Supplementary Material: Closed-Book Training to Improve Summarization Encoder Memory

Yichen Jiang and **Mohit Bansal**
UNC Chapel Hill
{yichenj, mbansal}@cs.unc.edu

1 Coverage Mechanism

See et al. (2017) apply coverage mechanism to the pointer-generator in order to alleviate repetition. They maintain a coverage vector c^t as the sum of attention distribution over all previous decoding steps $1:t-1$. This vector is incorporated in calculating the attention distribution at current step t:

$$c^t = \sum_{t'=0}^{t-1} a^{t'}$$
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + W_c c_i^t + b_{attn}) \quad (1)$$

where W_h, W_s, W_c, b_{attn} are learnable parameters. They define the coverage loss and combine it with the primary loss to form a new loss function, which is used to fine-tune a converged pointer-generator model.

$$\text{loss}_{cov}^t = \sum_i \min(a_i^t, c_i^t)$$
$$\mathcal{L}_{total} = \frac{1}{T}\sum_{t=1}^{T}(-\log P_{pg}^t(w|x_{1:t}) + \boldsymbol{\lambda}\text{loss}_{cov}^t) \quad (2)$$

2 Reinforcement Learning

To overcome the exposure bias (Bengio et al., 2015) between training and testing, previous works (Ranzato et al., 2016; Paulus et al., 2018) use reinforcement learning algorithms to directly optimize on metric scores for summarization models. In this setting, the generation of discrete words in a sentence is a sequence of *actions*. The decision to take what *action* is based on a *Policy Network* π_θ, which outputs a distribution of all possible *actions* at that step. In our case, π_θ is simply our summarization model.

The process of generating a summary $\mathbf{s}$ given the source passage $\mathbf{P}$ can be summarized as follows. At each time step t, we sample a discrete *action* $w_t \in \mathcal{V}$ - word in vocabulary, based on distribution from *policy* $\pi_\theta(\mathbf{P}, \mathbf{s}_t)$, where $\mathbf{s}_t = w_{1:t-1}$ is the sequence of *actions* sampled in previous steps. When we reach the end of the sequence at terminal step T (end-of-sentence marker is sampled from π_θ), we feed the entire sequence $\mathbf{s}_T = w_{1:T}$ into a reward function and get a reward $R(w_{1:T}|\mathbf{P})$.

In typical Reinforcement Learning, an agent with *policy* receives rewards at intermediate steps while the discount factor is used to balance long-term and short-term rewards. In our task, there is no intermediate rewards, only a final reward at terminal step T. Therefore, the value function of a partial sequence $\mathbf{c}_t = w_{1:t}$ is the expected reward at the terminal step.

$$V(w_{1:t}|\mathbf{P}) = \mathbb{E}_{w_{t+1:T}}[R(w_{1:t}; w_{t+1:T}|\mathbf{P})] \quad (3)$$

The objective of policy gradient is to maximize the average value starting from the initial state:

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N} V(w_0|\mathbf{I}) \quad (4)$$

where N is the total number of examples in training set. The gradient of $V(w_0|\mathbf{P})$ is computed as below (Williams, 1992):

$$\mathbb{E}_{w_{2:T}}[\sum_{t=1}^{T}\sum_{w_t\in\mathcal{V}} \nabla_\theta \pi_\theta(w_{t+1}|w_{1:t}, \mathbf{P}) \times Q(w_{1:t}, w_{t+1}|\mathbf{P})] \quad (5)$$

where $Q(w_{1:t}, w_t|\mathbf{P})$ is the *state-action* value for a particular *action* w_{t+1} at *state* $w_{1:t}$ given source passage $\mathbf{P}$, and should be calculated as follow:

$$Q(w_{1:t}, w_{t+1}|\mathbf{P}) = \mathbb{E}_{w_{t+2:T}}[R(w_{1:t+1}; w_{t+2:T}|\mathbf{P})] \quad (6)$$

Previous work (Liu et al., 2017) adopts Monte Carlo Rollout to approximate this expectation. Here we simply use the terminal reward $R(w_{1:T}|\mathbf{P})$ as an estimation with large variance. To compensate for the variance, we use a baseline estimator that doesn't change the validity of gradients (Williams, 1992). We further follow Paulus et al. (2018) to use the self-critical policy gradient training algorithm (Rennie et al., 2016; Williams, 1992). For each iteration, we sample a summary $y^s = w^s_{1:T+1}$, and greedily generate a summary $\hat{y} = \hat{w}_{1:T+1}$ by selecting the word with the highest probability at each step. Then these two summaries are fed to a reward function r that evaluates their closeness to the ground-truth. We choose ROUGE-L scores as the reward function r as in previous work (Paulus et al., 2018). The RL loss function is as follows:

$$\mathcal{L}_{RL} = \frac{1}{T}\sum_{t=1}^{T}(r(\hat{y}) - r(y^s))\log \pi_\theta(w^s_{t+1}|w^s_{1:t}) \tag{7}$$

where the reward for the greedily-generated summary ($r(\hat{y})$) acts as a baseline to reduce variance.

3 Training Details

We keep most of hyper-parameters and settings the same as in See et al. (2017). We use a bi-directional LSTM of 400 steps for the encoder, and a uni-directional LSTM of 100 steps for both decoders. All of our encoder and decoder LSTMs have hidden dimension of 256, and the word embedding dimension is set to 128. Our pre-set vocabulary has a total of 50k word tokens including special tokens for start, end, and out-of-vocabulary(OOV) signals. The embedding matrix is learned from scratch and shared between the encoder and two decoders.

All of our teacher forcing models reported are trained with Adagrad (Duchi et al., 2011) with learning rate of 0.15 and an initial accumulator value of 0.1. The gradients are clipped to a maximum norm of 2.0. The batch size is set to 16. Our model with closed-book decoder converged in about 200,000 to 240,000 iterations and achieved the best result on the validation set in another 2k~3k iterations with coverage loss added. We restore the best checkpoints (pre-coverage and post-coverage) and apply policy gradient (RL). For this phase of training, we choose Adam optimizer (Kingma and Ba, 2015) because of its time efficiency, and the learning rate is set to 0.000001. The RL-XE mixed-loss ratio (λ) is set to 0.9984.

4 Examples

We provide more example summaries generated by our 2-decoder and pointer-generator baseline (see Fig. 1, Fig. 2, Fig. 3).

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. 2017. Improved Image Captioning via Policy Gradient optimization of SPIDEr. In *International Conference on Computer Vision*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representation*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. In *Computer Vision and Pattern Recognition (CVPR)*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Original Text (truncated): lionel messi should be fit to play in barcelona's la liga game at celta vigo on sunday despite a scare over a possible foot injury , centre back gerard pique said on wednesday . messi , the top scorer in la liga , did not feature in either of argentina 's friendlies during the international break after sustaining a blow to his right foot in last month 's ` clasico ' against real madrid . ' i am optimistic about messi , i have spoken to him , ' pique told reporters at a promotional event . lionel messi -lrb- right -rrb- should be available for barcelona 's trip to celta vigo , according to gerard pique . ' i think that he can play at balaidos -lrb- celta 's stadium -rrb- , ' added the spain international , who came through barca 's youth academy with messi . ` in the end it is up to how he feels during the rest of the week . the medical staff are those who should decide . my feeling is that he will play . barca 's 2-1 win at home to real stretched their lead over their arch rivals at the top of la liga to four points with 10 games left . second-placed real , who host granada on sunday , have stuttered in recent weeks and the 2-1 defeat at the nou camp was their third loss in their last four outings in all competitions .

Reference summary:
lionel messi didn't feature in either of Argentina's recent friendlies .
messi suffered a foot injury in barcelona's win over real madrid last month .
barca sits four points clear of real in la liga with 10 games remaining.

Pointer-Gen baseline:
lionel messi should be available for barcelona‘s trip to celta vigo.
the **spain** international *did not feature in either of argentina‘s friendlies* during the international break.
messi, the top scorer in la liga, *did not feature in either of argentinal‘s friendlies*.

Pointer-Gen + closed-book decoder:
lionel messi should be available for barcelona 's trip to celta vigo.
messi , the top scorer in la liga , did not feature in either of argentina's friendlies during the international breakafter sustaining a blow to his right foot in last month's `clasico ' against real madrid.
barca 's 2-1 win at home to real stretched their lead over their arch rivals at the top of la liga to four points with 10 games.

Figure 1: The pointer-generator repeats itself (italic) and makes a factual error (red), while the 2-decoder (pointer-generator + closed-book decoder) generates the summary that recovers the salient information (highlighted) in the original text.

Original Text (truncated): a waitress has revealed how the new zealand prime minister had repeatedly given her unwanted attention while she was working at a cafe in auckland frequented by him and his wife . published on the daily blog on wednesday , the anonymous woman has recounted how john key kept playfully pulling her hair despite being told to stop during election time last year . however mr key defended his pranks as ' a bit of banter ' and said he had already apologised for his actions , stuff.co.nz reports . a waitress has revealed how the new zealand prime minister had repeatedly given her unwanted attention while she was working at a cafe in auckland frequented by him and his wife bronagh (pictured together) . the waitress had reportedly been working at a cafe called rosie (pictured) in parnell , east of auckland . the waitress - believed to be working at a cafe called rosie in parnell , east of auckland - wrote about she how made it very clear that she was unimpressed by mr key 's gestures . ' he was like the school yard bully tugging on the little girls ' hair trying to get a reaction , experiencing that feeling of power over her , ' she wrote on the blog . mr key kept being persistent with his hair-pulling antics , despite being told by his wife bronagh to stop . after dealing with the practical jokes over the six months he had visited the cafe , the waitress finally lost her cool ...

Reference summary:
amanda bailey , 26 , says she does n't regret going public with her story .
the waitress revealed in a blog how john key kept pulling her hair .
she wrote that she gained unwanted attention from him last year at a cafe .
ms bailey said mr key kept touching her hair despite being told to stop .
owners say they were disappointed she never told them of her concerns .
they further stated mr key is popular among the cafe staff .
the prime minister defended his actions , saying he had already apologised .
he also said his pranks were ` all in the context of a bit of banter '
the waitress was working at a cafe called rosie in parnell , east of auckland .

Pointer-Generator baseline:
waitress was working at a cafe in auckland frequented by him and his wife .
she was working at a cafe called rosie in parnell , east of auckland .
mr key defended his pranks as ' a bit of banter ' and said he had already apologised .

Pointer-Generator + closed-book decoder:
waitress has revealed how john key kept playfully pulling her hair despite being told to stop during election time last year .
however mr key defended his pranks as ' a bit of banter ' and said he had already apologised for his actions , stuff.co.nz reports .

Figure 2: The pointer-generator fails to address the most salient information from the original text, only mentioned a few unimportant points (where the waitress works), while the 2-decoder (pointer-generator + closed-book decoder) generates the summary that recovers the salient information (highlighted) in the original text.

Original Text (truncated): the suicides of five young sailors who served on the same base over two years has unearthed a shocking culture of ice taking , binge drinking , bullying and depression within the australian navy . the sailors were stationed or had been stationed at the west australian port of hmas stirling off the coast of rockingham , south of perth . their families did not learn of their previous attempts to take their own lives and their drug use until after their deaths , according to abc 's 7.30 program . scroll down for video . stuart addison was serving on hmas stirling off the coast of western australia when he took his own life . five of the sailors who committed suicide had been serving with the australian navy on hmas stirling ...

Reference summary:
five sailors took their own lives while serving on wa 's hmas stirling .
suicides happened over two years and some had attempted it before .
stuart addison 's family did n't know about his other attempts until his death .
it was a similar case for four other families , including stuart 's close friends .
revelations of ice use , binge drinking and depression have also emerged .

Pointer-Gen baseline:
stuart addison was serving on hmas stirling off the coast of western australia .
he was serving on hmas stirling off the coast of rockingham , south of perth .
their families did not learn of their previous attempts to take their own lives and their drug use until after their deaths .
their families did not learn of their previous attempts to take their own lives and their drug use until after their deaths .

Pointer-Gen + closed-book decoder:
the suicides of five young sailors who served on the same base over two years has unearthed a shocking culture of ice taking , binge drinking , bullying and depression within the australian navy .
the sailors were stationed at the west australian port of hmas stirling off the coast of rockingham , south of perth .
their families did not learn of their previous attempts to take their own lives and their drug use until after their deaths , according to abc 's 7.30 program .

Figure 3: The pointer-generator (non-coverage) repeats itself (italic), while the 2-decoder (pointer-generator + closed-book decoder) generates the summary that recovers the salient information (highlighted) in the original text as well as the reference summary.