A Survey into the Development of Multimodal NLP

Shreyans Sethi
University of California - Berkeley
Info 159 - Natural Language Processing

Abstract

[1] Multimodal NLP has become an important field of research over the last decade with models in both NLP and Computer Vision becoming so powerful that attempts to combine them have become common. Multimodal NLP can encompass many different modalities along with a variety of tasks (multimodal sentiment analysis, translation, generation, etc.). This survey will focus specifically on the methods and architectures by which textual and visual inputs are processed and used to influence each other. It will mainly use image captioning, visual question answering and visual commonsense reasoning as the tasks through which to explain these architectural advancements. With some of the papers mentioned in this survey having been published within a week of its writing, the Multimodal NLP field is very much active.

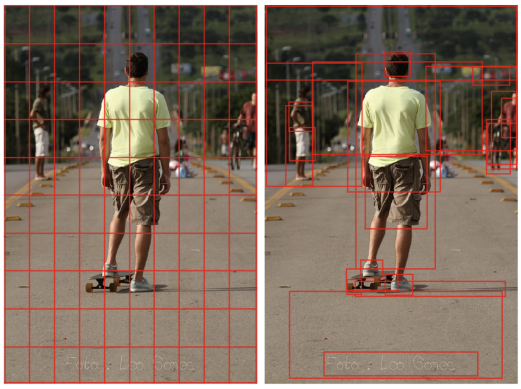

Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Figure 1: (Anderson et al., 2018)

1 Image Captioning and Encoder-Decoder Architecture

Image to text generation, or *image captioning*, was one of the first big focuses in multimodal NLP with (Vinyals et al., 2015) being historical as the first paper to build a neural caption generator. The paper used a deep convolutional network for the image-encoder and took its last layer outputs, or image embeddings, as input to an RNN (LSTM-based) which then sequentially generated the caption. By doing so, it combined state of the art subnetworks for vision and language with the advantage that these were already pre-trained on large corpora. It was trained to maximize the probability of a target sequence S given some image I and became the state-of-the-art for a non-template based approach to image captioning.

Despite being an important paper, there were immediate limitations found that were promptly addressed by other research. Anderson (2018) introduced a combined bottom-up and top-down mechanism that allows for image regions to be processed not as a rectangular grid but on the level of objects (Figure 1). The bottom-up approach uses a Faster R-CNN to propose a set of image regions on the object level (as bounding boxes) and the top down mechanism uses two LSTMs - one for attention weighing on the image regions and one for language generation.

Li (2017) similarly notes the lack of local and global features in Vinyals (2015) and concurrently uses a CNN (specifically, VGG16) for global feature extraction and a Faster R-CNN for object detection, before using LSTMs for feature weighting (across these two feature categories) and language generation. It is seen that the CNN-LSTM encoder-decoder model became generally adopted across all these papers for image captioning with changes made to focus more on specific object-detection.

[1] Word Count (Excluding Citations): 2195

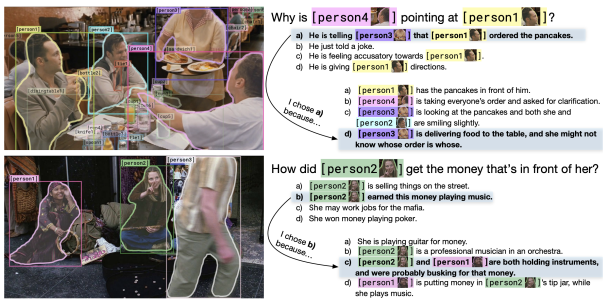

Figure 2: VCR Examples (Zellers et al., 2019)

2 VQA and Visual Commonsense Reasoning

With strong advancements being made in image captioning, with MSCOCO and Flickr30k being the common datasets for evaluation, there was also research into other useful visual-linguistic tasks. Agrawal (2016) created the task of "Visual Question Answering" (VQA) in which an image and a natural language question is fed as input to a model which must give an accurate language answer. The questions target different image areas and also require the model to understand the underlying context. Zellers (2019) built on this idea with the task of 'Visual Commonsense Reasoning' (VCR) which has 3 forms of tasks: Q→A, QA→R, and Q→AR [Q = Question, A = Answer, R = Reasoning]. The last of these tasks is the most challenging as the model must take an image and a challenging question and generate an answer and additionally, a rationale.

Both these tasks create an additional challenge over image captioning in that competent models must make complex inferences about the semantic relationships in the image and, since the inputs are now multimodal, they must generate embeddings for both types of input. These papers were extremely important in the push for *visual grounding* - Implicitly linking parts of the question to specific regions within the images. Huang (2019) came up with an approach to address VQA's need for object-level grounding by generating a caption for the input image and calculating its cosine similarity to the word-tokens of the input question. For highly similar pairs, the question's word-token can then be linked back to the entity in the image that generated that part of the caption. The paper used these grounding relationships to weigh the object and question features, which are processed to generate in the output space. On VQA v2, this model achieved a 67.41% accuracy.

However, as explored below, the norm to address multimodal inputs and VQA/VCR quickly became transformer-based models, leading to two schools of thought: early-fusion and late-fusion.

3 Transformer Models and Fusion Types

In 2017/2018, the research surrounding Transformers and BERT was released, showing the advantages of using self-attention and cross-attention on sequential data to model long range dependencies. Immediately, these methods became adopted in multimodal models as they integrated with the concept of visual grounding, with word-tokens being able to place a higher attention weight on an input region without having vanishing gradient issues. Early-fusion and late-fusion became the two schools of thought for integrating transformers for multimodal tasks which take text and image input (E.g. VCR/VQA). In early-fusion, a single stream architecture is used in which one transformer takes in both the image and textual embeddings, allowing for early cross modal interactions, and generates hidden layers that can then be used for the desired task. In late-fusion, two networks (often transformers) are used to separately process text and images and then a third transformer is used to combine the hidden layer information, and generate in the desired output space.

ViLBERT and LXMERT were two of the popular models that followed the two-stream architecture. ViLBERT (Vision and Language BERT) believed that using separate processing streams allowed for variable depths for each modality and the integration of large, unimodal models (Lu et al., 2019). The model uses more processing layers for the text stream and generates BERT embeddings which are then combined using co-attentional layers with the image regions (generated using Faster R-CNN). LXMERT also uses a 2 stream, late-fusion design but adds additional pre-training tasks (Specifically image question answering and region-of-interest feature regression) along with more cross-modality attention layers (Tan and Bansal, 2019). ViLBERT and LXMERT achieve 70.55% and 72.5% accuracy on VQA respectively and ViLBERT achieves 54.04% on the Q→AR part of VCR.

B2T2, VisualBERT, VL-BERT, Unicoder VL, and UNITER are all models released in 2019/2020 and prefer the single stream architecture as it gives the model sufficient capacity to learn multimodal

information. B2T2 (Bouncing Boxes in Text Transformer) embeds visual features on the same level as the word tokens and also includes features based on image bounding boxes. All these embeddings will be processed and contextualized together through a BERT network and then the CLS token is used for classification and generation (Alberti et al, 2019). VisualBERT also uses a single cross-modal Transformer but focuses more on syntactic grounding and Li (2019/2020) shows that the VisualBERT model is able to process a phrase like "man walking" and the attention heads for the walking token will attend to the image region of the man, because even without supervision it understands that there exists an 'nsubj' relation there.

VL-BERT is extremely similar to the other early-fusion models but it pre-trains on text-only datasets in addition to visual-linguistic ones as Su (2019) believed this joint pre-training helps the model deal with longer, complex sentences. Furthermore, the parameters of the Faster R-CNN used to find image regions are not frozen but actually updated. Unicoder-VL also adhered to the early-fusion school of thought but differed to other models in that it was trained using the objective of taking a caption and identifying the corresponding image from a set of candidates (Li et al., 2020). Also, Unicoder-VL has a fine tuning layer which means it is often fine-tuned for image-text retrieval tasks. UNITER, another single-stream, early fusion model, differentiates itself by using *conditional masking* while pre-training - instead of randomly masking out parts of both the image and text, it does masked language/region modeling based on full observation of the other modality (Chen et al., 2020)

Generally, the early fusion models perform better on VQA and VCR tasks with B2T2, VisualBERT, VL-BER, Unicoder VL, and UNITER having Q→AR scores of 55.0%, 52.4%, 59.7%, 54.9% and 58.20% respectively and UNITER being considered state-of-the-art on VCR. Despite the larger number of parameters and amount of compute required to do so, the models are able to benefit from learning cross-modal interactions earlier on.

4 Other Fusion Ideas

An interesting approach by Yu (2019) was to design a self-attention (SA) unit that models intra-modal connections (e.g. Image region-to-region) and a guide-attention (GA) unit that models inter-modal connections (word token to image region) using cross attention. These SA and GA units can be modularly combined to build a network layer, which can then be cascaded to build a "Modular Co-Attention Network" (MCAN). This MCAN has the benefits of both the 1-stream and 2-stream models and as such, is considered state-of-the-art performance on the VQA-2.0 dataset. Furthermore, Wu (2021) used an MCAN layout which uses multiple layers of GA units for co-attention but has some that take the previous GA-unit's output as its input whereas others re-input the original text/image embeddings. This layout with this architecture became very strong at fake news detection and could tell if the caption and/or the image input showed signs of being fake.

Similarly, Nagrani (2021) also tries to reap the benefits of both early and late fusion by allowing free attention flow within a modality at a given layer but introducing bottlenecks to restrict cross-modal attention flow between tokens. This means the model learns to condense relevant information in a modality and share what is necessary, which keeps parameter numbers down while allowing cross-modal interactions.

Liu (2018) approaches multimodal fusion as a dimension-reduction problem: For each modality, it passes in the tokens through their own networks to generate contextual embeddings. It then has a set of modality-specific factors (akin to basis vectors) which are used to decompose the embeddings into a smaller dimensionality. These low-rank embeddings can then be fused together to create a multimodal representation which will then be used to predict/generate. As such, fusion can happen at a lower parameter and compute cost.

5 Graph-Based Models and Other Novel Approaches

Beyond using the transformer-based encoder-decoder model, some papers have also incorporated graphs into their architecture to model complex relationships. Li (2019) encodes each image into a graph with detected objects as nodes, and then uses a relation encoder (a neural-network classifier) to draw semantic, spatial, or implicit relations between objects as edges. These relations are then embedded and concatenated with the question embeddings to complete the task. Shi (2020) uses a similar idea but creates a joint graph between the detected objects and the predicates in the textual

input before passing it into a graph convolutional network.

Wu (2019) has a novel approach in that it first generates a caption for the image which will help it answer the image-question. For example, for an image of a surfer and a question, "What color is the board?", it will generate "A young man riding a wave on a blue surfboard" based on the joint question and image embeddings. It will then process the caption to get an answer by using standard NLP QA models.

6 Multimodal Analysis Techniques

With the sudden burst of all these multimodal models, there has also been research into how to evaluate these models and determine the extent to which one modality influences the other.

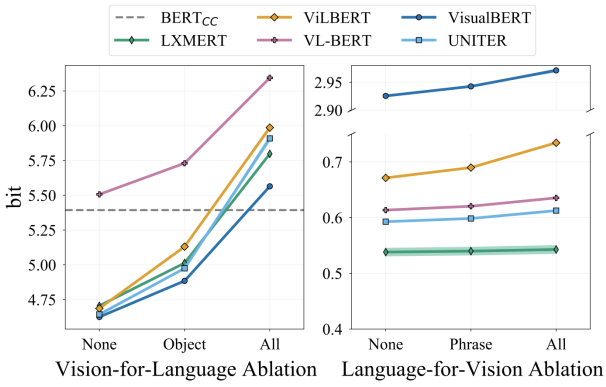

Figure 3: (Frank et al., 2021)

Ablation studies have been the main format to do such investigations - Part or all of one modality is hidden and it is seen how much performance suffers as a result. Frank (2021) conducted this strategy on ViLBERT, VisualBERT, LXMERT, VL-BERT and UNITER. It found across the board that the performance difference between masking nothing and masking the object region needed for the task is not as significant as the difference between masking the object region and masking everything (Figure 3). This suggests that models are not using grounding object information as much as initially suggested. Furthermore, performance suffers significantly more when visual input is ablated as compared to when language input is ablated, suggesting an asymmetrical modal relationship.

Ma (2022) builds on this research by proposing a training strategy that optimizes the transformer models on modal-complete and modal-incomplete data using multi-task optimization, thus making the model more robust to missing modalities. It also suggests adding a policy parameter to each layer of the model determining whether the two modalities should be fused on the given layer or not. By using the paper's differentiable search algorithm, models can go through this 2^L search space (for L layers) and find the optimal fusion strategy, a possible solution to the aforementioned late-fusion vs early-fusion debate.

7 Multimodal NLP's Modern Issue: Lack of Data

One of the issues with scaling multimodal models is the lack of available data to train the models on, especially for more complicated tasks such as VQA which require an image-question-answer triplet. Li (2020) suggests an approach to this problem by pre-training models on purely textual data and purely visual data, using *detector tags* to note when a similar concept is present in both (E.g. photo of a cake and the word 'cake') and if detected, then combining parameters for these tokens from both transformers.

Alternatively, new research by Changpinyo (2022) suggests taking image captions, extracting candidate answers based on the caption and then using a neural model to create a question for which it would have that candidate answer. This method is able to generate questions for which the answer is not in the caption verbatim and which requires a high level of inference about the image, and has generated hundreds of thousands of training examples to create a VQ^2A dataset. It is possible similar research could one day lead to the VCR dataset being grown in the same way.

8 References

*(Papers from ACL, EMNLP, NAACL, EACL, AACL, Transactions of the ACL or Computational Linguistics are marked with a ** at the end of the citation)*

1. Agrawal, Aishwarya, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. "VQA: Visual Question Answering." International Journal of Computer Vision 123, no. 1 (2016): 4–31. https://doi.org/10.1007/s11263-016-0966-6.

2. Alberti, Chris, Jeffrey Ling, Michael Collins, and David Reitter. "Fusion of detected objects in text for visual question answering." arXiv preprint arXiv:1908.05054 (2019).

3. Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6077-6086. 2018.

4. Changpinyo, Soravit, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding and Radu Soricut. "All You May Need for VQA are Image Captions." ArXiv abs/2205.01883 (2022): n. Pag. **

5. Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "Uniter: Universal image-text representation learning." In European conference on computer vision, pp. 104-120. Springer, Cham, 2020.

6. Frank, Stella, Emanuele Bugliarello, and Desmond Elliott. "Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers." arXiv preprint arXiv:2109.04448 (2021). **

7. Huang, Pingping, Jianhui Huang, Yuqing Guo, Min Qiao, and Yong Zhu. "Multi-Grained Attention with Object-Level Grounding for Visual Question Answering." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/p19-1349. **

8. Li, Gen, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 11336-11344. 2020.

9. Li, Linghui, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. "GLA: Global–Local Attention for Image Description." IEEE Transactions on Multimedia 20, no. 3 (2018): 726–37. https://doi.org/10.1109/tmm.2017.2751140.

10. Li, Linjie, Zhe Gan, Yu Cheng, and Jingjing Liu. "Relation-aware graph attention network for visual question answering." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10313-10322. 2019.

11. Li, Liunian Harold, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. "Unsupervised vision-and-language pre-training without parallel images and captions." arXiv preprint arXiv:2010.12831 (2020). **

12. Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

13. Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. "What Does Bert with Vision Look at?" Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.469.

14. Liu, Zhun, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. "Efficient low-rank multimodal fusion with modality-specific factors." arXiv preprint arXiv:1806.00064 (2018). **

15. Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in neural information processing systems 32 (2019). **

16. Ma, Mengmeng, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. "Are Multimodal Transformers Robust to Missing Modality?." arXiv preprint arXiv:2204.05454 (2022).

17. Nagrani, Arsha, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. "Attention bottlenecks for multimodal fusion." Advances in Neural Information Processing Systems 34 (2021).

18. Shi, Zhan, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. "Improving Image Captioning with Better Use of Caption." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.664. **

19. Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. "Vl-bert:

Pre-training of generic visual-linguistic representations." arXiv preprint arXiv:1908.08530 (2019).

20. Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." arXiv preprint arXiv:1908.07490 (2019). **

21. Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.

22. Wu, Jialin, Zeyuan Hu, and Raymond Mooney. "Generating Question Relevant Captions to Aid Visual Question Answering." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/p19-1348. **

23. Wu, Yang, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. "Multimodal Fusion with Co-Attention Networks for Fake News Detection." Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021. https://doi.org/10.18653/v1/2021.findings-acl.226. **

24. Yu, Zhou, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. "Deep modular co-attention networks for visual question answering." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6281-6290. 2019.

25. Zellers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "From recognition to cognition: Visual commonsense reasoning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6720-6731. 2019.