Ensemble based Discriminative Models for Visual Dialog Challenge 2018

Shubham Agarwal *
Raghav Goyal
Heriot Watt University, Edinburgh, UK
* Adeptmind Scholar, Adeptmind Inc., Montreal, Canada

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Overview

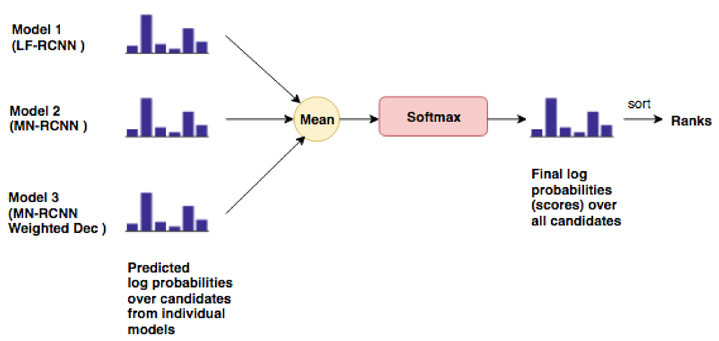
Goal: Visual Dialog Challenge - Conversational dialogue about visual content

Contributions:

- Ensemble of three discriminative models
  - LF-RCNN
  - MN-RCNN
  - MN-RCNN-Wt
- Final submission on 'test-std' split achieves NDCG score of 55.46 and MRR value of 63.77
- Faster RCNN [4] with ResNet-101 trained on Visual genome dataset [1,5] for object level image representations
- Memory Networks compared to concatenation for encoding dialogue history

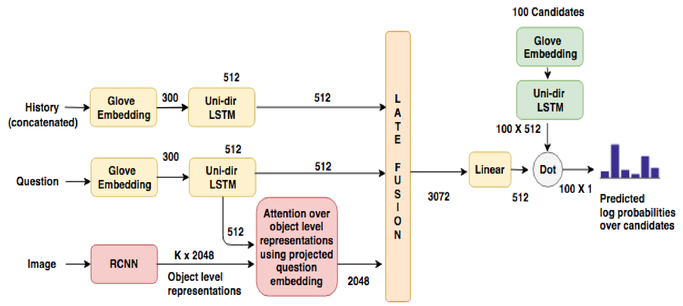
Ensembling

- Ensembled final layer's log-softmax output - distribution over candidate answers
- Mean of log probabilities from individual models
- Also tried taking maximum of results but mean performed better

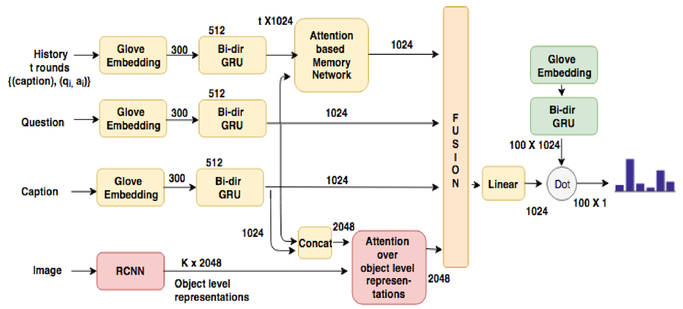


Individual Components

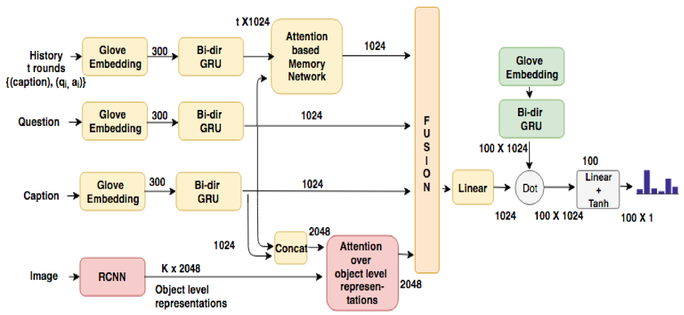
- **Late Fusion R-CNN (LF-RCNN)**
  - Late Fusion encoder [2] with concatenated history; Glove embeddings frozen, not fine-tuned
  - Object-level features are weighed using only question embeddings



- **Memory Network R-CNN (MN-RCNN)**
  - Memory Network encoder with bi-directional GRUs and word embeddings fine-tuned
  - Object-level features weighed by question and caption embeddings



- **Weighted MN-R-CNN (MN-RCNN-Wt)**
  - Additional gated linear layer applied to the dot product of candidate answer and encoder output



Results:

- Evaluation results of individual components on validation set

Model	MRR	R@1	R@5	R@10	Mean
Baseline	57.57	42.98	74.64	84.91	5.48
LF-RCNN	61.94	48.08	79.04	88.23	4.61
MN-RCNN	62.99	49.07	80.13	88.74	4.45
MN-RCNN-Wt	63.11	49.29	80.10	89.09	4.43

- Results for the challenge on test-std.
- Ensemble of best performing models for the final submission

Model	NDCG (x 100)	MRR (x 100)	R@1	R@5	R@10	Mean
LF-RCNN	51.69	61.03	47.03	77.83	87.55	4.70
MN-RCNN	53.59	61.25	46.78	79.43	87.93	4.63
MN-RCNN-Wt	53.20	61.50	47.10	78.7	88.38	4.54
Ensemble (all three)	**55.46**	**63.77**	**49.8**	**81.22**	**90.03**	**4.11**

Conclusions

- Object level image representations using Faster RCNN gave huge uplift
- Bi-directional GRUs constantly performed better than uni-directional LSTMs
- Memory Networks outperformed Late fusion encoders for encoding conversational history
- Fine-tuning Glove embeddings performed better than their counterparts
- Mean performed better than maximum for ensembling

References

[1] 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 visual question answering." In CVPR, 2018.
[2] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. "Visual Dialog." In CVPR, 2017.
[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In IJCV 2017.
[4] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks." In NIPS 2015.
[5] D. Teney, P. Anderson, X. He, and A. van den Hengel. "Tips and tricks for visual question answering: Learnings from the 2017 challenge."