CSCI 669: Project Survey

Hayley Song
USC ID: 5276048267
haejinso@usc.edu

1 Related Works

A major goal of our project is to express the semantics of an image (objects, attributes and relations) in the representation that can communicate with external data from the knowledge graph. Inspired by how our brain uses a conceptual representation, we aim to find the correct level of abstraction through which information can flow through different data modalities. In this section, we first introduce studies that provide neural evidence that our brain does indeed process multimodal information using a common semantic representation. Then, we introduce recent studies on the joint representation of image and text in a single semantic domain. Lastly, we introduce previous works that have incorporated external knowledge into visual task pipelines.

1.1 Neural evidence of a common representation of knowledge across multiple modalities

In neuroscience, studies on fMRI images have strongly suggested a common semantic representation of knowledge across different types of stimulus such as images and written texts. Shinkareva et al. (2011) shows the same neural patterns are observed when an image of an object and the word that identifies the object are presented. Simanova et al. (2014) expands the studies to 4 stimulus modalities (images, spoken and written names, and natural sounds) and shows a consistent neural activation pattern appears across all modalities. Simanova et al. (2014) further shows the same brain regions could predict semantic categories in the free recall session, which implies the common representation is accessed by all four modalities during information retrieval process as well. These studies suggest our brain maps knowledge from different modalities to a common representation space that is abstracted away from the format of inputs

1.2 Semantic representation of images

Driven by these intuitions, the AI community also have explored the idea of a semantic representation for visual (images or video) and linguistic inputs (spoken or written words), often referred to as an âĂIJembeddingâĂİ method. Most state-of-the-art embedding methods are based on neural networks such as Convolutional Neural Network (CNN) (LeCun et al., 2015), Recurrent Neural Network (RNN) (LeCun et al., 2015) and Autoencoder. These embeddings can be learned jointly with the network weights during the training process.

Another way to represent the semantics of an image is the semantic parsing, which is the process of decomposing images into semantic components and constructing the structure representation of the input. Johnson et al. (2015) introduces a data structure that supports this semantic parsing by encoding object instances, attributes of objects, and relationships between objects from images. Object instances may be any entity like people (“man”), locations(“swimming pool”), and parts of other objects (“nose”). Attributes describe the properties of an object such as its shape (“triangle”), color (“green”), or pose (“bent”). Relationships encode how two objects are related, for example, spatially (“behindâĂİ, “on top of”), hierarchically (“part of”), or through interactions (“a boy riding a monocycle”). There are different ways to perform the semantic parsing: Johnson et al. (2015) introduces a conditional random field model for reasoning about possible groundings of scene graphs, and Schuster et al. (2015) introduces a rule-based and classifier-based scene parser. More recently, Liang et al. (2017) introduces a method based on variation-structured reinforcement learning that incorporates a global interdependency in addition to the entities and their relations. The scene graph representation has been shown to improve image captioning (Farhadi et al., 2009; Gupta and Davis, 2008), and semantic image retrieval (Johnson et al., 2015; Schuster et al., 2015).

Krishna et al. (2016) is the first dataset that represents images semantically in a structured formalized way, in the form that is widely used in knowledge base representations in NLP. It uses the scene graph structure to express images in terms of their semantic components and maps each semantic component to its corresponding sense in WordNet (Miller, 1995). This mapping connects all images in Visual Genome and provides contextual information from multiple images. It also naturally provides a shared representation of a concept in its visual and linguistic form.

1.3 Use external knowledge for visual tasks

There has been an increasing interest in leveraging external knowledge to perform reasoning beyond the image contents. Li et al. (2017) incorporates external knowledge with dynamic memory networks to answer open-domain visual questions. Specifically, it embeds the KG in a continuous vector space while preserving the entity-relation structure, and uses the input image-question pair as the trigger to retrieve the relevant information from the embedded KG. Then, the dynamic networks perform reasoning over the contents of the input and the KG to generate an answer. For semantic image segmentation, Zand et al. (2016) proposes an ontology-based semantic segmentation approach that jointly models image segmentation and object detection. Specifically, a Dirichlet process mixture model is used to embed the image in a reduced feature space, and multiple CRFs are used to learn the weights of these features. The object inference is passed to an ontology model. This model resembles how human uses a combination of visual cues, context and rule-based reasoning to understand an image. For image classification, Marino et al. (2016) studies the use of structure prior knowledge in the form of knowledge graphs and introduces the Graph Search Neural Network as an end-to-end method to efficiently incorporate large KGs into a classification pipeline. This work is most similar to our project, yet since our dataset consists of overhead satellite images with vastly different types of objects and relations, we face a different challenge.

References

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pages 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772.

Abhinav Gupta and Larry S. Davis. 2008. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. Springer Berlin Heidelberg, Berlin, Heidelberg, volume 5302, pages 16–29. https://doi.org/10.1007/978-3-540-88682-2_3.

J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. 2015. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 3668–3678. https://doi.org/10.1109/CVPR.2015.7298990.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv:1602.07332 [cs]* ArXiv: 1602.07332. http://arxiv.org/abs/1602.07332.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553):436–444. https://doi.org/10.1038/nature14539.

Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks. *arXiv:1712.00733 [cs]* ArXiv: 1712.00733. http://arxiv.org/abs/1712.00733.

Xiaodan Liang, Lisa Lee, and Eric P. Xing. 2017. Deep Variation-structured Reinforcement Learning for Visual Relationship and Attribute Detection. *arXiv:1703.03054 [cs]* http://arxiv.org/abs/1703.03054.

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2016. The More You Know: Using Knowledge Graphs for Image Classification. *arXiv:1612.04844 [cs]* ArXiv: 1612.04844. http://arxiv.org/abs/1612.04844.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38(11):39–41. https://doi.org/10.1145/219717.219748.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*. Association for Computational Linguistics, pages 70–80. https://doi.org/10.18653/v1/W15-2812.

Svetlana V. Shinkareva, Vicente L. Malave, Robert A. Mason, Tom M. Mitchell, and Marcel Adam Just. 2011. Commonality of neural representations of words and pictures. *NeuroImage* 54(3):2418–2425. https://doi.org/10.1016/j.neuroimage.2010.10.042.

Irina Simanova, Peter Hagoort, Robert Oostenveld, and Marcel A. J. van Gerven. 2014. Modality-Independent Decoding of Semantic Information from the Human Brain. *Cerebral Cortex* 24(2):426–434. https://doi.org/10.1093/cercor/bhs324.

M. Zand, S. Doraisamy, A. Abdul Halin, and M. R. Mustaffa. 2016. Ontology-Based Semantic Image Segmentation Using Mixture Models and Multiple CRFs. *IEEE Transactions on Image Processing* 25(7):3233–3248. https://doi.org/10.1109/TIP.2016.2552401.