-CNN, Faster R-CNN, Bottom-Up and Top-Down Attention, LSTM, GRU, gated tanh, Drop out, Batch Normalization, Regularization, Word Embedding (WE), Word2Vec, Glove, Deep Learning, Machine Learning, artificial intelligence, AI

منابع

[1]. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in The IEEE International Conference on Computer Vision (ICCV), 2015.

[2]. Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: a survey of methods and data sets," Computer Vision and Image Understanding, 2017

[3]. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L.Zhang, "Bottom-up and top-down attention for image captioning and VQA," arXiv Preprint, arXiv:1707.07998, 2017