Visual question answering (VQA) untuk objek pariwisata monas menggunakan deep learning = Visual question answering (VQA) for monas tourism objects using deep learning.

Siregar, Ahmad Hasan, author

Abstrak

Visual Question Answering (VQA) adalah sebuah tugas pembelajaran mesin di mana diberikan pasangan gambar dan pertanyaan visual dalam bahasa natural, mesin harus memprediksi jawaban yang tepat. Kesulitan dari tugas VQA adalah masukan melibatkan dua media informasi (modality), yaitu gambar dan teks. VQA masih merupakan bidang penelitian yang aktif yang setiap tahunnya berbagai peneliti mempublikasikan model VQA, sebuah respons terhadap VQA challenge, dengan akurasi state-of-the-art tahun 2016 di 66.47% dan akurasi state-of-ther-art terakhir tahun 2019 masih di 75.23%. Diketahui bahwa tidak ada data VQA yang tersedia dalam bahasa Indonesia, data VQA Monas disusun dalam bahasa tersebut dengan fokus Monas sebagai konteksnya yang merupakan objek pariwisata di Jakarta. Metode pembelajaran mesin multimodal diajukan menggunakan CNN sebagai image embedding dan beberapa teknik di bidang linguistik sebagai sentence embedding, yaitu Bag-of-Words, fastText, BERT, dan [Bi-]LSTM. Akurasi sebesar 68.39% dicapai pada model dengan performa terbaik. Studi ablasi juga dilaporkan untuk menganalisis pengaruh dari sebuah lapisan individu terhadap akurasi model secara keseluruhan.
.....Visual Question Answering (VQA) is a machine learning task, given a pair of image and natural language visual question, machine should predict an accurate answer. Difficulty of VQA lies in the fact that the inputs has two information media (modality), i.e. image and text. VQA is an active research field as each year researchers still publish VQA models, a response to a VQA challenge, with state-of-the-art accuracy in 2016 at 66.47% and the latest state-of-the-art accuracy in 2019 is still at 75.23%. Known that there is no VQA dataset available in Bahasa Indonesia, a VQA Monas dataset is established in that language with focus on Monas as the context, a Jakarta tourism object. A multimodal machine learning method is proposed based on CNN for image embedding and several techniques in linguistic field for sentence embedding, i.e. Bag-of-Words, fastText, BERT, and [Bi-]LSTM. Accuracy of 68.39% is achieved on the best performing model. Ablation studies is also shown to analyze the impact of a layer to model’s accuracy as a whole.