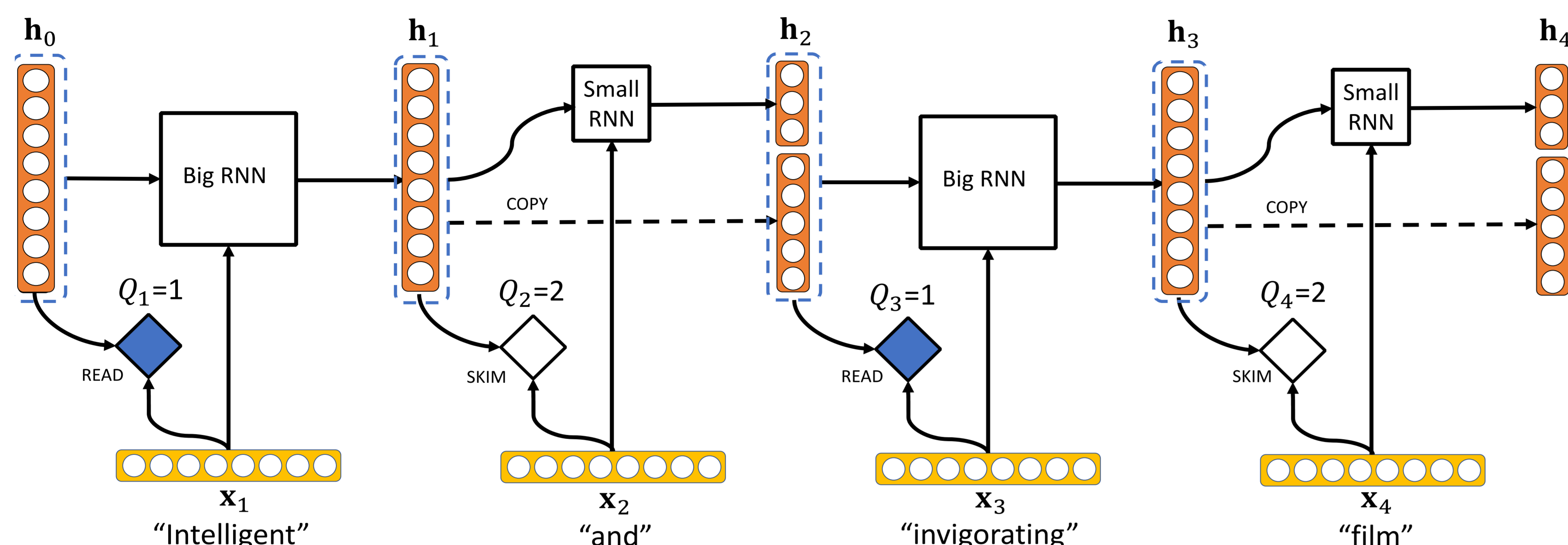


Neural Speed Reading via Skim-RNN

Minjoon Seo^{1,2*}, Sewon Min^{3*}, Ali Farhadi^{1,4,5}, Hannaneh Hajishirzi¹
 University of Washington¹, NAVER Clova², Seoul National University³, Allen Institute for AI⁴, XNOR.AI⁵

ICLR 2018 @Vancouver, Canada - April 30 (Mon)



Motivation

RNN on CPUs

- RNNs are slow on CPUs/GPUs
- CPUs are often more desirable than GPUs

Human Speed Reading

- Skim** unimportant words and **fully read** important words

Our contributions

- Dynamically decide to skim by choosing a small or big RNN with a discrete random variable
- Same interface as standard RNN → Easily Replaceable
- Computational advantage over standard RNN on CPUs with comparable/better accuracy

Model

Consists of two RNNs, sharing hidden state

- Big RNN** (hidden size of d) updates the entire hidden state. $O(d^2)$
- Small RNN** (hidden size of d' , where $d \gg d'$) updates a small portion of the hidden state. $O(d'd)$

Formulation

$\mathbf{x}_t, \mathbf{h}_t$: Input/Hidden state at t

Q_t : Random variable for skim decision at t

$$\mathbf{p}_t = \text{softmax}(\alpha(\mathbf{x}_t, \mathbf{h}_{t-1}))$$

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

$$\mathbf{h}_t = \begin{cases} f(\mathbf{x}_t, \mathbf{h}_{t-1}), & \text{if } Q_t = 1 \\ [f'(\mathbf{x}_t, \mathbf{h}_{t-1}); \mathbf{h}_{t-1}[d'+1:d]], & \text{if } Q_t = 2 \end{cases}$$

(f, f' : big and small RNN, respectively)

Training

$$L'(\theta; Q) = L(\theta; Q) + \gamma \frac{1}{T} \sum_{t=1}^T (-\log(\Pr[Q_t = 2]))$$

(γ : hyperparameter, encourages skimming)

Goal: minimize $\mathbb{E}[L'(\theta)] = \sum_Q L'(\theta; Q) \Pr(Q)$

→ sample space = 2^T for $Q = [Q_1, \dots, Q_T]$

Gumbel-Softmax

- Biased but low variance, good empirical results
- Reparameterize $\mathbf{h}_t = \mathbf{r}_t^1 \mathbf{h}_t^1 + \mathbf{r}_t^2 \mathbf{h}_t^2$, where

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

($g \sim \text{Gumbel}(0, 1)$, τ : hyperparameter)

- Slowly anneal (decrease τ), making the distribution more discrete to allow differentiation with stochasticity

Related Work

LSTM-Jump (Yu et al., 2017)

- Skip rather than skim
- No outputs for skipped time steps

Variable-Computation RNN (VCRNN) (Jernite et al., 2017)

- Use 1 RNN, and controls number of hidden units to update by making a d -way decision
- Required to be invariant of number of hidden units being updated
- Computing loss is intractable (d^T vs 2^T)

Experiments

Text Classification (w/ LSTM)

Dataset	SST	Rotten Tomatoes	IMDb	AGNews
Regular	86.4	82.5	91.1	93.5
LSTM-Jump	-	79.3 / 1.6x Sp	89.4 / 1.6x Sp	89.3 / 1.1x Sp
VCRNN	81.9 / 2.6x FLOP-R	81.4 / 1.6x Sp	-	-
Skim-RNN	86.4 / 3.0x FLOP-R	84.2 / 1.3x Sp	91.2 / 2.3x Sp	93.6 / 1.0x Sp

- ▲ Accuracy, Floating point operation reduction (FLOP-R) and Speed-up (Sp).
- Skim-RNN achieves comparable/better accuracy than others, with up to 3.0x FLOP-R.

Question Answering (SQuAD) (w/ LSTM+Attention)

	F1	Exact Match	FLOP-R
Regular	75.5	67.0	1.0x
VCRNN	74.9	65.4	1.0x
Skim-RNN	75.0	66.0	2.3x

- ▲ Skim-RNN achieves comparable result to Regular RNN with 2.3x FLOP-R on SQuAD.

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!

- ▲ Examples on IMDb. Skim-RNN skims **black words** and fully reads **blue words**.

