Visual Analysis of Recurrent Neural Networks

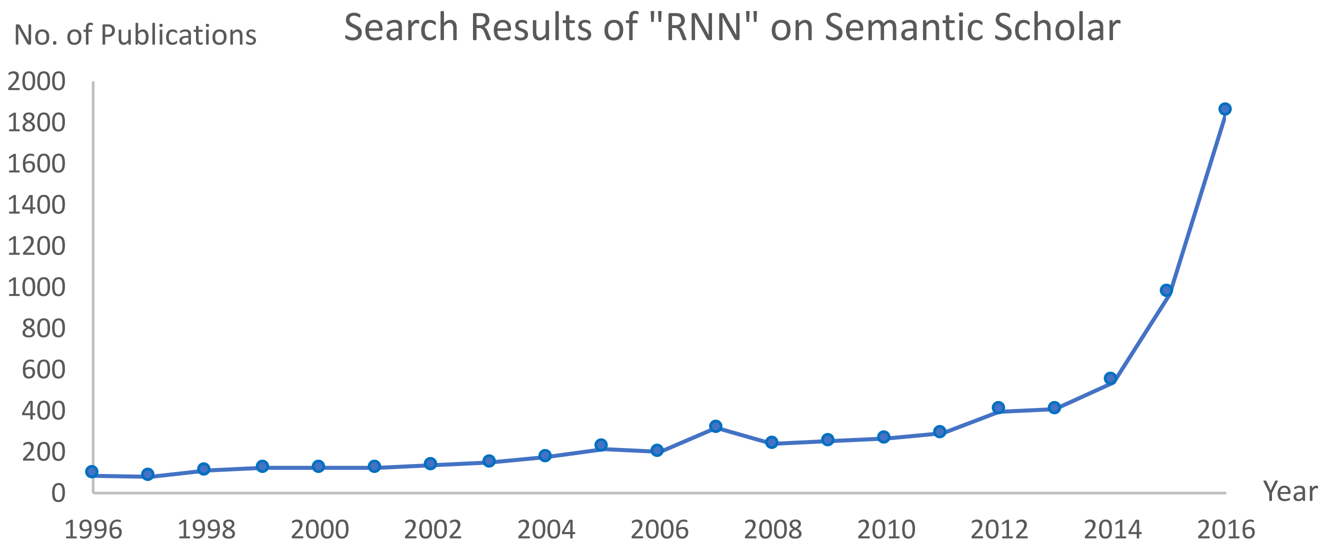
Yao MING

Jan 11, 2017

Introduction

Recurrent Neural Networks (RNNs)

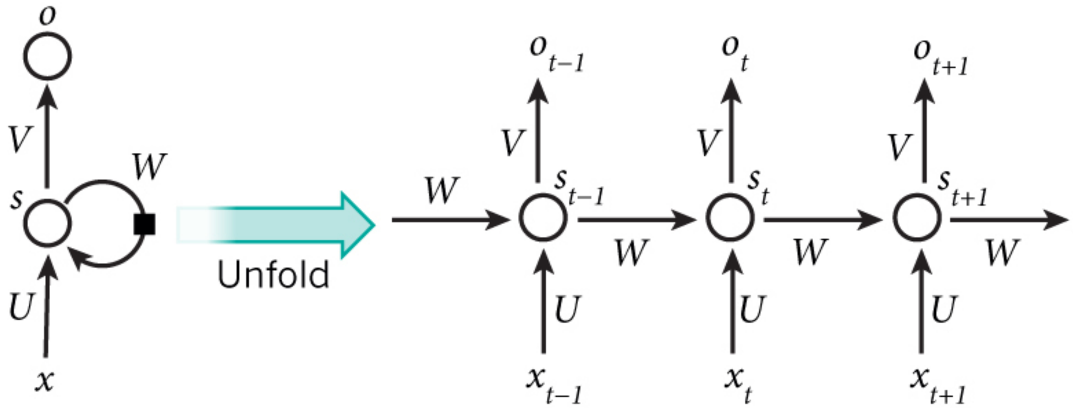
- More general and flexible than Feedforward Neural Network or CNN, operates on **sequential data** like text and audios
- Effective in: language modeling, machine translation, speech recognition, sentiment analysis, image captioning…
- With fast growing research attention



Introduction

Vanilla RNN (single layered)

- Like a scanner, each time step, update **hidden state** s_t using input x_t and previous s_{t-1}, and output o_t.
- Weights are shared over time steps.



$$s_t = f(Ux_t + Ws_{t-1})$$

$$o_t = softmax(Vs_t)$$

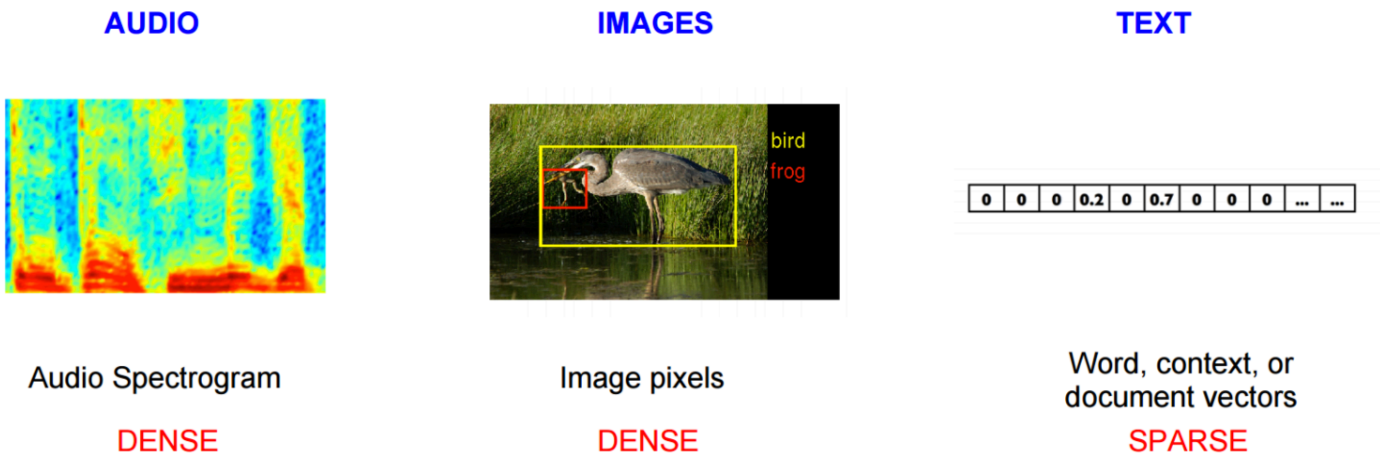
f is nonlinear e.g.: tanh

Variants

- Long Short Term Memory networks (LSTMs)
- Gated Recurrent Unit (GRU)
- Bi-directional LSTMs

Introduction

Word Embedding (compress input of RNNs)



AUDIO

Audio Spectrogram

DENSE

IMAGES

Image pixels

DENSE

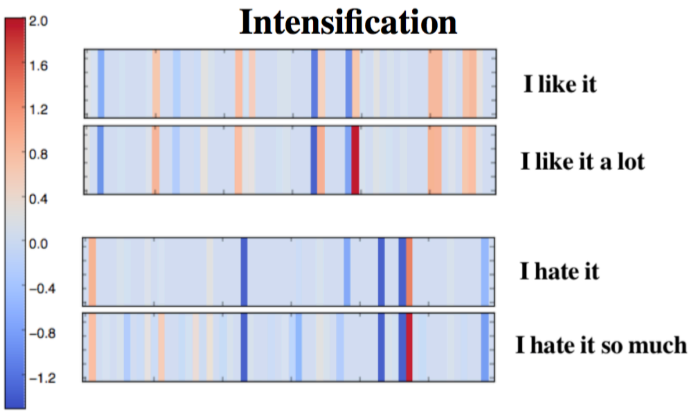
TEXT

0	0	0	0.2	0	0.7	0	0	0	...	...

Word, context, or document vectors

SPARSE

- Compress word vector from **several thousands** of dimensions to **tens or hundreds** of dimensions
- Some of the dimensions may contain semantic meaning or sentiments



The Problem is

In one sentence,

- How to **compare** different RNNs and better **understand** their **inner mechanisms** (capture keyword, long term memory) to help **improvements** and **understanding** of RNNs

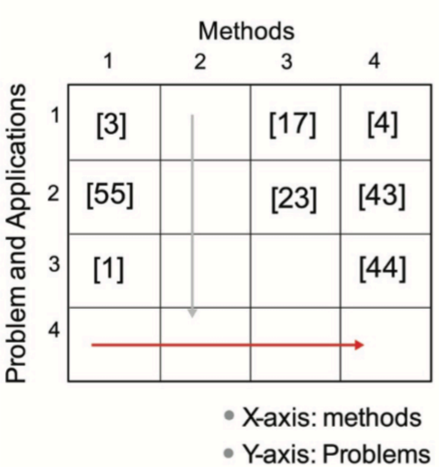
Specifically,

- How to encode the **hidden states** and **gate activation** information into visualization to **intuitively** presents what RNNs learned from text (or audios).
- Design more convenient visualization for **comparatively** analyzing model's **performance/error** on **datasets** for improvements.

The Problem is

An "Old" Problem

- Known issues of interpretability of what deep Neural Networks (with **millions** of parameters) have learned
- Poor knowledge in the **source** of RNNs' impressive performance and the **shortcomings** of different RNNs (Karpathy et. al, 2015)
- Visualizing RNNs are less studied than CNNs, and are more challenging (Images are visualizations, but text/audios are not.)
- Existing work (LSTMVis, 2016) only focus on exploring hidden state patterns of pre-trained models.



	1	2	3	4
1	[3]		[17]	[4]
2	[55]		[23]	[43]
3	[1]			[44]
4				

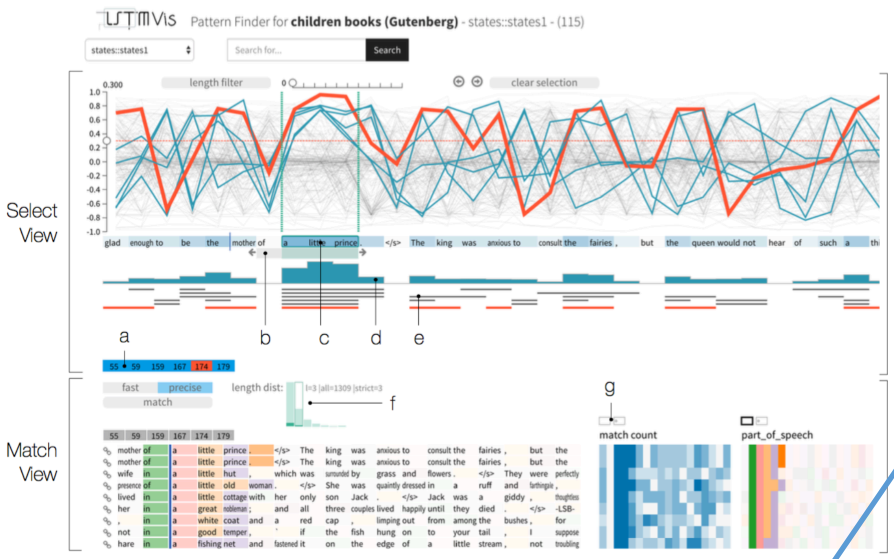
- X-axis: methods
- Y-axis: Problems

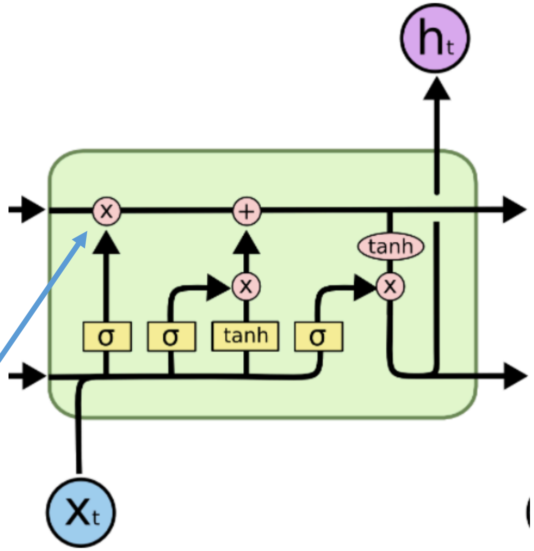
Contributions

- Clearer understanding on RNNs' inner mechanism (long and short term learning ability, hidden state interpretation, gating mechanisms, etc.), which drives further improvements on RNNs' architecture
- A user-friendly visualization tool for:
    - analyzing model's comprehensive performance (debugging)
    - comparing pros and cons of different models (understanding)
    - exploring datasets and alternative models (improvements)

Related Works

Closely Related Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. (arxiv Jun, 2016)





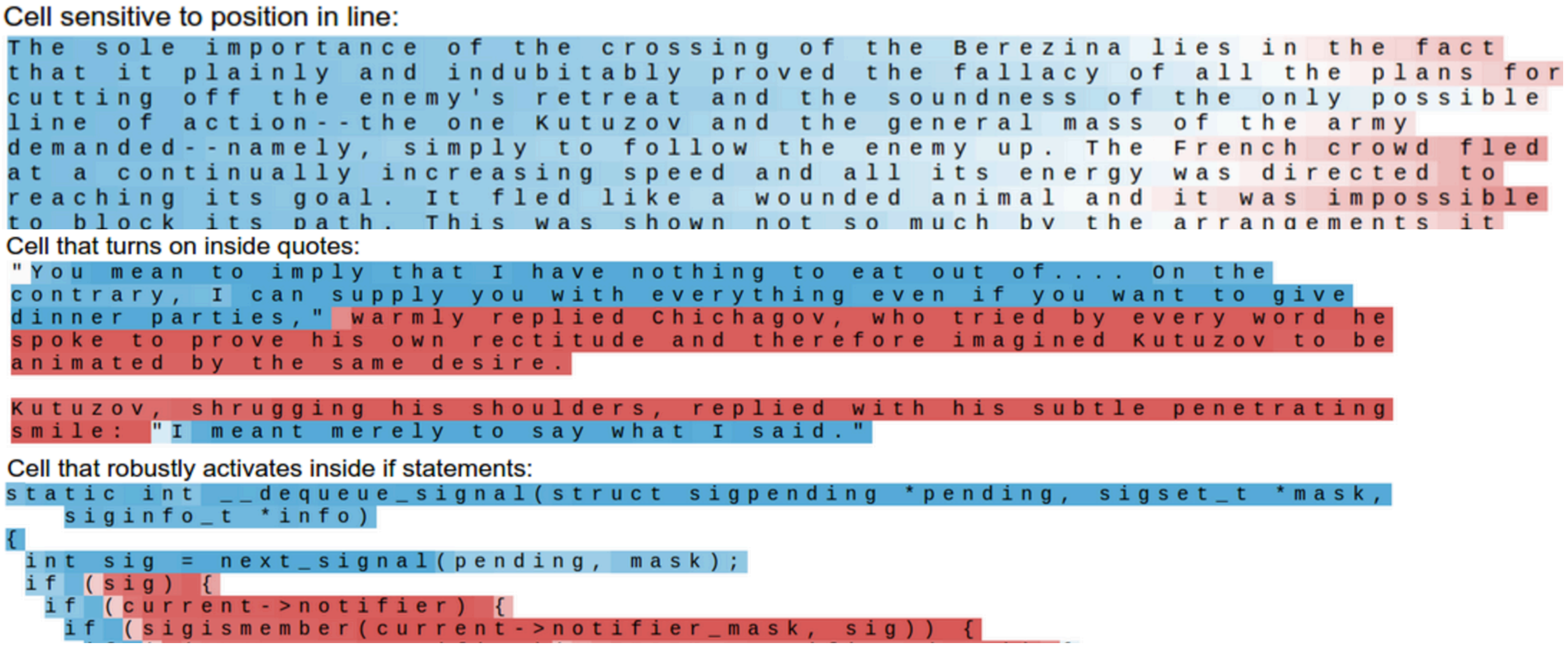
- It fails to analyze **gating structure** -- a **key** design of popular RNNs (LSTM, GRU) to control the update of hidden states and output.
- Also fails to analyze **error/performance** of a RNN (which is important in model evaluation).

Related Works

Understanding RNNs

Visualizing and Understanding Recurrent Networks (Karpathy, A., Johnson, J. and Fei-Fei, L., arxiv 2015 / ICLR 2016):

- a practical **error analysis method** which break errors into different categories to guide model improvements;
- model interpretations which can be of reference in case study.



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

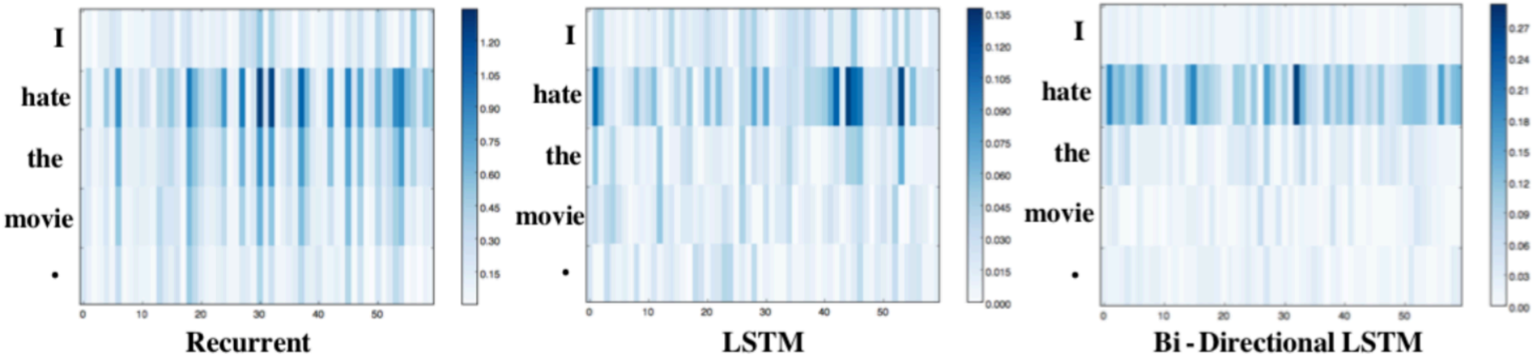
```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
   siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
```

Related Works

Understanding RNNs

Visualizing and Understanding Neural Models in NLP (Li, J. et. al., NAACL-HLT 2016):

- explorations of models' capability in several language phenomenon (concessive sentence, negation, intensification) (for case study)
- visual designs to compare models ability in learning word embedding and capturing keywords (design alternatives)



Recurrent

LSTM

Bi - Directional LSTM

Related Works

Visual analysis of Neural Networks

- Visualizing the Hidden Activity of Artificial Neural Networks (VAST 2016)

  t-SNE projection of neuron activations is inspiring, maybe used in the exploration phase.

- Towards Better Analysis of Deep CNNs (VAST 2016)

  Formulate CNN as DAG and bi-clusters, structural visualization.

Others:

- Model Performance Analysis: Squares (VAST 2016) …
- Interactive Machine Learning…

Design Tasks

High

Priority

1. Presents the learned features of RNNs (hard)
    a) Need to be intuitively **interpretable** (language phenomenon etc.)
    b) Should provide convenient **exploration** interactions (many dimensions of hidden states make no sense)
    c) Experts are interested in **hidden state** and **gate activation** information (haven't validated with experts yet)
2. Comprehensively analyze model's **error/performance**
3. **Comparative** analysis of different models
4. Visualization of RNNs' training process and debugging information

Low

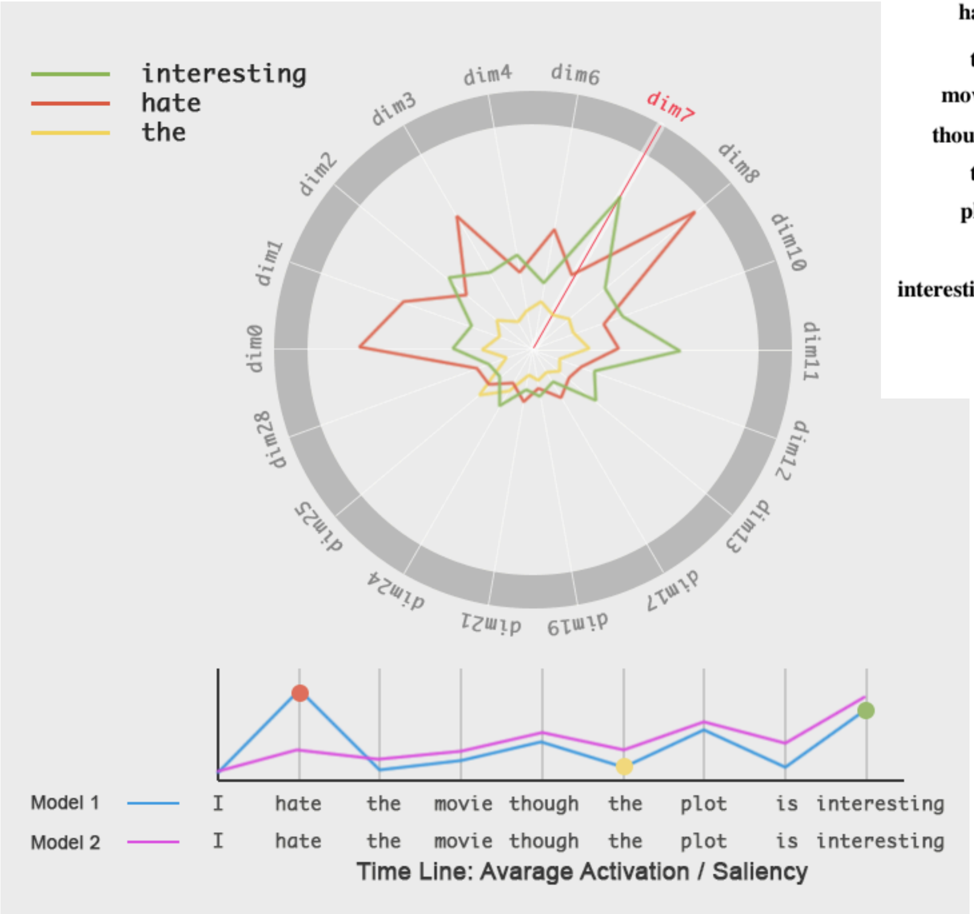
Visual Designs

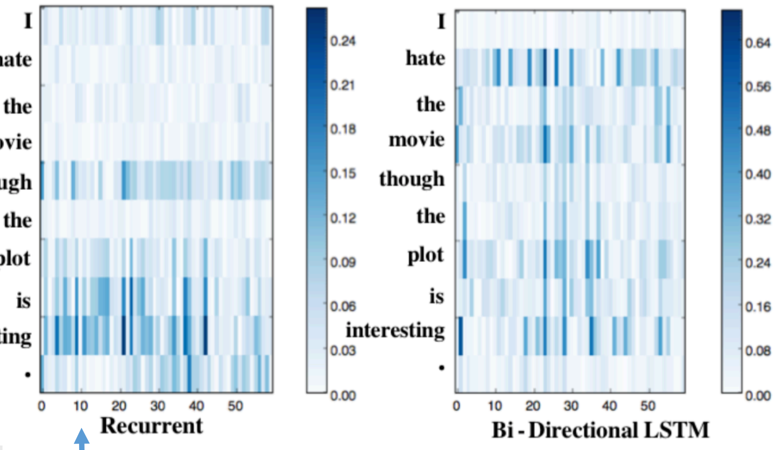
- Summary View (T2, T4):
  - Error: simple pie chart and data sample for categorical error analysis
  - Training Information: simple line chart for loss curve
- Exploration View (T1.b,c):
  - Exploring interpretable hidden states, salient dimensions of word embedding and **analyze gate activation**.
- Detail View (T2):
  - Use heat map to show detailed hidden state change and gate activations

For T3, add compare (side by side / overlay) to each views

Visual Designs

Exploration View – Semantic Radar

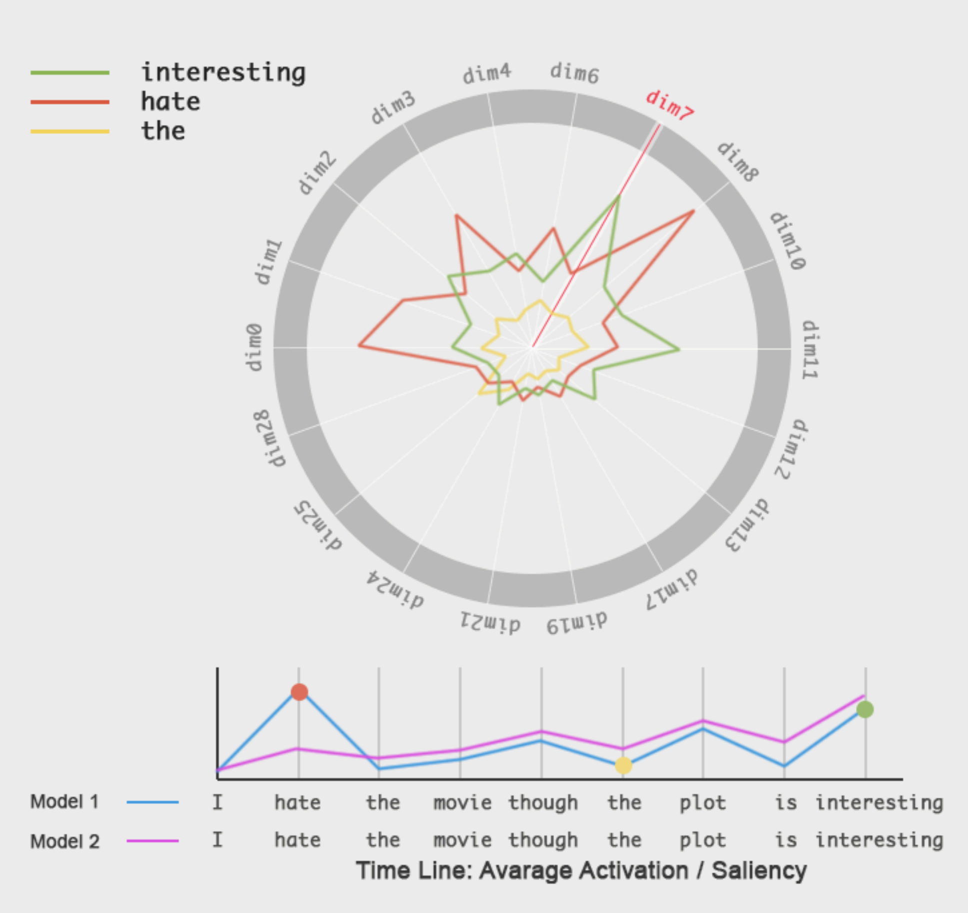




(Li et. al. 2016) Hard to find which dim of word embedding is more salient

Visual Designs

Exploration View – Semantic Radar



Functions:

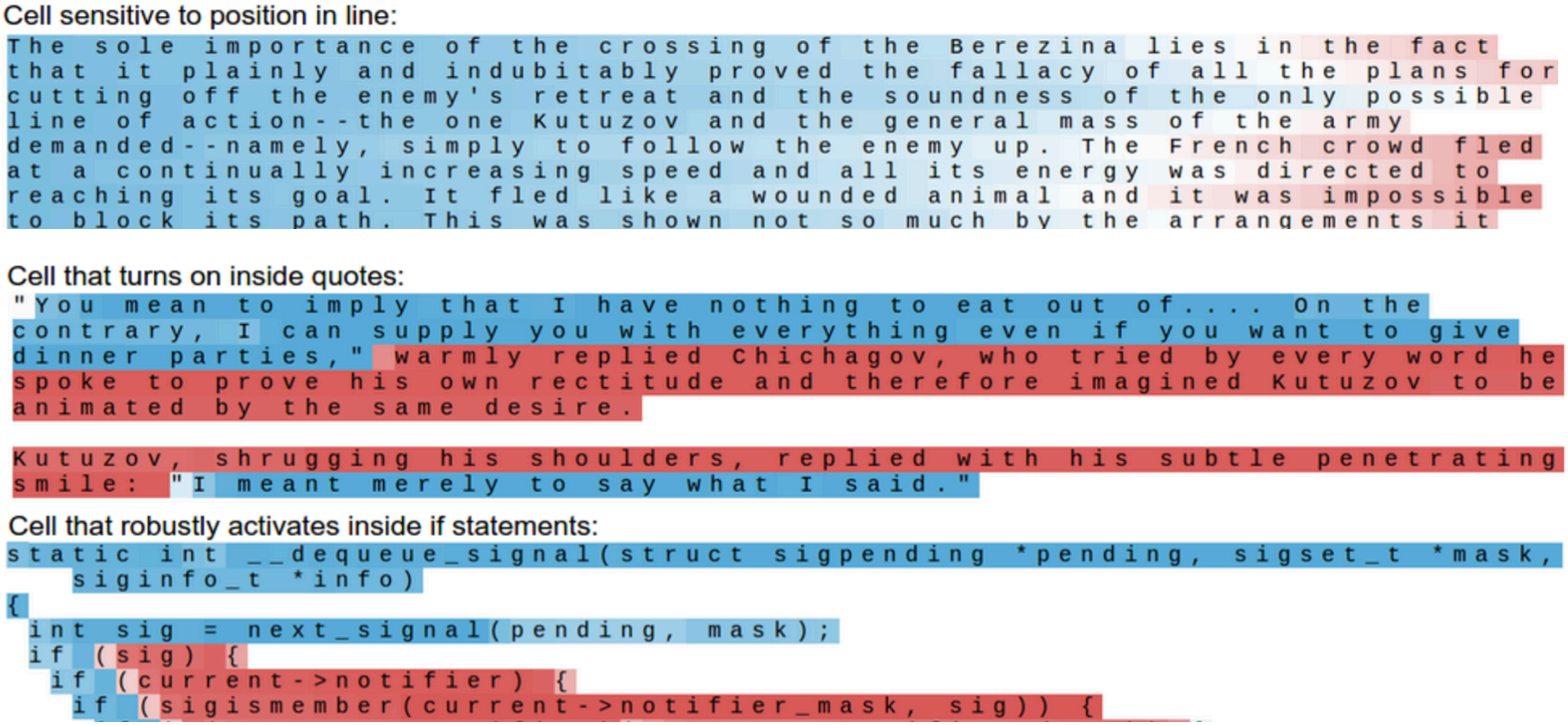
- Finding **interpretable** states
- Validate Gate Mechanism
- Compare **keyword** capturing capability

Options:

- **Hidden states** (50~1000d/lr.)
- **Gate Activations** (same as ↑)
- **Saliency scores** of each dim of word embedding (50~1000d)

Visual Designs

Detail View – Heat map (Karpathy et. al. 2015)



Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
    siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
```

Expected Results

Tools

- RNN implementation: TensorFlow
- Visualization: D3.js
- Framework: Flask (app), Vue.js(front-end)

Datasets:

- Treebanks: Penn Treebank, Stanford Sentiment Treebank
- Plain Text and source code: Linux Kernel Code
- …

Remaining Tasks

Unsolved Major Problems

- Complete and refine visual design
- More literature reviews on existing methods in understanding RNNs and text visualization
- …

Workload:

- Coding
  - RNN and Data Preprocessing ★★★?
  - Visual System ★★★★ ? ?
- Visual Design and Evaluation ★★★?
- Writing: ★★★

Milestones

	Idea	Coding	Writing
Feb 1st	Fix problem definition & visual designs	Prototype system framework & back-end RNN and data module	Review Related Works
Feb 15th	Refine visual design, compare design alternative	Finish back-end module & Finish Summary View & Prototype Exploration View and Detail View	
Mar 1st		Refine Visualization	Intro, related works & background
Mar 15th		Finish Interactions	System Overview & Visual design
Mar 23rd		Finish experiment and cases	Case study
Mar 29th			Finish all writings