# Artificial Neural Network Final Project

Minlie Huang

aihuang@tsinghua.edu.cn

# Outline

- Overview
- Project Topics
- Important Dates
- Submission Guideline
- Grading

# Overview

- **Apply what you have learned to a real problem of your interest**
- **A chance to show your engineering/research potential!**
- **Each team consists of 2-3 students**
- Two types of projects are welcome:
  - **Applications**: Apply existing models to an existing or new problem
  - **Innovations**: Build a new model (algorithm) with nn, or a new variant of existing models
- Making contributions to CoTK will be awarded with **bonus** (at most 5 points)
- A **reproduction challenge** are also welcome this semester
- It can be your research project undergoing, but must relate to
  - **Neural Network, Deep Learning**
  - **Deep reinforcement learning**

# Recommended Project Topics

- Object Recognition
- Object Detection
- Image Captioning
- Image Generation
- Text Classification
- Machine Comprehension
- Language Generation
- Other Options

**You are not limited to these topics**

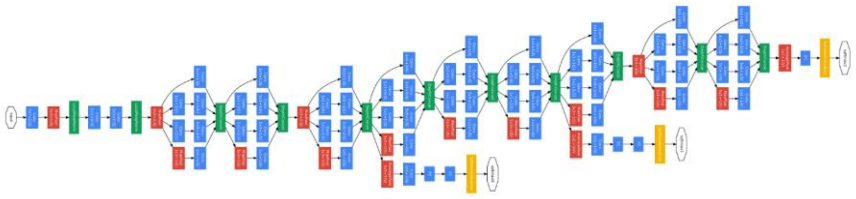
# Object Recognition

- Also known as image classification
- Single label/Multi-label classification
- Well-studied, hard to improve performance
- But new ideas, new tasks are worth of study


Cifar-10 Dataset

# Object Recognition Models

- VGGnet, GoogleNet, ResNet, DenseNet
- Batch Normalization, Dropout
- Fine-grained object recognition



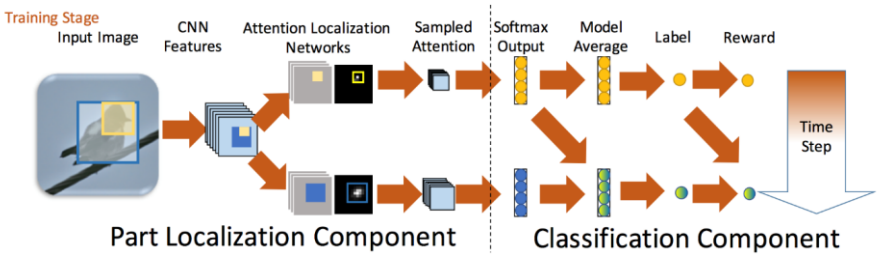
**Input:** Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: $\gamma$, $\beta$
**Output:** $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$



# Object Recognition Datasets

- For general object recognition, we recommend the established Cifar10 dataset
  - It is a subset of the tiny images dataset and consists of 60,000 32 × 32 color images containing one of 10 object classes, with 6000 images per class.
- Cifar100
  - This dataset is just like CIFAR-10, except it has 100 classes containing 600 images each.
- For fine-grained recognition, we recommend Stanford Dogs dataset Stanford Dogs dataset
  - includes over 22,000 annotated images of dogs belonging to 120 species.
- Choose your datasets if you want a novel task

# Object Recognition Results

Cifar10, Cifar10 with augmentation error rates

| Method | Depth | Params | C10 | C10+ |
|---|---|---|---|---|
| Network in Network [22] | - | - | 10.41 | 8.81 |
| All-CNN [31] | - | - | 9.08 | 7.25 |
| Deeply Supervised Net [20] | - | - | 9.69 | 7.97 |
| Highway Network [33] | - | - | - | 7.72 |
| FractalNet [17] | 21 | 38.6M | 10.18 | 5.22 |
| with Dropout/Drop-path | 21 | 38.6M | 7.33 | 4.60 |
| ResNet [11] | 110 | 1.7M | - | 6.61 |
| ResNet (reported by [13]) | 110 | 1.7M | 13.63 | 6.41 |
| ResNet with Stochastic Depth [13] | 110 | 1.7M | 11.66 | 5.23 |
| | 1202 | 10.2M | - | 4.91 |
| Wide ResNet [41] | 16 | 11.0M | - | 4.81 |
| | 28 | 36.5M | - | 4.17 |
| with Dropout | 16 | 2.7M | - | - |
| ResNet (pre-activation) [12] | 164 | 1.7M | 11.26* | 5.46 |
| | 1001 | 10.2M | 10.56* | 4.62 |
| DenseNet ($k = 12$) | 40 | 1.0M | **7.00** | 5.24 |
| DenseNet ($k = 12$) | 100 | 7.0M | **5.77** | **4.10** |
| DenseNet ($k = 24$) | 100 | 27.2M | **5.83** | **3.74** |
| DenseNet-BC ($k = 12$) | 100 | 0.8M | **5.92** | 4.51 |
| DenseNet-BC ($k = 24$) | 250 | 15.3M | **5.19** | **3.62** |
| DenseNet-BC ($k = 40$) | 190 | 25.6M | - | **3.46** |

| **Stanford Dogs** | Accuracy(%) | Acc w. Box(%) |
|---|---|---|
| Gavves *et al.* [37] | - | 50.1 |
| Simon & Rodner [39] | 68.1 | - |
| Sermanet *et al.* [17] | 76.8 | - |
| Zhang *et al.* [45] | 79.9 | - |
| Krause *et al.* [40] | 82.6 | - |
| Our Model | **88.9** | - |

Table 3. Comparison to related work on Stanford Dogs dataset.

| Method | Depth | Params | C10 | C100 |
|---|---|---|---|---|
| Wide ResNet | 28 | 36.5M | 3.8 | 18.3 |
| ResNeXt-29, 16x64d | 29 | 68.1M | 3.58 | 17.31 |
| DenseNet-BC (k=40) | 190 | 25.6M | 3.46 | 17.18 |
| C10 Model S-S-I | 26 | 26.2M | **2.86** | - |
| C100 Model S-E-I | 29 | 34.4M | - | **15.85** |

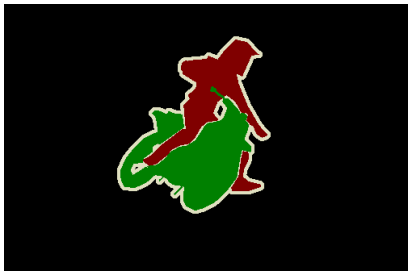
Test error (%) and model size on CIFAR.[1]

# Object Recognition

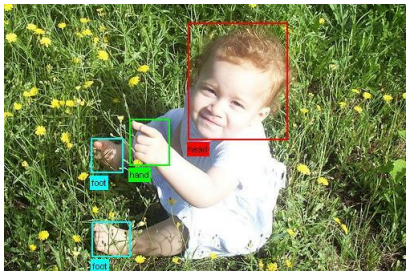

1. Gastaldi, X. (2017). Shake-Shake regularization. *arXiv preprint arXiv:1705.07485*.
2. Zagoruyko, S., & Komodakis, N. (2016). Wide Residual Networks. arXiv preprint arXiv:1605.07146.
3. Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. (2016). Deep networks with stochastic depth. arXiv preprint arXiv:1603.09382.
4. Liu, X., Xia, T., Wang, J., & Lin, Y. (2016). Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. arXiv preprint arXiv:1603.06765.
5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
6. Huang, G., Liu, Z., Maaten, L.V., & Weinberger, K.Q. (2016). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261-2269.

# Object Detection

- Detect instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos.
- Well-researched domains of object detection include person layout and pedestrian detection.



Segmentation

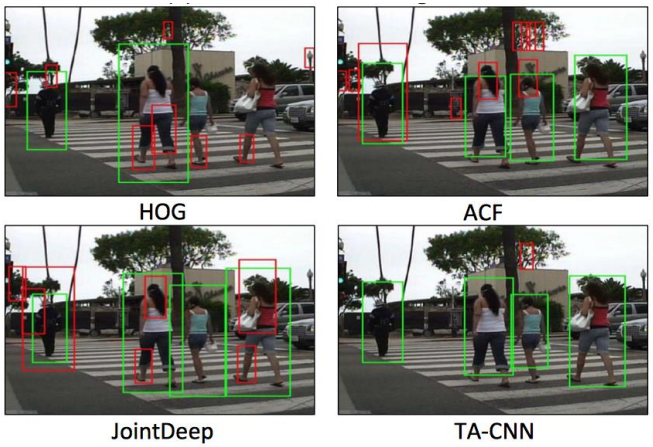


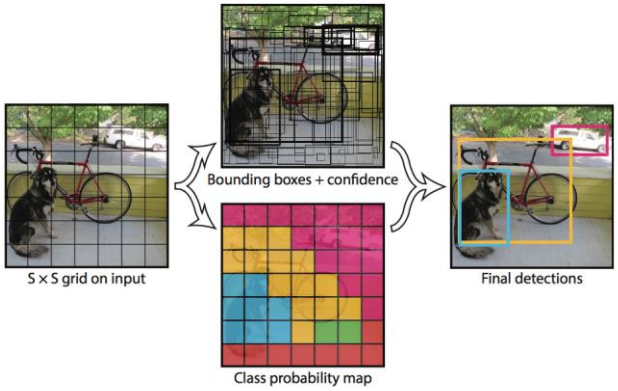

Person Layout

Pedestrian Detection

# Object Detection Models

- Mask R-CNN
- YOLO: Real-time Object Detection




R-CNN (2013.11) → OverFeat (ICLR' 14) → MultiBox (CVPR' 14) → SPP-Net (ECCV' 14) → MR-CNN (ICCV' 15) → DeepBox (ICCV' 15) → AttentionNet (ICCV' 15) →
Fast R-CNN (ICCV' 15) → DeepProposal (ICCV' 15) → Faster R-CNN (NIPS' 15) → OHEM (CVPR' 16) → YOLO v1 (CVPR' 16) → G-CNN (CVPR' 16) → AZNet (CVPR' 16) →
Inside-OutsideNet(ION) (CVPR' 16) → HyperNet (CVPR' 16) → CRAFT (CVPR' 16) → MultiPathNet(MPN) (BMVC' 16) → SSD (ECCV' 16) → GBDNet (ECCV' 16) →
CPF (ECCV' 16) → MS-CNN (ECCV' 16) → R-FCN (NIPS' 16) → PVANET (NIPSW' 16) → DeepID-Net (PAMI' 16) → NoC (TPAMI' 16) → DSSD (arXiv' 17) → TDM (CVPR' 17) → YOLO v2 (CVPR' 17) →
Feature Pyramid Net(FPN) (CVPR' 17) → RON (CVPR' 17) → DCN (ICCV' 17) → DeNet (ICCV' 17) → CoupleNet (ICCV' 17) → RetinaNet (ICCV' 17) → DSOD (ICCV' 17) →
Mask R-CNN (ICCV' 17) → SMN (ICCV' 17) → YOLO v3 (arXiv' 18) → SIN (CVPR' 18) → STDN (CVPR' 18) → RefineDet (CVPR' 18) → MLKP (CVPR' 18) → Relation-Net (CVPR' 18) →
Cascade R-CNN (CVPR' 18) → RFBNet (ECCV' 18) → CornerNet (ECCV' 18) → PFPNet (ECCV' 18) → Pelee (NIPS' 18) → HKRM (NIPS' 18) → R-DAD (AAAI' 19) → M2Det (AAAI' 19) ···

# Object Detection Datasets

- For general object detection, we recommend Pascal VOC2007 dataset.
  - In total there are 9,963 images, containing 24,640 annotated objects ranging from 20 classes like person, animals, vehicles and indoor objects.
- For human detection, we recommend Caltech Pedestrian Detection
  - includes about 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 bounding boxes and 2300 unique annotated pedestrians.

# State-of-the-art Object Detection

| usage of deformable convolution (# layers) | DeepLab | | class-aware RPN | | Faster R-CNN | | R-FCN | |
|---|---|---|---|---|---|---|---|---|
| | mIoU@V (%) | mIoU@C (%) | mAP@0.5 (%) | mAP@0.7 (%) | mAP@0.5 (%) | mAP@0.7 (%) | mAP@0.5 (%) | mAP@0.7 (%) |
| none (0, baseline) | 69.7 | 70.4 | 68.0 | 44.9 | 78.1 | 62.1 | 80.0 | 61.8 |
| res5c (1) | 73.9 | 73.5 | 73.5 | 54.4 | 78.6 | 63.8 | 80.6 | 63.0 |
| res5b,c (2) | 74.8 | 74.4 | 74.3 | 56.3 | 78.5 | 63.3 | 81.0 | 63.8 |
| res5a,b,c (3, default) | **75.2** | **75.2** | 74.5 | 57.2 | 78.6 | 63.3 | 81.4 | 64.7 |
| res5 & res4b22,b21,b20 (6) | 74.8 | 75.1 | **74.6** | **57.7** | **78.7** | **64.0** | **81.5** | **65.4** |

Table 1: Results of using deformable convolution in the last 1, 2, 3, and 6 convolutional layers (of $3 \times 3$ filter) in ResNet-101 feature extraction network. For *class-aware RPN*, *Faster R-CNN*, and *R-FCN*, we report result on VOC 2007 test.

| Method | mAP | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fast R-CNN[10] | 70.0 | 77.0 | 78.1 | 69.3 | 59.4 | 38.3 | 81.6 | 78.6 | 86.7 | 42.8 | 78.8 | 68.9 | 84.7 | 82.0 | 76.6 | 69.9 | 31.8 | 70.1 | 74.8 | 80.4 | 70.4 |
| Faster R-CNN[23] | 73.2 | 76.5 | 79.0 | 70.9 | 65.5 | 52.1 | 83.1 | 84.7 | 86.4 | 52.0 | **81.9** | 65.7 | 84.8 | 84.6 | 77.5 | 76.7 | 38.8 | 73.6 | 73.9 | 83.0 | 72.6 |
| SSD300[19] | 72.1 | 75.2 | 79.8 | 70.5 | 62.5 | 41.3 | 81.1 | 80.8 | 86.4 | 51.5 | 74.3 | 72.3 | 83.5 | 84.6 | 80.6 | 74.5 | 46.0 | 71.4 | 73.8 | 83.0 | 69.1 |
| SSD500[19] | 75.1 | 79.8 | 79.5 | 74.5 | 63.4 | 51.9 | 84.9 | **85.6** | 87.2 | 56.6 | 80.1 | 70.0 | 85.4 | 84.9 | 80.9 | 78.2 | 49.0 | **78.4** | 72.4 | 84.6 | 75.5 |
| RON320 | 74.2 | 75.7 | 79.4 | 74.8 | 66.1 | 53.2 | 83.7 | 83.6 | 85.8 | 55.8 | 79.5 | 69.5 | 84.5 | 81.7 | 83.1 | 76.1 | 49.2 | 73.8 | 75.2 | 80.3 | 72.5 |
| RON384 | 75.4 | 78.0 | 82.4 | 76.7 | 67.1 | 56.9 | 85.3 | 84.3 | 86.1 | 55.5 | 80.6 | 71.4 | 84.7 | 84.8 | 82.4 | 76.2 | 47.9 | 75.3 | 74.1 | 83.8 | 74.5 |
| RON320++ | 76.6 | 79.4 | **84.3** | 75.5 | **69.5** | 56.9 | 83.7 | 84.0 | **87.4** | 57.9 | 81.3 | **74.1** | 84.1 | 85.3 | **83.5** | 77.8 | 49.2 | 76.7 | **77.3** | **86.7** | 77.2 |
| RON384++ | **77.6** | **86.0** | 82.5 | **76.9** | 69.1 | **59.2** | **86.2** | 85.5 | 87.2 | **59.9** | 81.4 | 73.3 | **85.9** | **86.8** | 82.2 | **79.6** | **52.4** | 78.2 | 76.0 | 86.2 | **78.0** |

Detection results on PASCAL VOC 2007 test set. The entries with the best APs for each object category are bold-faced.[1]

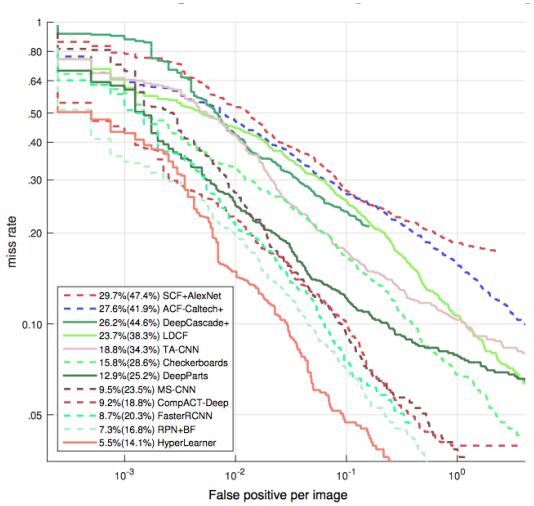



Figure 7. Detection quality on Caltech test set (reasonable, $MR^N_{-2}(MR^N_{-4})$), evaluated on the new annotations [34]. We achieve state-of-the-art results on both evaluation metrics.

# Object Detection References


1. Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., & Chen, Y. (2017). RON: Reverse Connection with Objectness Prior Networks for Object Detection. *arXiv preprint arXiv:1707.01691*.
2. Bell, S., Lawrence Zitnick, C., Bala, K., & Girshick, R. (2016). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2874-2883).
3. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, *39*(6), 1137-1149.

# Image Captioning

- Describe an image with a natural language description

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.


A stop sign is on a road with a mountain in the background.

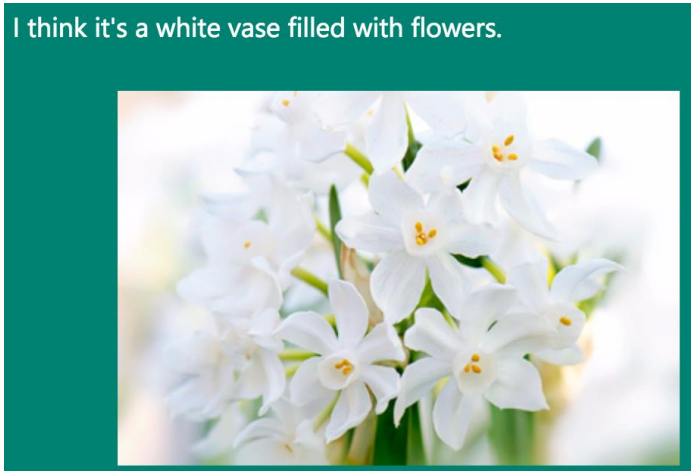
A little girl sitting on a bed with a teddy bear.

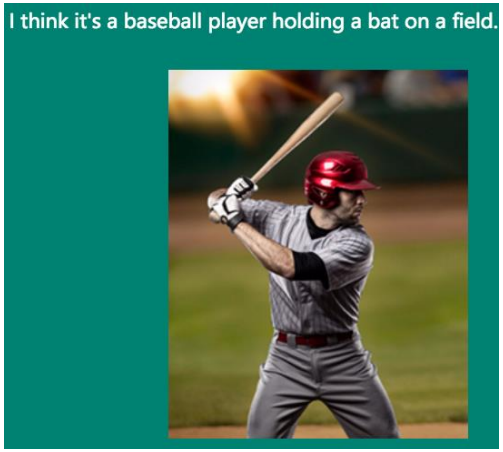
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

# Image Captioning

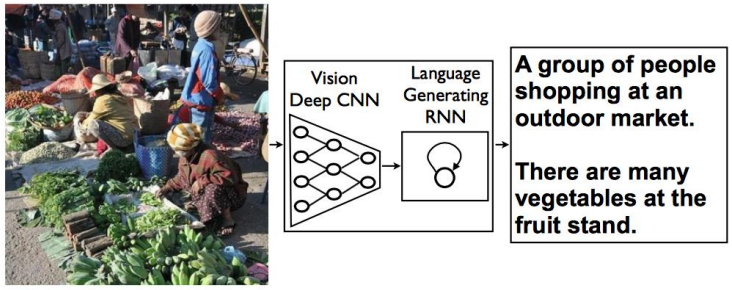
- https://www.captionbot.ai/

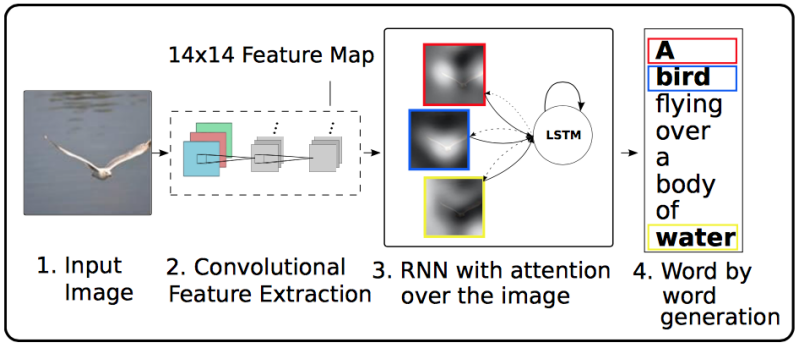



# Image Captioning

- Show and tell[1]
- Show, attend and tell[2]



# Image Captioning

- We recommend popular Flickr8k and Flickr30k dataset which has 8,000 and 30,000 images respectively, each paired with five different captions which provide clear descriptions of the salient entities and events.
- A more challenging dataset is Microsoft COCO. For more detailed information and evaluation protocol, you can refer to [3].

| COCO Online Testing Server C5 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Method | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE_L | CIDEr |
| Google NIC [23] | 0.713 | 0.542 | 0.407 | 0.309 | 0.254 | 0.530 | 0.943 |
| Hard-Attention[24] | 0.705 | 0.528 | 0.383 | 0.277 | 0.241 | 0.516 | 0.865 |
| AdaptiveAttention [15] | 0.735 | 0.569 | 0.429 | 0.323 | 0.258 | 0.541 | 1.001 |
| AdpativeAttention + CL (Ours) | 0.742 | 0.577 | 0.436 | 0.326 | **0.260** | 0.544 | 1.010 |
| PG-BCMR [14] | **0.754** | **0.591** | **0.445** | **0.332** | 0.257 | **0.550** | **1.013** |
| ATT-FCN† [26] | 0.731 | 0.565 | 0.424 | 0.316 | 0.250 | 0.535 | 0.943 |
| MSM† [25] | 0.739 | 0.575 | 0.436 | 0.330 | 0.256 | 0.542 | 0.984 |
| AdaptiveAttention† [15] | 0.746 | 0.582 | 0.443 | 0.335 | 0.264 | 0.550 | 1.037 |
| Att2in† [19] | - | - | - | 0.344 | 0.268 | 0.559 | 1.123 |
| COCO Online Testing Server C40 | | | | | | | |
| Method | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE_L | CIDEr |
| Google NIC [23] | 0.895 | 0.802 | 0.694 | 0.587 | 0.346 | 0.682 | 0.946 |
| Hard-Attention [24] | 0.881 | 0.779 | 0.658 | 0.537 | 0.322 | 0.654 | 0.893 |
| AdaptiveAttention [15] | 0.906 | 0.823 | 0.717 | 0.607 | 0.347 | 0.689 | 1.004 |
| AdaptiveAttention + CL (Ours) | **0.910** | **0.831** | **0.728** | **0.617** | **0.350** | **0.695** | **1.029** |
| PG-BCMR [14] | - | - | - | - | - | - | - |
| ATT-FCN† [26] | 0.900 | 0.815 | 0.709 | 0.599 | 0.335 | 0.682 | 0.958 |
| MSM† [25] | 0.919 | 0.842 | 0.740 | 0.632 | 0.350 | 0.700 | 1.003 |
| AdaptiveAttention† [15] | 0.918 | 0.842 | 0.740 | 0.633 | 0.359 | 0.706 | 1.051 |
| Att2in† [19] | - | - | - | - | - | - | - |

Table 2: This table lists published results of state-of-the-art image captioning models on the online COCO testing server. † indicates ensemble model. "-" indicates not reported. In this table, CL improves the base model (AdaptiveAttention [15]) to gain the best results among all single models on C40.

# Image Captioning


1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, *39*(4), 652-663.

2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., & Salakhutdinov, R., et al. (2015). Show, attend and tell: neural image caption generation with visual attention. *Computer Science*, 2048-2057.

3. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.

4. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.

5. Fortunato, M., Blundell, C., & Vinyals, O. (2017). Bayesian Recurrent Neural Networks. *arXiv preprint arXiv:1704.02798*.

6. Gu, J., Cai, J., Wang, G., & Chen, T. (2017). Stack-Captioning: Coarse-to-Fine Learning for Image Captioning. *arXiv preprint arXiv:1709.03376*.

7. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems* (pp. 2352-2360).

8. Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep Reinforcement Learning-based Image Captioning with Embedding Reward. *arXiv preprint arXiv:1704.03899*.

# Image Generation

- Generate a natural image
- Or filling blanks in images

(a) Input context

(d) Context Encoder ($L2$ + Adversarial loss)

Figure 1: Class-conditional samples generated by our model.

# Image Generation Models

- Auto-Encoder
  - Denoising / Variational
- Generative adversarial networks (GAN)
  - Train a generator and a discriminator
- Pixel-CNN/RNN
  - Generate pixel by pixel

# State-of-the-art for Image Generation

| Ch. | Param (M) | Shared | Skip-$z$ | Ortho. | FID | IS | (min FID) / IS | FID / (max IS) |
|---|---|---|---|---|---|---|---|---|
| 64 | 317.1 | ✗ | ✗ | ✗ | 48.38 | 23.27 | 48.6/23.1 | 49.1/23.9 |
| 64 | 99.4 | ✓ | ✓ | ✓ | 23.48 | 24.78 | 22.4/21.0 | 60.9/35.8 |
| 96 | 207.9 | ✓ | ✓ | ✓ | 18.84 | 27.86 | 17.1/23.3 | 51.6/38.1 |
| 128 | 355.7 | ✓ | ✓ | ✓ | 13.75 | 30.61 | 13.0/28.0 | 46.2/47.8 |

BigGAN

Figure 3: Samples (128x128) from a VQ-VAE with a PixelCNN prior trained on ImageNet images. From left to right: kit fox, gray whale, brown bear, admiral (butterfly), coral reef, alp, microwave, pickup.

# Image Generation References


- *Aäron van den Oord, Oriol Vinyals, Koray Kavukcuoglu: Neural Discrete Representation Learning. NIPS 2017: 6306-6315*
- *Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. CoRR, abs/1511.06434.*
- *Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. ArXiv, abs/1606.03498.*
- *Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. NIPS.*
- *Salimans, T., Karpathy, A., Chen, X., & Kingma, D.P. (2017). PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. ArXiv, abs/1701.05517.*

# Text Classification

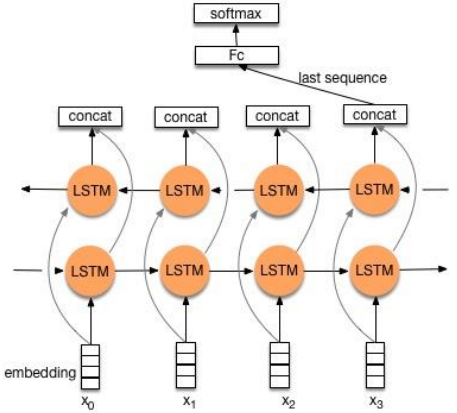
- Document-level classification
- Sentence-level classification

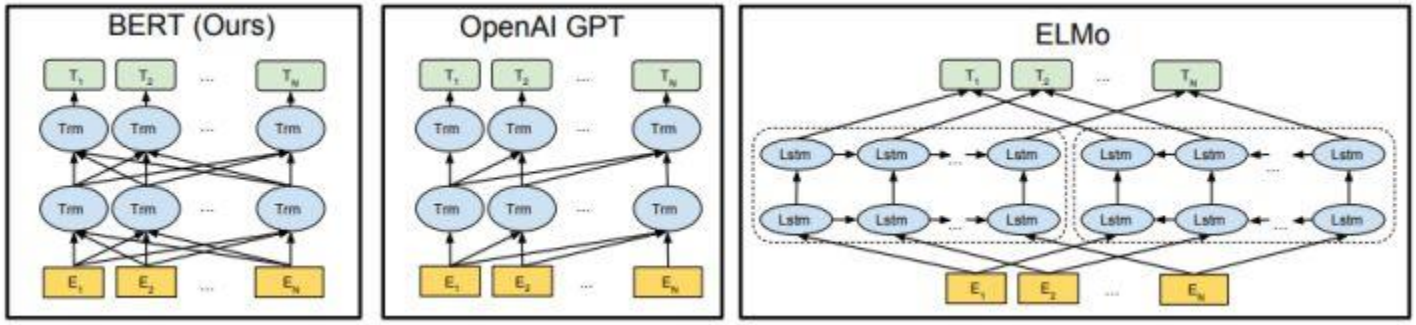
- Topic classification: sports, economics
- Sentiment classification
  - Positive vs. negative
  - Subjective vs. objective
  - Emotion
- Relation classification

- Multi-class / Fine-grained classification

# Text Classification Models

- BERT / GPT / Elmo
- Bi-directional LSTM / Stacked LSTM



# Datasets for Text Classification

- YELP / IMDB Large-scale datasets
- GLEU dataset  Multi-task language dataset
- Stanford Sentiment Treeebank (SST) dataset contains 11,855 sentences, and has splited into the training/validation/test parts, respectively containing 8,544/1,101/2,210 sentences.
  - Specially, the dataset has provided phrase-level annotation on all inner nodes, so that it is more suitable for Tree-structed Model.

# Datasets for Text Classification

| Corpus | #docs | #s/d | #w/d | $\|V\|$ | #class | Class Distribution |
|---|---|---|---|---|---|---|
| Yelp 2013 | 335,018 | 8.90 | 151.6 | 211,245 | 5 | .09/.09/.14/.33/.36 |
| Yelp 2014 | 1,125,457 | 9.22 | 156.9 | 476,191 | 5 | .10/.09/.15/.30/.36 |
| Yelp 2015 | 1,569,264 | 8.97 | 151.9 | 612,636 | 5 | .10/.09/.14/.30/.37 |
| IMDB | 348,415 | 14.02 | 325.6 | 115,831 | 10 | .07/.04/.05/.05/.08/.11/.15/.17/.12/.18 |

Table 1: Statistical information of Yelp 2013/2014/2015 and IMDB datasets. #docs is the number of documents, #s/d and #w/d represent average number of sentences and average number of words contained in per document, $|V|$ is the vocabulary size of words, #class is the number of classes.

# Datasets for Text Classification

| **Data** | $c$ | $l$ | $N$ | $\|V\|$ | $\|V_{pre}\|$ | *Test* |
|---|---|---|---|---|---|---|
| MR | 2 | 20 | 10662 | 18765 | 16448 | CV |
| SST-1 | 5 | 18 | 11855 | 17836 | 16262 | 2210 |
| SST-2 | 2 | 19 | 9613 | 16185 | 14838 | 1821 |
| Subj | 2 | 23 | 10000 | 21323 | 17913 | CV |
| TREC | 6 | 10 | 5952 | 9592 | 9125 | 500 |
| CR | 2 | 19 | 3775 | 5340 | 5046 | CV |
| MPQA | 2 | 3 | 10606 | 6246 | 6083 | CV |

Table 1: Summary statistics for the datasets after tokenization. $c$: Number of target classes. $l$: Average sentence length. $N$: Dataset size. $|V|$: Vocabulary size. $|V_{pre}|$: Number of words present in the set of pre-trained word vectors. *Test*: Test set size (CV means there was no standard train/test split and thus 10-fold CV was used).

# State-of-the-art for Text Classification

| Models | MR | SST | Subj | AG |
|---|---|---|---|---|
| LSTM | 77.4* | 46.4* | 92.2 | 90.9 |
| biLSTM | 79.7* | 49.1* | 92.8 | 91.6 |
| CNN | 81.5* | 48.0* | 93.4* | 91.6 |
| RAE | 76.2* | 47.8 | 92.8 | 90.3 |
| Tree-LSTM | 80.7* | **50.1** | 93.2 | 91.8 |
| Self-Attentive | 80.1 | 47.2 | 92.5 | 91.1 |
| ID-LSTM | 81.6 | 50.0 | 93.5 | 92.2 |
| HS-LSTM | **82.1** | 49.8 | **93.7** | **92.5** |

Tianyang Zhang, Minlie Huang, Li Zhao.
Learning Structured Representation for Text Classification via Reinforcement Learning. AAAI 2018.

| System | MNLI-(m/mm) 392k | QQP 363k | QNLI 108k | SST-2 67k | CoLA 8.5k | STS-B 5.7k | MRPC 3.5k | RTE 2.5k | Average - |
|---|---|---|---|---|---|---|---|---|---|
| Pre-OpenAI SOTA | 80.6/80.1 | 66.1 | 82.3 | 93.2 | 35.0 | 81.0 | 86.0 | 61.7 | 74.0 |
| BiLSTM+ELMo+Attn | 76.4/76.1 | 64.8 | 79.8 | 90.4 | 36.0 | 73.3 | 84.9 | 56.8 | 71.0 |
| OpenAI GPT | 82.1/81.4 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| $\text{BERT}_{\text{BASE}}$ | 84.6/83.4 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| $\text{BERT}_{\text{LARGE}}$ | **86.7/85.9** | **72.1** | **92.7** | **94.9** | **60.5** | **86.5** | **89.3** | **70.1** | **82.1** |

# Text Classification Papers


- Minlie Huang, Qiao Qian, and Xiaoyan Zhu. 2017. Encoding Syntactic Knowledge in Neural Networks for Sentiment Classification. ACM Trans. Inf. Syst. 35, 3, Article 26 (June 2017), 27 pages. DOI: https://doi.org/10.1145/3052770
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
- Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011b. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. Association for Computational Linguistics, 151–161.

# Reading Comprehension

- QA-style RC
  - natural language form questions
  - answer is a word/span from passages

---

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

# Reading Comprehension

- Cloze-style RC
  - question is marked by a blank space;
  - single-word answer from candidates is selected;

# Reading Comprehension

- Choice-style RC
  - natural language form questions
  - Few answer sentences

Story: ..... *I just wanted to take a few minutes to meet with everyone to make sure your class presentations for next week are all in order and coming along well. And as you know, you're supposed to report on some areas of recent research on genetics, something, you know, original.* .....(manual transcription)

Question: *Why does the professor meet with the student?*

Choices:

**A. To determine if the student has selected an appropriate topic for his class project**

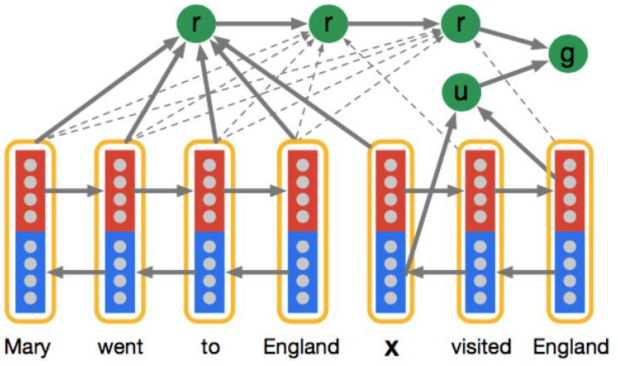
B. To find out if the student is interested in taking part in a genetics project

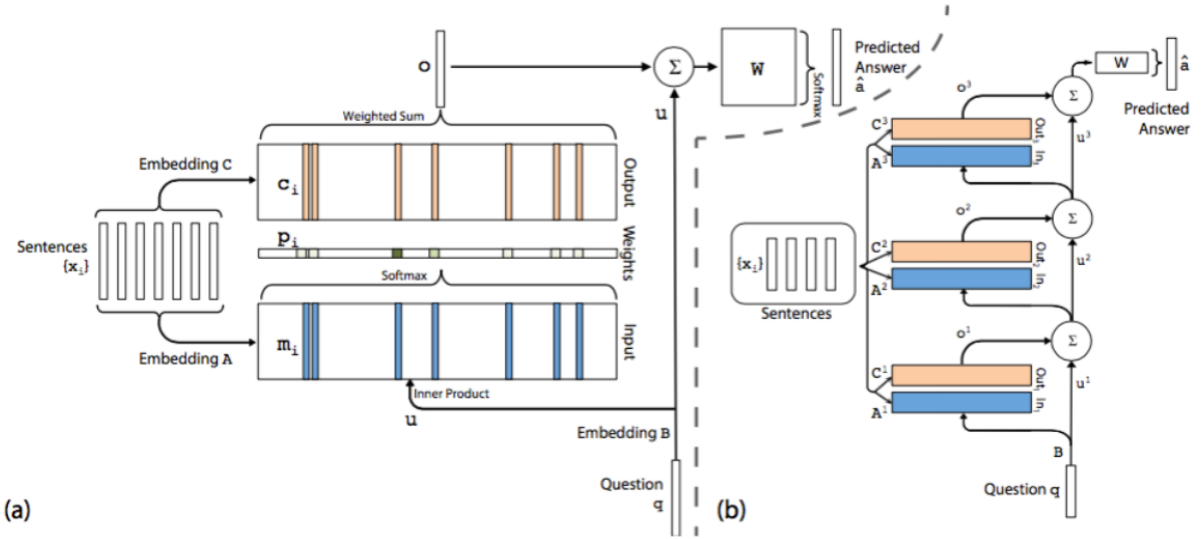
C. To discuss the student's experiment on taste perception

D. To explain what the student should focus on for his class presentation

# Reading Comprehension Models

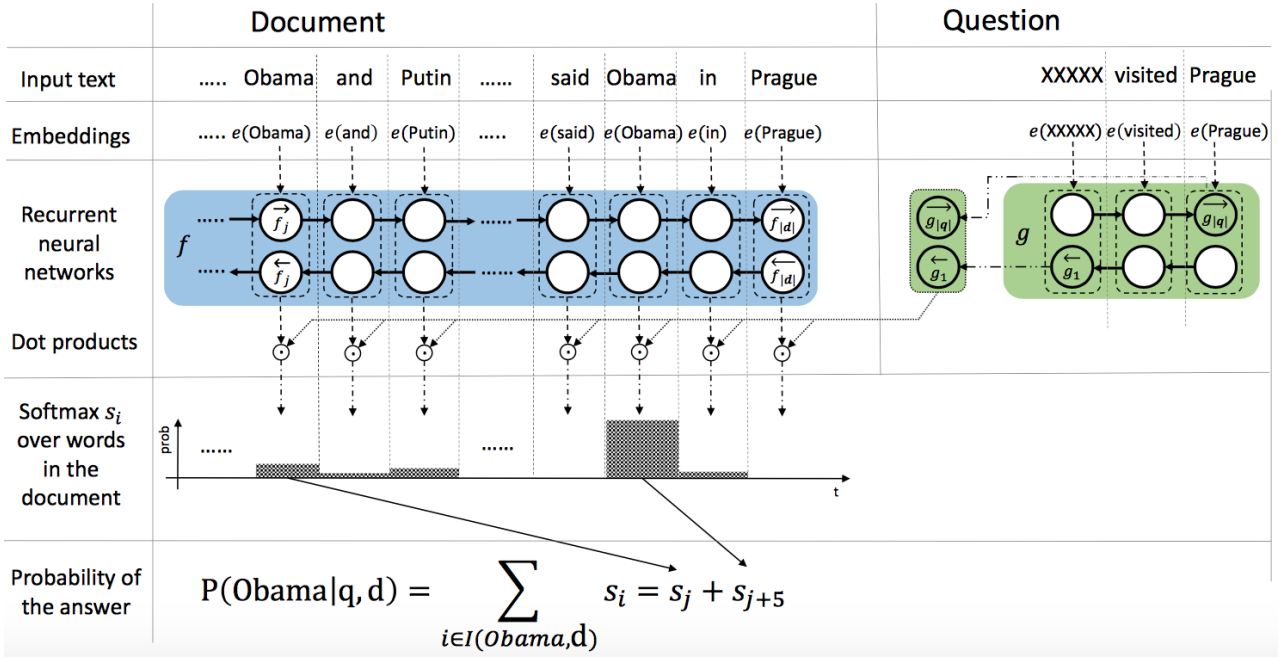
- BERT
- Attention
- Multi-hops
- Memory Network



# Reading Comprehension Models

- Attention
- Multi-hops
- Memory Network



# Datasets for Reading Comprehension

- For QA style Reading Comprehension, we recommend SQuAD dataset
  - includes 107785 question-answer pairs on 536 Wikipedia articles
  - answer is a word/span in passage

- For Cloze style Reading Comprehension, we recommend CNN/ DailyMail dataset
  - includes 93k articles for CNN and 220k articles for DailyMail
  - 1M anonymized named-entity bullet points

# Datasets for Reading Comprehension

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

| **Original Version** | **Anonymised Version** |
|---|---|
| **Context** | |
| The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ... | the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ... |
| **Query** | |
| Producer **X** will not press charges against Jeremy Clarkson, his lawyer says. | producer **X** will not press charges against *ent212* , his lawyer says . |
| **Answer** | |
| Oisin Tymon | *ent193* |

# State-of-the-art for Reading Comprehension

| Rank | Model | EM | F1 |
|---|---|---|---|
| | Human Performance<br>*Stanford University*<br>(Rajpurkar & Jia et al. '18) | 86.831 | 89.452 |
| 1<br>Jul 22, 2019 | XLNet + DAAF + Verifier (ensemble)<br>*PINGAN Omni-Sinitic* | **88.592** | **90.859** |
| 2<br>Jul 26, 2019 | UPM (ensemble)<br>*Anonymous* | 88.231 | 90.713 |
| 3<br>Aug 04, 2019 | XLNet + SG-Net Verifier (ensemble)<br>*Shanghai Jiao Tong University & CloudWalk*<br>https://arxiv.org/abs/1908.05147 | 88.174 | 90.702 |
| 4<br>Aug 04, 2019 | XLNet + SG-Net Verifier++ (single model)<br>*Shanghai Jiao Tong University & CloudWalk*<br>https://arxiv.org/abs/1908.05147 | 87.238 | 90.071 |

| Model | CNN | | Daily Mail | |
|---|---|---|---|---|
| | Valid | Test | Valid | Test |
| Human(query) | - | - | - | - |
| Human(context+query) | - | - | - | - |
| LSTM(context+query) | - | - | - | - |
| Deep LSTM Reader | 55.0 | 57.0 | 63.3 | 62.2 |
| Attentive Reader | 61.6 | 63.0 | 70.5 | 69.0 |
| Impatient Reader | 61.8 | 63.8 | 69.0 | 68.0 |
| Memory Networks(single) | 63.4 | 66.8 | - | - |
| Memory Networks(ensemble) | 66.2 | 69.4 | - | - |
| Attention Sum Reader(single) | 68.6 | 69.5 | 75.0 | 73.9 |
| Attention Sum Reader(ensemble) | 73.9 | 75.4 | 78.7 | 77.7 |
| Dynamic Entity Representation | 71.3 | 72.9 | - | - |
| Gate Attention Reader(single) | 73.0 | 73.8 | 76.7 | 75.7 |
| Gate Attention Reader(ensemble) | 76.4 | **77.4** | 79.1 | **78.1** |
| Iterative Alternating Attention(single) | 72.6 | 73.3 | - | - |
| Iterative Alternating Attention(ensemble) | 73.6 | 74.0 | - | - |

# Reading Comprehension References


- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman,and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Proc. of NIPS, pages 1684–1692.
- Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, andJan Kleindienst. 2016. Text understanding with the attention sum reader network. ACL
- Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al.2015. End-to-end memory networks. In Proc. of NIPS,pages 2431–2439.
- Cui Y, Chen Z, Wei S, et al. Attention-over-Attention Neural Networks for Reading Comprehension[J]. meeting of the association for computational linguistics, 2017: 593-602.
- Shen Y, Huang P, Gao J, et al. ReasoNet: Learning to Stop Reading in Machine Comprehension[J]. SIGKDD 2017.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In Proceedings of ICLR
- Microsoft Research Asia. 2017. R-NET: MACHINE READING COMPREHENSION WITH SELF-MATCHING NETWORKS. In Proceedings of ACL

# Text Generation

- Language Models
  - Generating grammatical language.
- Translation Models
  - Translating a sequence of words from Language A to Language B.
- Conversation Models
  - Generating relative and appropriate responses given user’s query.

# Text Generation Tasks (Examples)

- Text generation tasks with different controlled attributes

| Post | Model | | Response(Original) |
|---|---|---|---|
| 比情人节更庸俗的事就是，对情人节的接待规格待遇提出要求。<br>What is more vulgar than Valentine's day is to make specific demands on the reception of the Valentine's day. | Seq2Seq | | 这才是真正的情人节。 |
| | ECM | *Like* | 情人节快乐！ |
| | | *Happiness* | 哈哈，这也太浪漫了吧！ |
| | | *Sadness* | 我也想要这样的情人节。 |
| | | *Disgust* | 这就是所谓的情人节！ |
| | | *Anger* | 这是赤裸裸的炫耀啊！ |
| 知道真相的我眼泪笑下来。<br>Knowing the truth, I laughed myself to tears. | Seq2Seq | | 知道真相的我眼泪掉下来。 |
| | ECM | *Like* | 太可爱了，太可爱了！ |
| | | *Happiness* | 笑得我眼泪都出来了。 |
| | | *Sadness* | 知道真相的我眼泪掉下来。 |
| | | *Disgust* | 这是真的还是假的？ |
| | | *Anger* | 什么意思？没看懂。 |

**Sentiment** (negative, positive)

+

**Tense** (past, present, future)

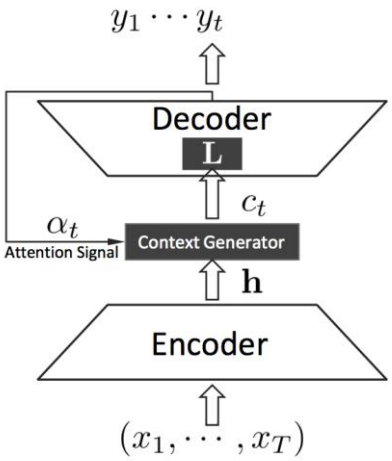
**Emotion**

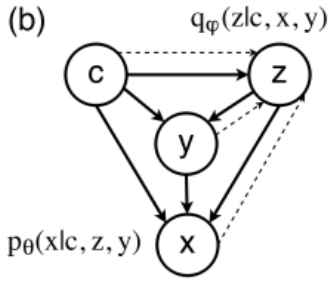
(like, happiness, sadness, disgust, anger)

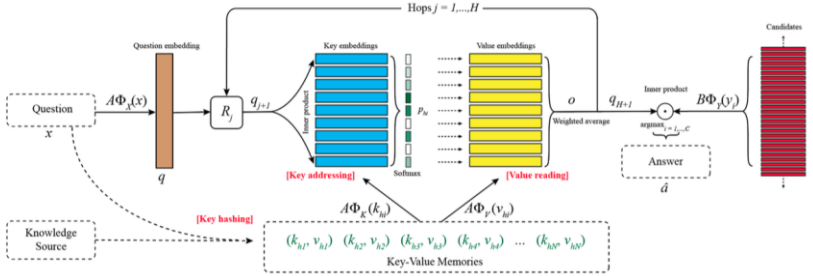
| **Varying the unstructured code $z$** | |
|---|---|
| *("negative", "past")* | *("positive", "past")* |
| the acting was also kind of hit or miss . | his acting was impeccable |
| i wish i 'd never seen it | this was spectacular , i saw it in theaters twice |
| by the end i was so lost i just did n't care anymore | it was a lot of fun |
| *("negative", "present")* | *("positive", "present")* |
| the movie is very close to the show in plot and characters | this is one of the better dance films |
| the era seems impossibly distant | i 've always been a big fan of the smart dialogue . |
| i think by the end of the film , it has confused itself | i recommend you go see this, especially if you hurt |
| *("negative", "future")* | *("positive", "future")* |
| i wo n't watch the movie | i hope he 'll make more movies in the future |
| and that would be devastating ! | i will definitely be buying this on dvd |
| i wo n't get into the story because there really is n't one | you will be thinking about it afterwards, i promise you |

# Text Generation Models

- Language Models
  - RNN / Transformer
  - Autoencoder or VAE
  - GAN
- Translation Models
  - Seq2Seq
- Conversation Models
  - Seq2Seq
  - Memory Networks
  - VAE
- Pretrained Models
  - GPT2



# Datasets for Text Generation

- Google 1 Billion Word Corpus dataset makes available a standard corpus of reasonable size (0.8 billion words) to train and evaluate language models.
- Workshop on Statistical Machine Translation (WMT) dataset contains training data for five language pairs, and a common framework (including a baseline system). The task is to improve methods current methods.
- Ubuntu Dialogue Corpus, a dataset containing almost 1 million multi-turn dialogues, with a total of over 7 million utterances and 100 million words.
- OpenSubtitles Dialogs Corpus contains a large collection of conversations extracted from raw movie scripts

# Datasets for Dialogue Generation (by CoAI)

- **Emotional Conversation Dataset** contains 1,110,000 post-response pairs collected from Weibo. Each post/response is tagged with an emotion category (happy, sad, like, disgust, angry, others) by an emotion classifier.
- **Commonsense Conversation Dataset** contains one-turn post-response pairs with the corresponding commonsense knowledge graphs. Each pair is associated with some knowledge graphs retrieved from ConceptNet.
- **Dialogue Question Generation Dataset** contains post-response pairs collected from Weibo. Each response is a question, detected with manually-crafted templates.
- **Chinese Dialogue Dataset with Sentence Function** contains post-response pairs annotated with sentence function labels.

http://coai.cs.tsinghua.edu.cn/hml/dataset/

# CoTK (recommended)

- A tools supporting NLP, especially language generation tasks
- Contact TAs / post issues at Github if you meet any problems

- Report bugs / suggestions (1 ~ 2 points bonus)
- Make contributions to CoTK codes (2 ~ 5 points bonus)
- Write a Baseline for CoTK (3 ~ 5 points bonus)

# State-of-the-art for Text Generation

## WMT 2014 EN-DE

Models are evaluated on the English-German dataset of the Ninth Workshop on Statistical Machine Translation (WMT 2014) based on BLEU.

| Model | BLEU | Paper / Source |
|---|---|---|
| Transformer Big + BT (Edunov et al., 2018) | 35.0 | Understanding Back-Translation at Scale |
| DeepL | 33.3 | DeepL Press release |
| DynamicConv (Wu et al., 2019) | 29.7 | Pay Less Attention With Lightweight and Dynamic Convolutions |
| Transformer Big (Ott et al., 2018) | 29.3 | Scaling Neural Machine Translation |

# Text Generation references


- Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (2014), pp. 3104–3112.
- Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (2015).
- Chung, J., Cho, K., and Bengio, Y. A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147 (2016).
- Dong, D., Wu, H., He, W., Yu, D., and Wang, H. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (2015), pp. 1723–1732.
- Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In Conference on Empirical Methods in Natural Language Processing (2015).
- Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. Coverage-based neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016).
- Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In ACL, pages 1577–1586, 2015.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. InEMNLP,pages 1192–1202, 2016.
- Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. In ACL, pages 1631–1640,2016.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, Bing Liu. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI 2018.
- Hao Zhou, Tom Yang, Minlie Huang, Haizhou Zhao, Jingfang Xu, Xiaoyan Zhu. Commonsense Knowledge Aware Conversation Generation with Graph Attention. IJCAI-ECAI 2018.
- Pei Ke, Jian Guan, Minlie Huang, Xiaoyan Zhu. Generating Informative Responses with Controlled Sentence Function. ACL 2018.
- Yansen Wang, Chenyi Liu, Minlie Huang, Liqiang Nie. Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. ACL 2018.

# Other Options

- Actually neural network can do more than above, for example
- Generative adversial network: "What I cannot create, I do not understand."
- Deep reinforcement learning: Deepmind's Apollo project for general AI
- ……

# Reproduction Challenge

- Reproduction for language generation tasks
- You must write the model with CoTK (or adapt the codes)
- You will be also awarded with bonus if you make contributions to CoTK
  - Help us add new datasets / metrics
- Important:
  - You can use the codes accessible online. (But only the adaption will be counted as your work, so you should do detailed tests or reproduce more models.)
  - Format your reports and codes well (at least like this)
  - Do ablation tests
  - If possible, try your novel ideas
  - Make sure you spend enough efforts
  - Negative results is also acceptable

# Reproduction Challenge

- We recommend the following papers / codes
  - Texygen: A Benchmarking Platform for Text Generation Models
  - RelGAN
  - GPT 2.0 + finetuning
  - Transformer: Language Model / seq2seq / Machine Translation
  - Seq2seq-rl
  - Seq2seq-exposure-bias
  - Text style transfer
  - Image Caption
  - Text Summarization
- One line isn't strictly one project
- If you want to reproduce the other papers, you have to contact TAs before going on.

# Important Dates

- **Project Proposal**: 第8周，11月3日，星期日，网络学堂提交
- **Milestone Report**: 第11周，11月24日，星期日，网络学堂提交
- **Project Presentation (报名参加)**: 第16周，12月29日，星期日
- **Final Report**: 第17周，2019年1月5日，星期日

- **请用中文写所有的报告，推荐使用Latex;助教会提供模板**

# How to Choose a Good Project?

- **Significance and importance**: why do I choose this topic?
- **Novelty**: anything new?
- **Substances**: how much work I did?
- **Results**: How well I did?
  - Lessons, experiences learned?
  - New, interesting discoveries?
  - Inspiration to others? Any vision?

# Remember 5 Key Steps

- 1. Idea
- 2. Math
- 3. Pseudocodes →Codes
- 4. Experiments
- 5. Report/Paper

# Some Notes

- It is hard to overperform baselines in some tasks
  - But the results of baselines are always needed
  - Novel ideas are more important
- Don’t waste too much time on setting up environment
  - Use the existing tools
  - Only data labeling can’t be your main part of work
- If someone in your group didn’t do anything, you can tell TAs privately. Every one will be scored separately.

# Submission Guideline~~Proposal

- Your project **Proposal** should be at most 2 pages using the provided template. The following is a suggested structure for your proposal:
- **Title, Author**(s)
- **Task and Problem Definition**: Define what you will do precisely. Is it a well-define problem? Try to use math symbols to formulate the task.
- **Dataset**: Existing dataset or create a new one?
- **Challenges and Baselines**: What is difficult for you? State-of-the-art?
- **Proposal**: State your idea very clearly. Connect the idea with the math and even the codes
- **Feasibility**: State why this is possible for your team

# Submission Guideline~~Milestone report

- Your project **milestone report** should be between 2 - 4 pages using the provided template. The following is a suggested structure for your report:
- **Title, Author**(s)
- **Introduction**: this section introduces your problem, and the overall plan for approaching your problem
- **Problem statement**: Describe your problem precisely specifying the dataset to be used, expected results and evaluation
- **Technical Approach**: Describe the methods you intend to apply to solve the given problem
- **Intermediate/Preliminary Results**: State and evaluate your results up to the milestone

# Submission Guideline~~Final Report

- Your **final write-up** should be between 6 - 8 pages using the provided template. The following is a suggested structure for the report:
- **Title, Author**(s)
- **Abstract**: It should not be more than 300 words
- **Introduction**: this section introduces your problem, and the overall plan for approaching your problem
- **Background/Related Work**: This section discusses relevant literature for your project
- **Approach**: This section details the framework of your project. Be specific, which means you might want to include equations, figures, plots, etc
- **Experiment**: This section begins with what kind of experiments you're doing, what kind of dataset(s) you're using, and what is the way you measure or evaluate your results. It then shows both quantitative evaluations (show numbers, figures, tables, etc) as well as qualitative results (show images, example results, etc).
- **Conclusion**: What have you learned? Suggest future ideas.
- **References**: This is absolutely necessary

- Describe teamwork in another files

# Grading

- 40 in total
- Milestone report (5)
- Project presentation (Bonus at most 5)
- Final submission (35)
  - Codes
  - Report
    - Clarity, structure, language
    - Background literature survey, good understanding of the problem
    - Innovation, correctness
    - Sound evaluation metrics, good results and performance
    - Good insights and discussion of methods, analysis, results, etc.

# Q&A

Tracing the most recent papers on your topic!