A Gripping Story - Written by Deep Neural Network

Natural Language Processing

July 2, 2020

Team Members
Deepan Chakravarthi Padmanabhan
Venkata Santosh Sai Ramireddy Muthireddy
Vahid MohammadiGahrooei

Introduction

Neural-storyteller

- Input: meaningful image. Output: story
- Automatic image indexing, entertainment, vision-language assistance [1]

Generated story about image
Model: Romantic Novels

"He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder.

He wanted to strangle me, considering the beatiful boy I'd become wearing his boxers."

[1] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CSUR) 51, no. 6, pp. 1-36. 2019.

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Overview of Neural-storyteller

Pipeline

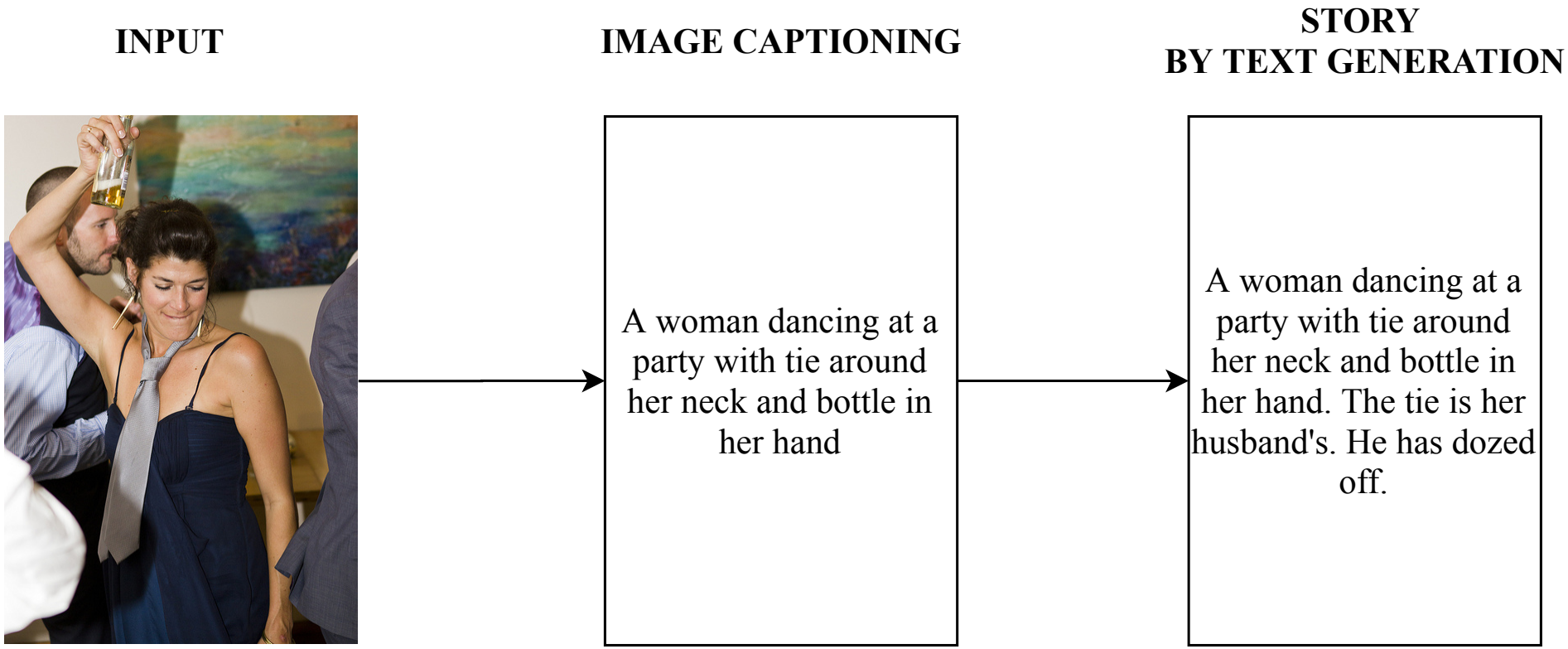


Figure 1: Two-stages of storyteller

Photo credits: [3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

Image captioning: overview

- Encoder: vision embeddings & Decoder: Recurrent Neural Network (RNN)
- Attention: Bahdanau attention

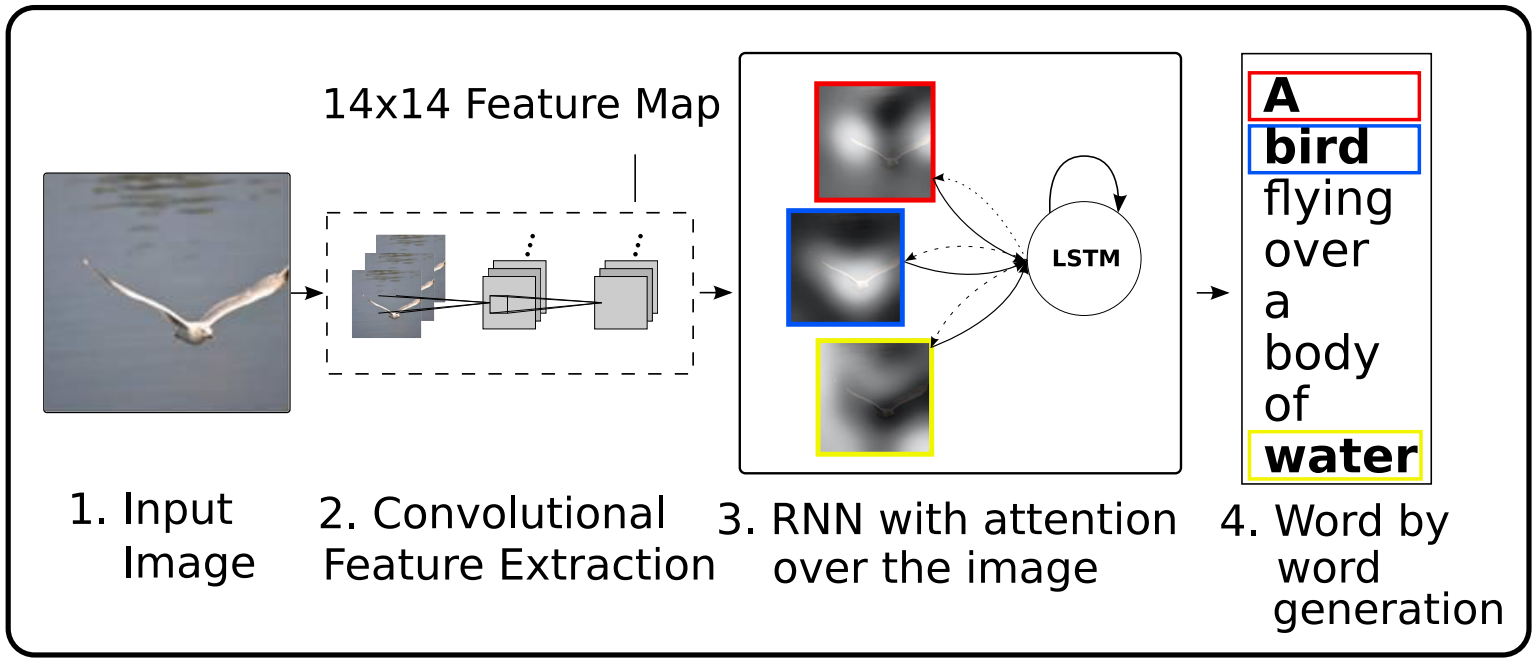


Figure 2: Model learns word-image alignment. Image from [4]

[4] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.

Image captioning: overview

- Encoder: vision embeddings & Decoder: Recurrent Neural Network (RNN)
- Attention: Bahdanau attention

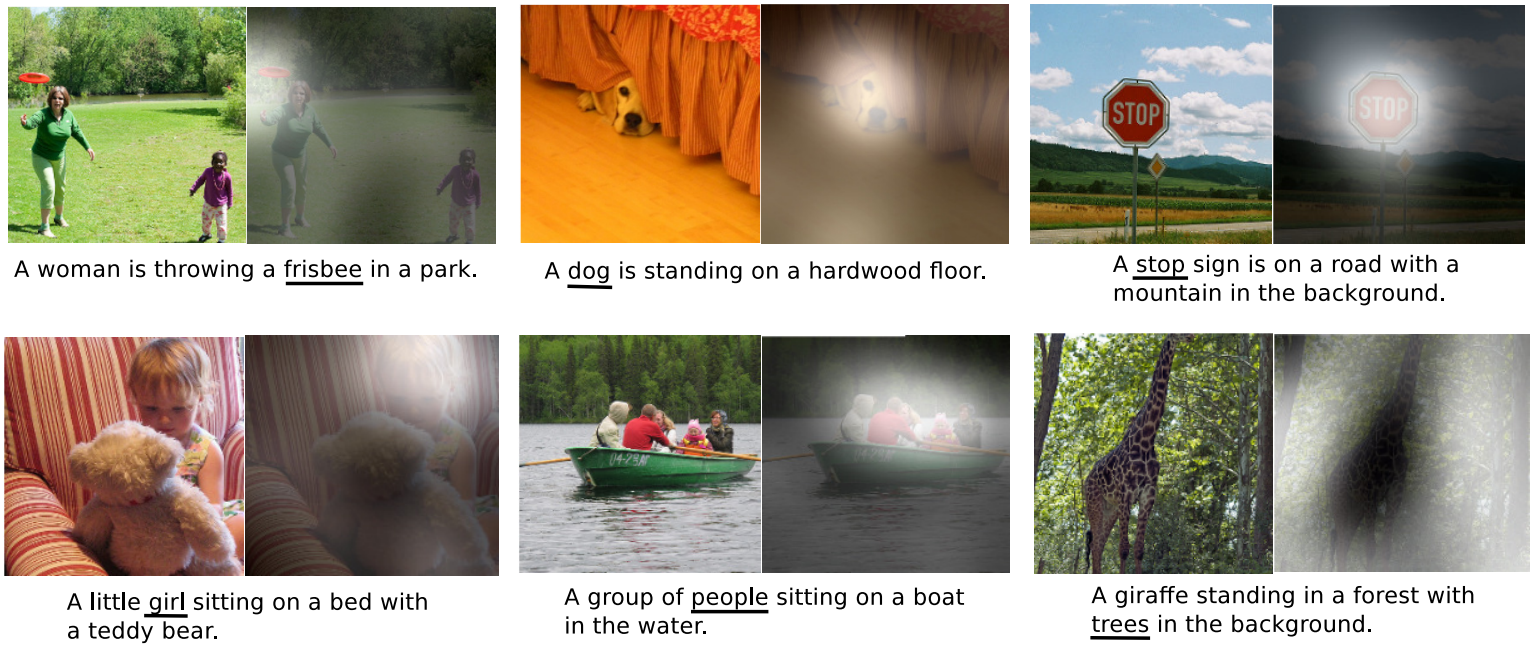


Figure 3: Attention to underlined words. Image from [4]

[4] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.

Image captioning: experimental setup

- Dataset: Microsoft COCO [3]
  - (train, validation, test) - (10000, 3000, 1500)

[3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

Image captioning: experimental setup

- Dataset: Microsoft COCO [3]
  - (train, validation, test) - (10000, 3000, 1500)
- Training procedure:
  - 5000 vocabulary size

[3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

Image captioning: experimental setup

- Dataset: Microsoft COCO [3]
  - (train, validation, test) - (10000, 3000, 1500)
- Training procedure:
  - 5000 vocabulary size
  - teacher forcing

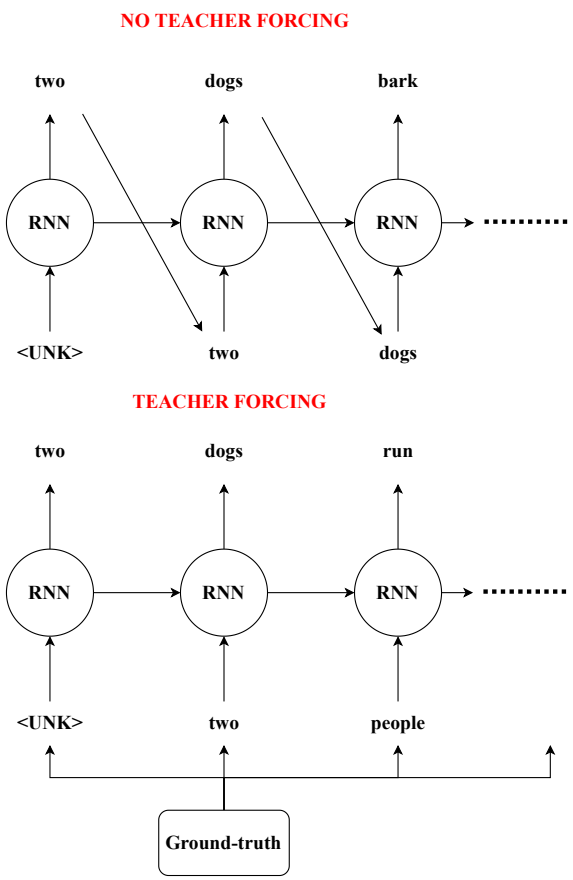


Figure 4: Illustration of training with teacher forcing. <UNK> is any word in training sequence

[3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

Image captioning: experimental setup

- Dataset: Microsoft COCO [3]
  - (train, validation, test) - (10000, 3000, 1500)
- Training procedure:
  - 5000 vocabulary size
  - teacher forcing
  - early stopping

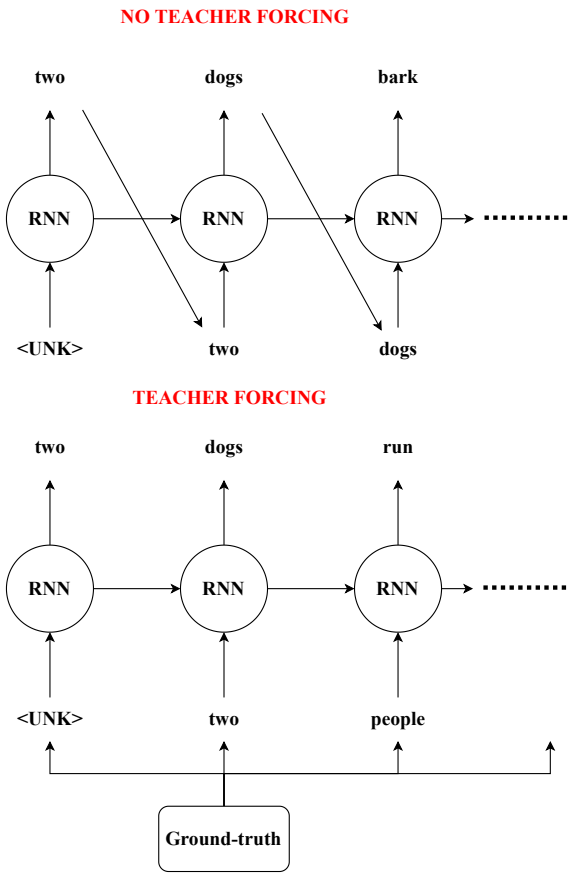


Figure 4: Illustration of training with teacher forcing. <UNK> is any word in training sequence

[3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

Image captioning: experimental setup

- Dataset: Microsoft COCO [3]
  - (train, validation, test) - (10000, 3000, 1500)
- Training procedure:
  - 5000 vocabulary size
  - teacher forcing
  - early stopping
  - TensorFlow embedding layer
  - sparse categorical cross entropy

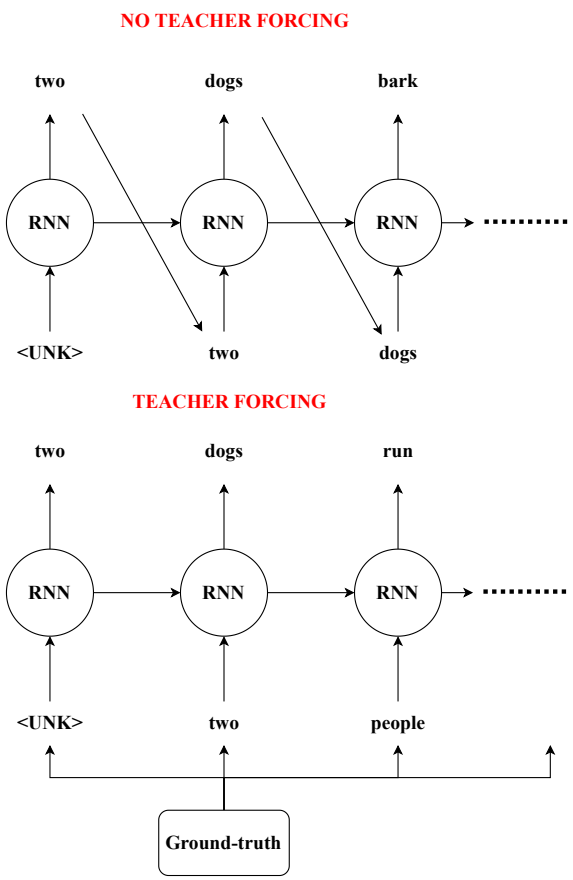
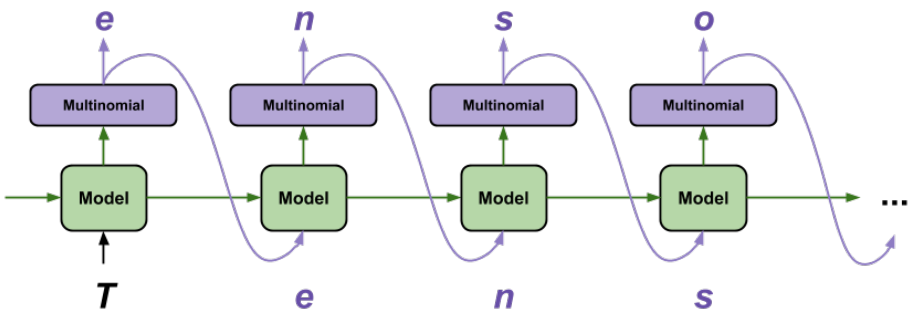


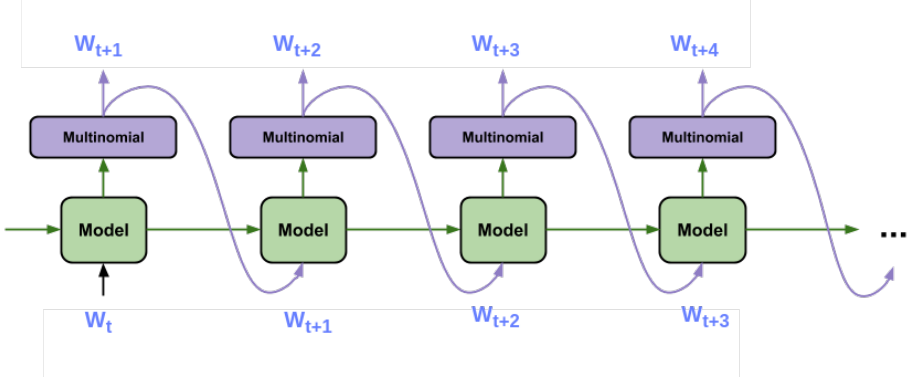
Figure 4: Illustration of training with teacher forcing. <UNK> is any word in training sequence

[3] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

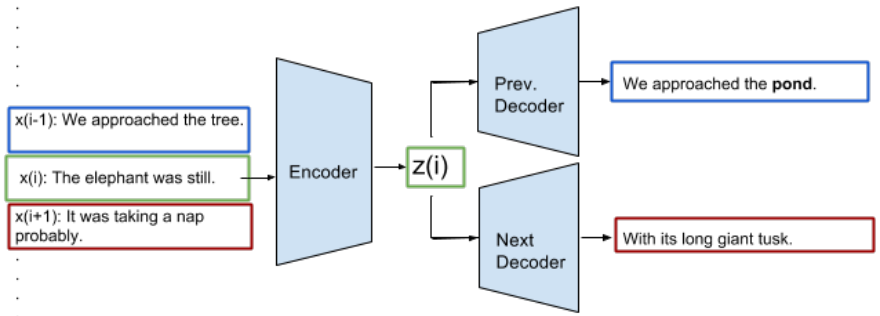
Story generation: overview



(a) Character embedding [5]



(b) Word embedding [5]



(c) Sentence embedding (skip-thought vectors) [6]

Figure 5: Different approaches of story generation from captions

Photo credits: [5] Tensorflow, "Text generation with an RNN." Accessed on: 20-06-20. [Online] url: https://www.tensorflow.org/tutorials/text/text_generation

Photo credits: [6] Sanyam Agarwal , "My thoughts on Skip-Thoughts." Accessed on: 20-06-20. [Online] url: https://medium.com/@sanyamagarwal/my-thoughts-on-skip-thoughts-a3e773605efa

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

- Training Procedure
  - Preprocess dataset

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

- Training Procedure
  - Preprocess dataset
  - Build tokenizer

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

- Training Procedure
  - Preprocess dataset
  - Build tokenizer
  - Convert text to sequences

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

- Training Procedure
  - Preprocess dataset
  - Build tokenizer
  - Convert text to sequences
  - Train

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Story generation: experimental setup

- Dataset: only part of dataset is used

Table 1: Statistical summary of Book Corpus dataset [2]

Books	Sentences	Words	Unique words	Mean words per sentence
11,038	74,004,228	984,846,357	1,316,420	13

- Training Procedure
  - Preprocess dataset
  - Build tokenizer
  - Convert text to sequences
  - Train
  - Validate

Photo credits: [2] Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In Proceedings of the IEEE international conference on computer vision, pp. 19-27. 2015.

Results

Image captioning: choice of optimizer & feature extractor

- Bilingual Evaluation Understudy (BLEU) score
- Feature extractor from Tensorflow

Table 2: Inception-ResNet with Adam optimizer selected to feed story generator

Feature extractor	**Optimizer lr = 0.001**	**Test set - BLEU score**			
		1	**2**	**3**	**4**
Inception	Adam	**47.3**	**23.8**	**10.7**	**7.12**
Inception	SGD	23.1	11.9	4.81	1.23
Inception	RMSprop	17.8	9.23	2.29	0.78
Inception-ResNet	Adam	**51.2**	**31.1**	**20.4**	**11.3**
Inception-ResNet	SGD	28.3	9.12	3.17	2.55
Inception-ResNet	RMSprop	12.2	8.12	1.97	0.39

Image captioning: effect of batch size

- No early stopping and until train error converges
- Smaller batch provide faster convergence
- Overfitting for larger batch size

Table 3: Batch size (BS) tuning. T- Train, V- Validation, E- Error

Feature extractor	**BS = 128**			**BS = 16**		
	VE	**TE**	**Epoch**	**VE**	**TE**	**Epoch**
Inception	6.32	0.47	24	0.78	0.32	10
Inception-ResNet	4.74	0.51	27	0.54	0.43	12

Story generation: effect of embedding size and dataset

- Reduced batch size for embedding size of 1200 due to memory allocation problem

Table 4: Impact of skip-thought embedding size on training time and performance using 20000 sentences. 70%-30% train-val split

Embedding size	Epochs	Batch size	Train loss	Test loss	Time per epoch (seconds)
120	50	128	13.77	13.41	67
240	50	128	14.32	13.66	90
480	50	128	14.61	14.2	170
1200 (as per [7])	50	32	-	-	91350

Table 5: Impact of dataset size on training time and performance of word-level model. 70%-30% train-val split

Sentences	Epochs	Loss	Accuracy	Time per epoch (seconds)
30000	100	4.71	19.1	800
60000	100	5.1	17.3	1800

[7] Kiros, Ryan, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Skip-thought vectors." In Advances in neural information processing systems, pp. 3294-3302. 2015.

Results: image captioning

(a) **Predicted**: an airplane sitting on the parked next to a runway <end>
Real:<start> an airplane parked on asphalt surrounded by grass <end>

(b) **Predicted**: several colorful bananas <end>
Real: <start> a plate full of bananas sitting on a table <end>

Figure 6: Implementation results on test set

Results: image captioning

(a) **Predicted**: a large woman standing next to the throw each other guitar smiling video playing will playing other other game him <end>
Real: <start>onlookers watch as two men play a video game<end>

(b) **Predicted**: two man in the water in image of their catch a wave in the water <end>
Real: <start> a man catching a frisbee in the water <end>

Figure 7: Implementation results on test set

Results: story generation

Table 6: Story for Figure 7(a) caption prediction

Approach	Story
Character	**an airplane sitting on the parked next to a runway**t ln b baneyee oodoiitt aeenwdsd ae eww e r ewtept pymteetanbna aateee ny sryuanwelr dlast
Word	**an airplane sitting on the parked next to a runway** and i could see the mutans were coming from the water i was a little feeling of the ground above the door and the mutans were still still in the ground and the mutans were still down the floor and i could see the mutans to get the other side of the water and the rest of the mutans were still still in the ground with a matter of the water and the other foot of the ancients had been a larger chance to the union soldiers and i could see the mutans were still still in the ground
Sentence	**an airplane sitting on the parked next to a runway** spoke armed gps crone able amounted refused database should soderburg cast bridle temporarily explaining diving bum widen cancer dishes inspected unfortunate gets folder animal argued neither heavenly melon stripping stood beloved singer penance team hopes tourist petey outcome run duffel dry advantage dim solid saving platters hospital uniform appearing fingernail harm lungs usually torch photo options answering easy interrupted concussion posed assassination starlings popular mecca their wished gallo perform marker straw sandwich text seated tweed

Subjective evaluation

- Readability: quality of being easy or enjoyable to read
- Semantic relatedness: relationship between two or more words based on their meaning
- Range of each metric: 0 (low) - 10 (high)

Table 7: Evaluation taken from 10 humans for 10 stories

Approach	**Mean readability**	**Mean semantic relatedness**
Character	0	0
Word	7.4	6.1
Sentence	5.9	4.3

Conclusion

Contributions

- Implemented a storyteller in TensorFlow 2.2*
- Compared the image captioning implementation with two feature extractors
- Implemented character-level, word-level, and sentence-level text generation to form story
- Performed subjective evaluation of storyteller implemented

* GitHub repository: https://github.com/DeepanChakravarthiPadmanabhan/Image_Storyteller

Challenges

- Compatible dataset for captioning and story generation
- Improving generalization by hyperparameter tuning in limited time frame
- Character-level model training requires more epochs
- Efficient dataloader for large datasets

Lessons learned

- Implement - **Debug and Test** - Conduct experiment
- Intermediate checkpoints to overcome cluster time out
- Novel loss functions for significant increase in performance and reduce manual labor
- Callback functions to monitor individual losses for generalization
- Preparing raw text dataset for different level embedding like character, word and sentences

Future work

- End-to-end learning with multi-loss function
- Compare different attention mechanisms and embeddings in image captioning

Takeaways

- Tune hyperparameters: find sweet spot in batch size, embedding size, tokenizer word count
- Trade-off between embedding size and performance
- **Inception-ResNet** vs **Inception**
- **Adam** vs **SGD** vs **RMSProp**
- Intermediate checkpoints for training

References I

[1] MD Zakir Hossain et al. "A comprehensive survey of deep learning for image captioning". In: ACM Computing Surveys (CSUR) 51.6 (2019), pp. 1–36.

[2] Yukun Zhu et al. "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books". In: arXiv preprint arXiv:1506.06724 (2015).

[3] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: European conference on computer vision. Springer. 2014, pp. 740–755.

[4] Kelvin Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. 2015. arXiv: 1502.03044 [cs.LG].

[5] Tensorflow. Tensorflow. Accessed on: 20-06-2020. [Online]. 2018. URL: https://www.tensorflow.org/tutorials/text/text_generation.

[6] Sanyam Agarwal. My thoughts in Skip-thoughts. Accessed on: 20-06-2020. [Online]. 2018. URL: https://medium.com/@sanyamagarwal/my-thoughts-on-skip-thoughts-a3e773605efa.

[7] Ryan Kiros et al. "Skip-Thought Vectors". In: arXiv preprint arXiv:1506.06726 (2015).

[8] Gabriel Loye. Attention mechanisms. Accessed on: 20-06-2020. [Online]. 2019. URL: https://blog.floydhub.com/attention-mechanism.

Thank you for your time!

Extras

Bahdanau attention

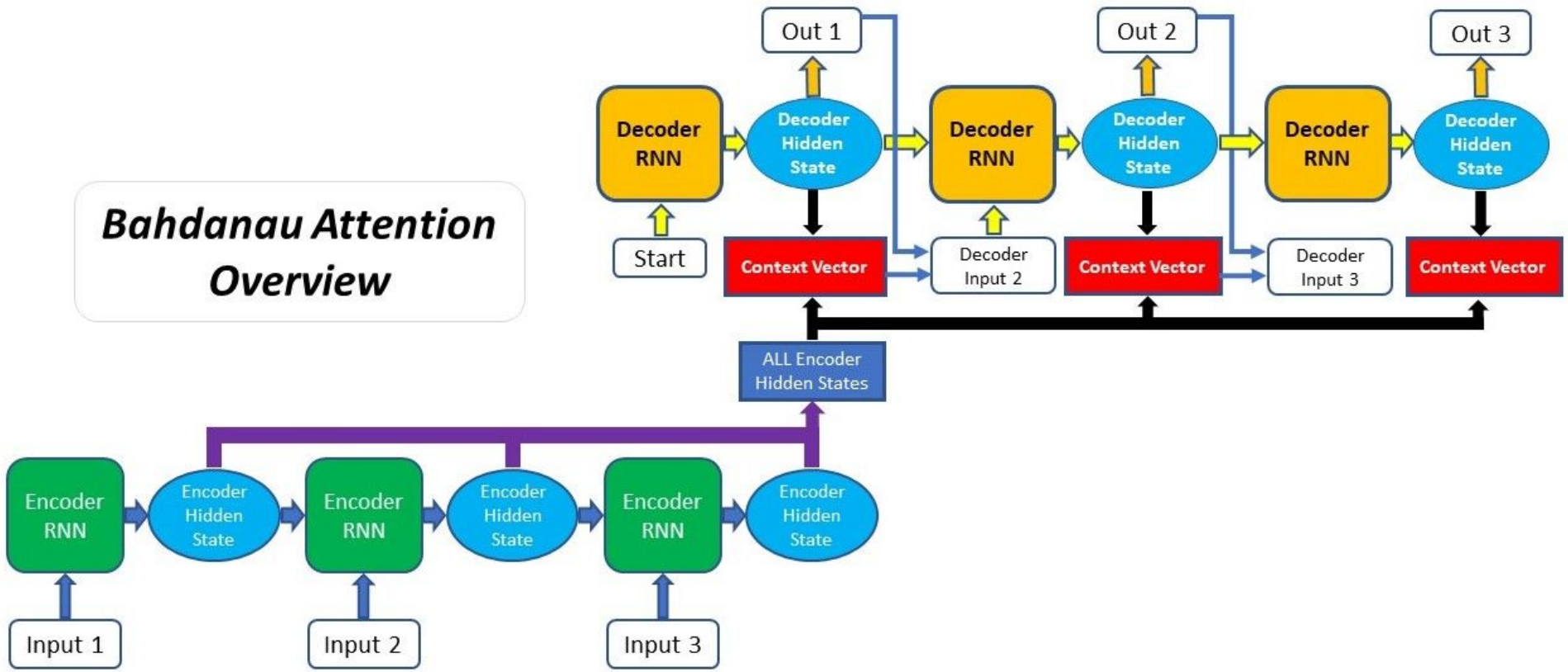


Figure 8: Context vector is used for alignment

Photo credits: [8] Gabriel Loye. FloydHub. "Attention mechanisms". Accessed on: 20-06-20. [Online]. 2019. url: https://blog.floydhub.com/attention-mechanism.