Neural Speed Reading via Skim-RNN

Artificial Intelligence Laboratory
오 주민

Introduction

- ICLR 2018에 게제된 논문
- LSTM-Jump는 Token을 Skip하는 형태
- Skip이 아닌 Skim(드문드문 읽기)를 이용하여 Computation speed를 높이고 Loss를 낮추자

Introduction

- 목표

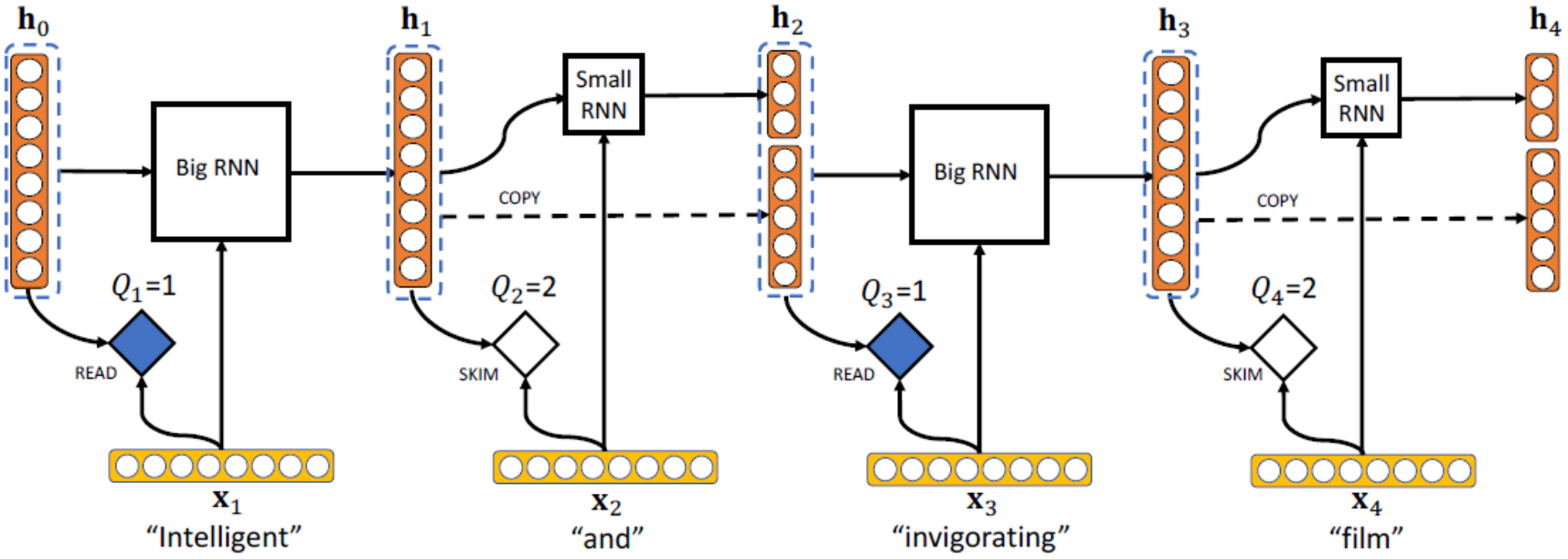
① 중요한 정보는 열심히 읽고 덜 중요한 정보는 대충 읽자

② Learning Speed와 Loss 두 마리 토끼를 잡아보자

➢ Skimming를 통해 해결하자!

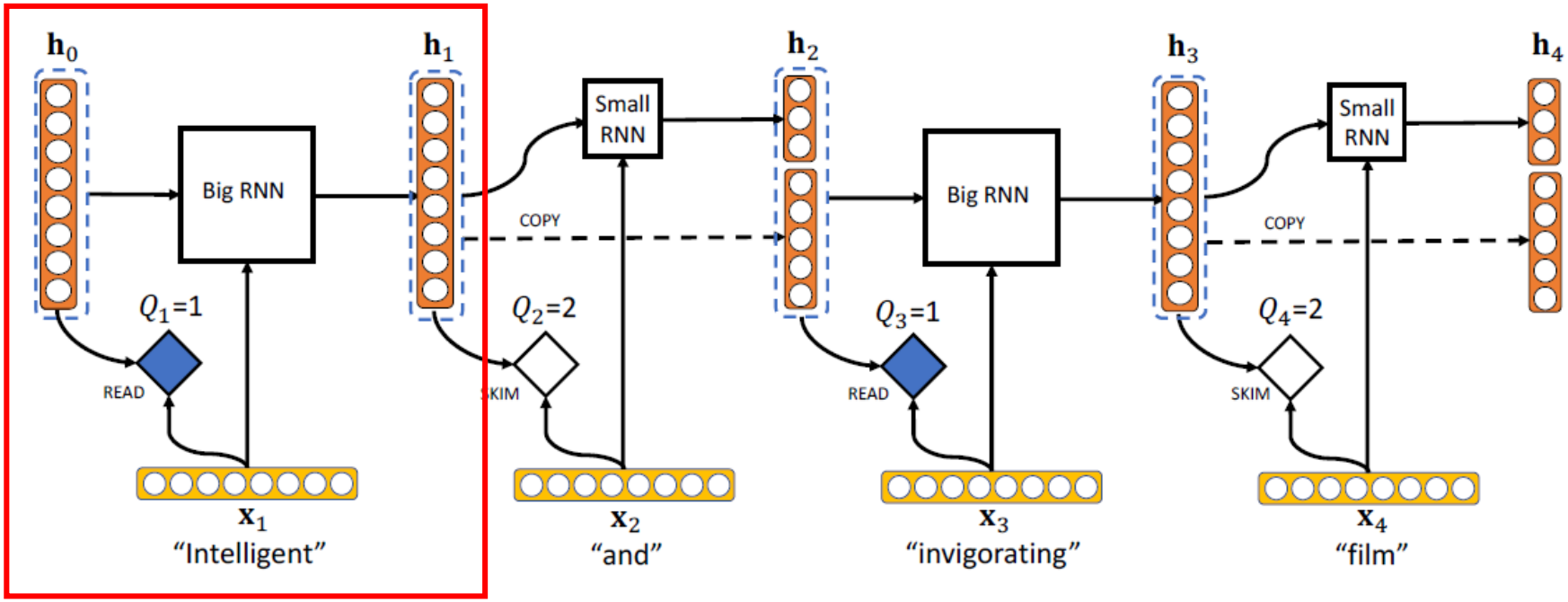
Model

- 두 개의 RNN을 이용해보자!
  - Big RNN & Small RNN
  - 두 RNN의 parameter는 독립적이며 각각 d, d'의 사이즈를 가짐(d >> d')



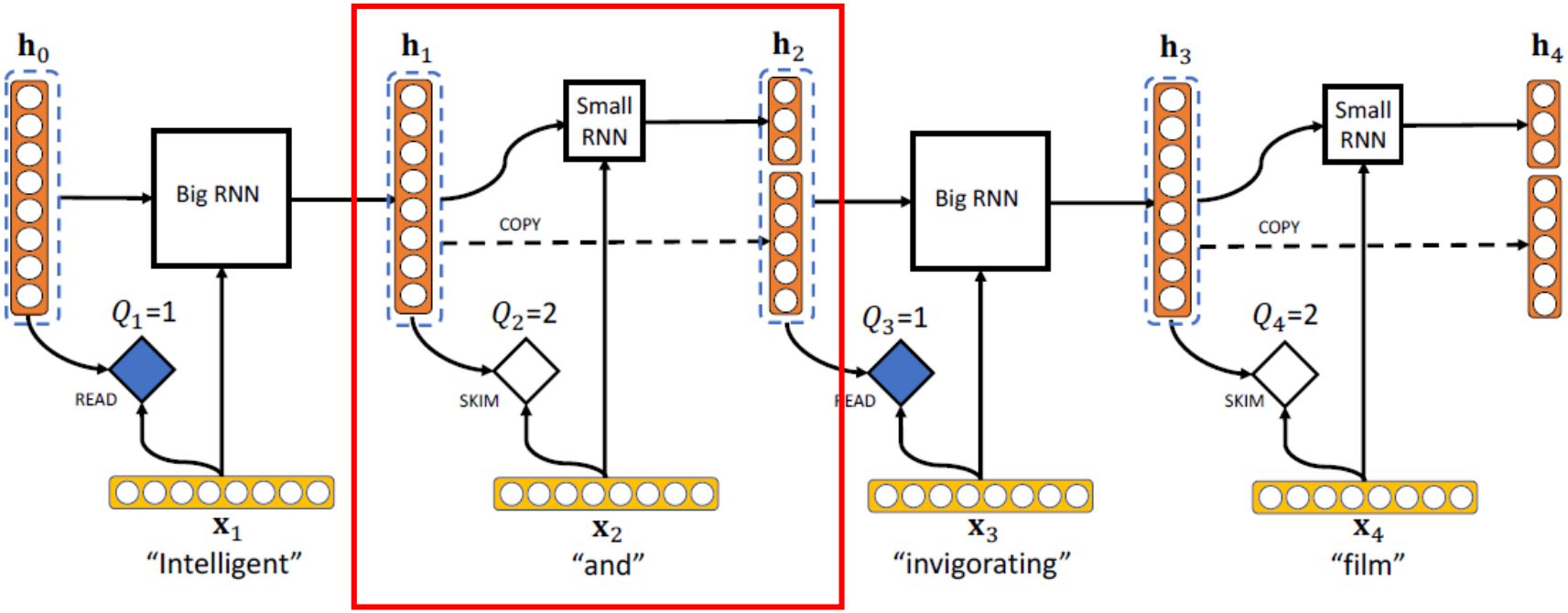
Model

- Big RNN
  - 중요한 Input에 대해서 처리하는 RNN
  - 현재 Step의 Input과 이전 Step의 Hidden state를 통해 계산(기존과 동일)



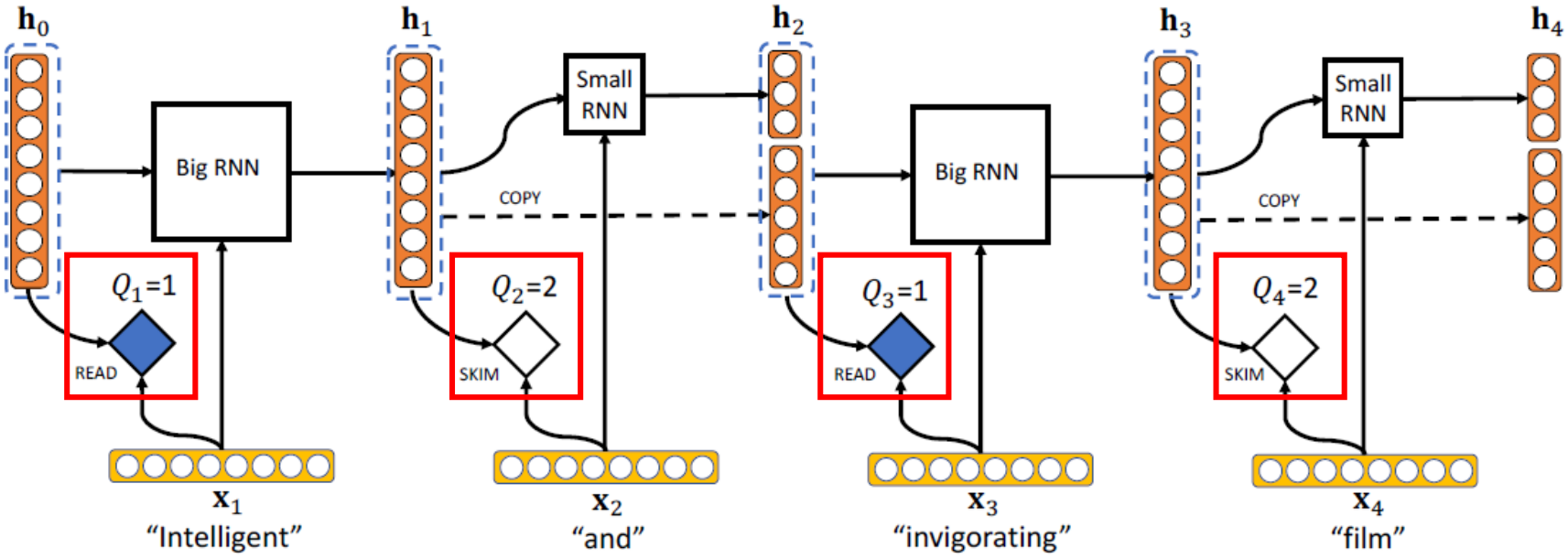
Model

- Small RNN
  - 덜 중요한 Input에 대해서 처리하는 RNN
  - 현재 Step의 Input과 이전 Step의 Hidden state를 통해 계산을 하되 Big RNN의 출력 Size와 맞추기 위해 이전 Step의 Hidden state를 Copy하여 Concatenate



Model

- Decision to skim
  - 각 Step에서 Big RNN을 사용할 지, Small RNN을 사용할 지 결정
  - Q_t를 계산하여 그 값에 의해 결정
  - Q_t는 어떻게 계산할까?



Model

- Q_t는 p_t에 의해 결정되는 Multinomial random variable

- $$\mathbf{p}_t = \text{softmax}(\alpha(\mathbf{x}_t, \mathbf{h}_{t-1})) = \text{softmax}(\mathbf{W}[\mathbf{x}_t; \mathbf{h}_{t-1}] + \mathbf{b}) \in \mathbb{R}^k$$

  본 논문에서는 Computation advantage를 위해 α함수를 Input vector와 Hidden state를 Concatenate하고 projection 했음

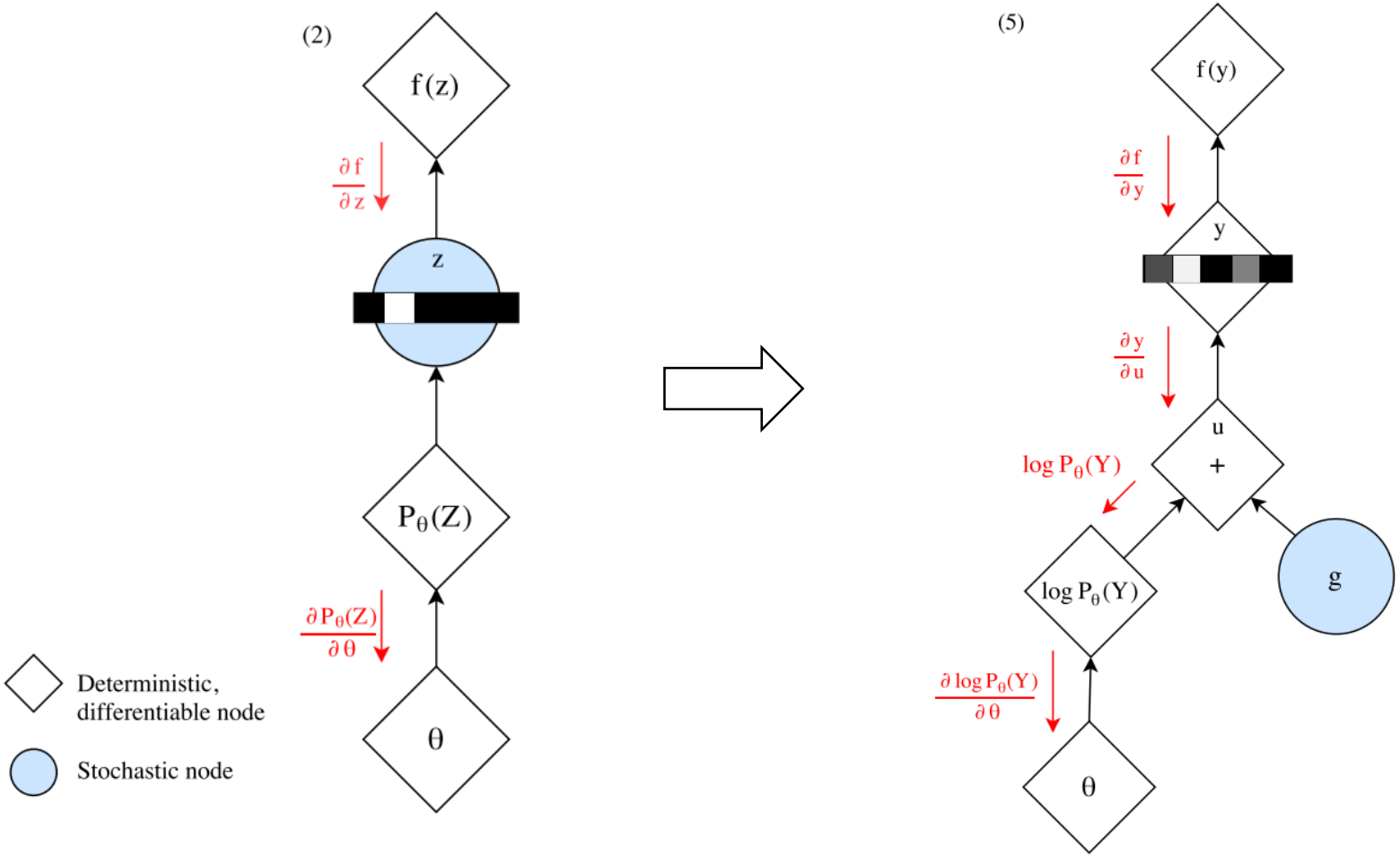
- $$Q_t \sim \text{Multinomial}(\mathbf{p}_t)$$

  p_t에 근거한 Multinomial distribution에서 sampling을 하여 Q_t 결정

- 근데 학습 단계에서 Sampling한 값을 역전파시키냥?

Gumbel-softmax

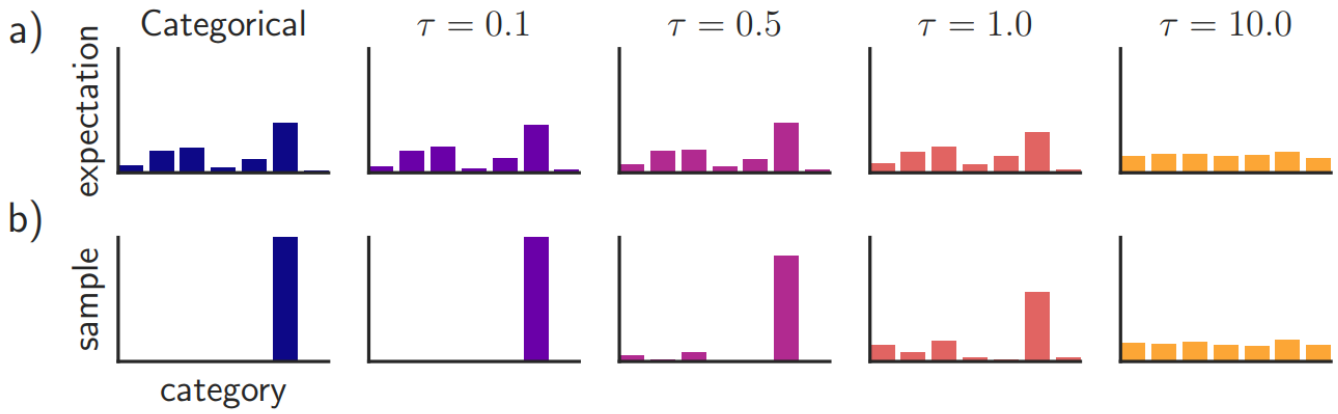
- Stochastic variable의 back-propagation이 가능하도록 reparameterised form으로 변경



Gumbel-softmax

- $$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)}$$

- τ : Smooth relaxation parameter(a.k.a Temperature)

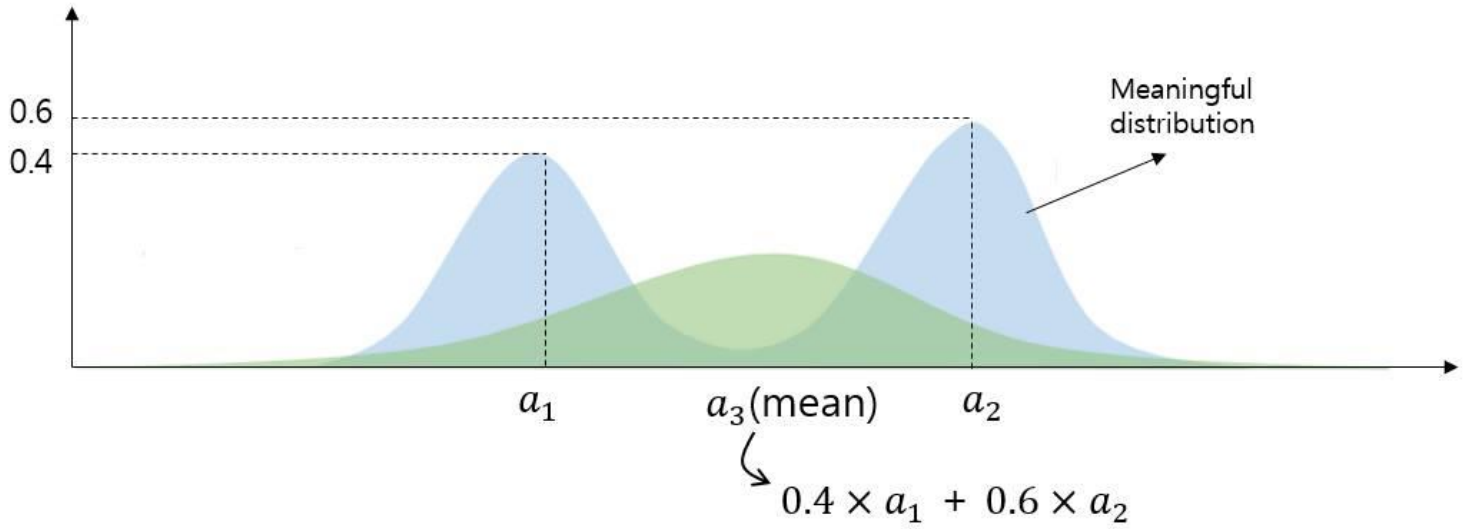


➢More discrete하고 Q_t의 분포에 따라가도록 $\tau \to 0$

Gumbel-softmax

- 그냥 Softmax 값 쓰면 안되요?
  - 응 안돼

- Expectation과 Sampling의 차이



Model

- Reparameterized distribution r_t

$$\mathbf{r}_t^i = \frac{\exp((\log(\mathbf{p}_t^i) + g_t^i)/\tau)}{\sum_j \exp((\log(\mathbf{p}_t^j) + g_t^j)/\tau)}$$

- Inference 할 때는 Q_t를 쓰고 Training 할 때는 r_t쓰자

- $\mathbf{h}_t = \sum_i \mathbf{r}_t^i \tilde{\mathbf{h}}_t^i$

  where, $\tilde{\mathbf{h}}_t^1 = f(\mathbf{x}_t, \mathbf{h}_{t-1})$

  $\tilde{\mathbf{h}}_t^2 = [f'(\mathbf{x}_t, \mathbf{h}_{t-1}); \mathbf{h}_{t-1}(d'+1:d)]$

Model

- Loss

$$L'(\theta) = L(\theta) + \gamma \frac{1}{T} \sum_{t} -\log(\mathbf{p}_t^2)$$

Skim이 할만 할 때, 더 자주 Skim하라고
Skimming의 Negative log probability 추가

Experiment and Result

Dataset	task type	answer type	Number of examples	Avg. Len	vocab size
SST	Sentiment Analysis	Pos/Neg	6,920 / 872 / 1,821	19	13,750
Rotten Tomatoes	Sentiment Analysis	Pos/Neg	8,530 / 1,066 / 1,066	21	16,259
IMDb	Sentiment Analysis	Pos/Neg	21,143 / 3,857 / 25,000	282	61,046
AGNews	News classification	4 categories	101,851 / 18,149 / 7,600	43	60,088
CBT-NE	Question Answering	10 candidates	108,719 / 2,000 / 2,500	461	53,063
CBT-CN	Question Answering	10 candidates	120,769 / 2,000 / 2,500	500	53,185
SQuAD	Question Answering	span from context	87,599 / 10,570 / -	141	69,184

- Natural Language Processing 분야의 다양한 문제에 대한 실험

- Four for Text Classification
- Three for QA

Experiment and Result

LSTM Model	d'/γ	SST Acc	SST Sk	SST Flop-r	SST Sp	Rotten Tomatoes Acc	Rotten Tomatoes Sk	Rotten Tomatoes Flop-r	Rotten Tomatoes Sp	IMDb Acc	IMDb Sk	IMDb Flop-r	IMDb Sp	AGNews Acc	AGNews Sk	AGNews Flop-r	AGNews Sp
Standard		86.4	-	1.0x	1.0x	82.5	-	1.0x	1.0x	91.1	-	1.0x	1.0x	93.5	-	1.0x	1.0x
Skim	5/0.01	**86.4**	58.2	2.4x	1.4x	**84.2**	52.0	2.1x	1.3x	89.3	79.2	4.7x	2.1x	**93.6**	30.3	1.4x	1.0x
Skim	10/0.01	85.8	61.1	2.5x	1.5x	82.5	58.5	2.4x	1.4x	**91.2**	83.9	5.8x	2.3x	93.5	33.7	1.5x	1.0x
Skim	5/0.02	85.6	62.3	2.6x	1.5x	81.8	63.7	2.7x	1.5x	88.7	63.2	2.7x	1.5x	93.3	36.4	1.6x	1.0x
Skim	10/0.02	**86.4**	68.0	3.0x	1.7x	82.5	63.0	2.6x	1.5x	90.9	90.7	9.5x	2.7x	92.5	10.6	1.1x	0.8x
LSTM-Jump		-	-	-	-	79.3	-	-	1.6x	89.4	-	-	1.6x	89.3	-	-	1.1x
VCRNN		81.9	-	2.6x	-	81.4	-	1.9x	-	-	-	-	-	-	-	-	-
SOTA		89.5	-	-	-	83.4	-	-	-	94.1	-	-	-	93.4	-	-	-

- Result of Text classification
  - d' : Small RNN Size
  - γ : Skim loss ratio
  - Standard는 LSTM cell을 사용한 RNN
  - Sk : Skimming rate (percentage)
  - Flop-r : float operating reduction rate
  - Sp : Speed up rate

Experiment and Result

- Example of Sentimental Analysis
  - 파란 단어는 Fully Read 한 것
  - 검은 단어는 Skimming 한 것

Positive	I **liked** this movie, not because Tom Selleck was in it, but **because** it was a **good** story about baseball and it also had a semi-over **dramatized** view of some of the issues that a BASEBALL player coming to the end of their time in Major League sports must face. I also **greatly enjoyed** the cultural differences in American and Japanese baseball and the small facts on how the games are played differently. **Overall**, it is a **good movie** to watch on Cable TV or rent on a cold winter's night and watch about the "Dog Day's" of summer and know that spring training is only a few months away. A **good** movie for a baseball fan as well as a good "DATE" movie. Trust me on that one! *Wink*
Negative	**No**! **no** - **No** - **NO**! My **entire** being is **revolting** against this **dreadful** remake of a classic movie. I **knew** we were heading for trouble from the moment Meg Ryan appeared on screen with her **ridiculous** hair and clothing - literally looking like a scarecrow in that garden she was digging. Meg Ryan playing Meg Ryan - how **tiresome** is that?! And it **got worse** ... so much **worse**. The **horribly cliché** lines, the stock characters, the increasing sense I was **watching** a spin-off of "The First Wives Club" and the ultimate **hackneyed** schtick in the delivery room. How many times have I seen this movie? Only once, but it **feel** like a dozen times - **nothing** original or fresh about it. For shame!

Experiment and Result

- Example of Sentimental Analysis
  - 파란 단어는 Fully Read 한 것
  - 검은 단어는 Skimming 한 것

Positive	I **liked** this movie, not because Tom Selleck was in it, but **because** it was a **good** story about baseball and it also had a semi-over **dramatized** view of some of the issues that a BASEBALL player coming to the end of their time in Major League sports must face. I also **greatly enjoyed** the cultural differences in American and Japanese baseball and the small facts on how the games are played differently. **Overall**, it is a **good movie** to watch on Cable TV or rent on a cold winter's night and watch about the "Dog Day's" of summer and know that spring training is only a few months away. A **good** movie for a baseball fan as well as a good "DATE" movie. Trust me on that one! *Wink*
Negative	**No**! **no** - **No** - **NO**! My **entire** being is **revolting** against this **dreadful** remake of a classic movie. I **knew** we were heading for trouble from the moment Meg Ryan appeared on screen with her **ridiculous** hair and clothing - literally looking like a scarecrow in that garden she was digging. Meg Ryan playing Meg Ryan - how **tiresome** is that?! And it **got worse** ... so much **worse**. The **horribly cliché** lines, the stock characters, the increasing sense I was **watching** a spin-off of "The First Wives Club" and the ultimate **hackneyed** schtick in the delivery room. How many times have I seen this movie? Only once, but it **feel** like a dozen times - **nothing** original or fresh about it. For shame!

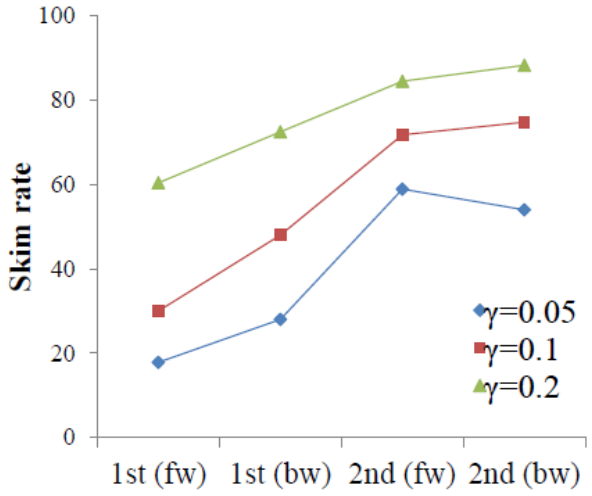
Experiment and Result

- Result of QA

Model	γ	F1	EM	Sk	Flop-r
LSTM+Att (1 layer)	-	73.3	63.9	-	1.3x
LSTM+Att ($d = 50$)	-	74.0	64.4	-	3.6x
LSTM+Att	-	75.5	**67.0**	-	1.0x
Sk-LSTM+Att ($d' = 0$)	0.1	**75.7**	66.7	37.7	1.4x
Sk-LSTM+Att ($d' = 0$)	0.2	75.6	66.4	49.7	1.6x
Sk-LSTM+Att	0.05	75.5	66.0	39.7	1.4x
Sk-LSTM+Att	0.1	75.3	66.0	56.2	1.7x
Sk-LSTM+Att	0.2	75.0	66.0	76.4	2.3x
VCRNN	-	74.9	65.4	-	1.0x
BiDAF ($d = 30$)	-	74.6	64.0	-	9.1x
BiDAF ($d = 50$)	-	75.7	65.5	-	3.7x
BiDAF	-	**77.3**	**67.7**	-	1.0x
Sk-BiDAF	0.01	76.9	67.0	74.5	2.8x
Sk-BiDAF	0.001	77.1	67.4	47.1	1.7x
SOTA (Wang et al., 2017)		79.5	71.1	-	-

Experiment and Result

- Result of QA



  - Layer가 쌓여갈수록 높은 Layer가 Skim을 결정하는데 더욱 Confident

Conclusion

- Skim-RNN을 이용할 때의 장점
  - 높은 정확도
  - 빠른 계산 (Advantage in Computation)

- Future Work
  - Hidden state size가 큰 application(Video understanding)에 적용해볼 예정

- 우리가 얻을 수 있는 Insight
  - Gumbel-Softmax를 이용한 discrete category 처리
  - 더 빠르게 동작하는 RNN Model