Importance of Self-Attention for Sentiment Analysis

Gaël Letarte*, Frédérik Paradis*, Philippe Giguère, François Laviolette

Department of Computer Science and Software Engineering, Big Data Research Center; Université Laval
*Authors contributed equally to this work.

Abstract

- We propose the Self-Attention Network (SANet), a flexible and interpretable architecture for text classification.
- Experiments indicate that gains obtained by self-attention is task-dependent.
- Interpretability brought forward by our architecture highlighted the importance of neighboring word interactions to extract sentiment.

Architecture

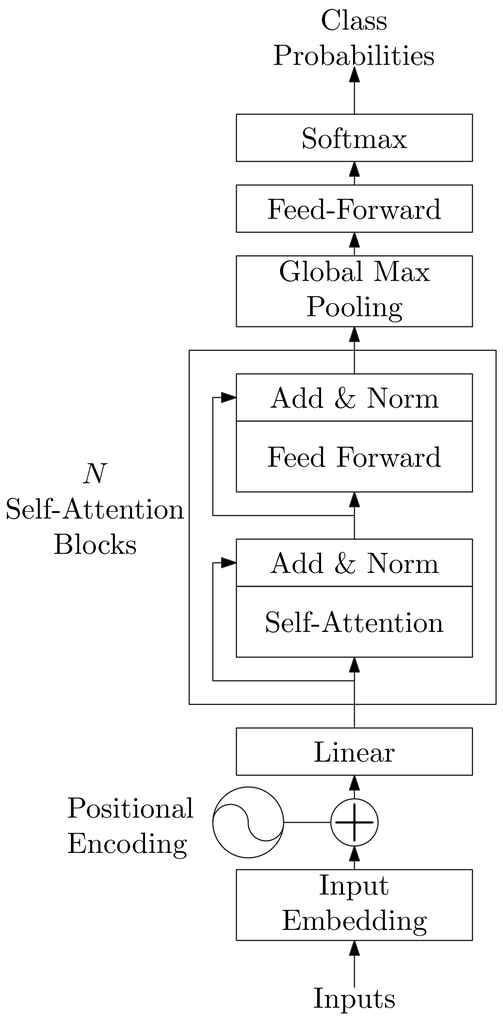


Figure 1: Our Self-Attention Network (SANet), derived from the Transformer architecture [1].

- **Self-Attention**(X)

$$= \text{Attention}(XW_Q, XW_K, XW_V)$$
$$= \text{softmax}\left(XW_{QK}X^T\right)XW_V$$

- **Positional encoding**:

$$\text{PE}_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$
$$\text{PE}_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

- No recurrent or convolutional layers.
- **Length-agnostic** contrary to some approaches based on CNN, where sequences are truncated or padded.
- **Global max pooling** yields a fixed-size representation of the sequence.

Model Configurations

Model	N	Hidden Size	Embedding
Base	1	128	100
Big	2	256	200

References

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6000–6010.

[2] Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification". In: *Advances in neural information processing systems*. 2015, pp. 649–657.

Acknowledgements

NSERC CRSNG

Groupe de Recherche en Apprentissage Automatique de Laval

Contact: {gael.letarte, frederik.paradis}.1@ulaval.ca

Data

- Seven large scale text classification datasets [2] grouped in two tasks: **Topic Classification** (TC) and **Sentiment Analysis** (SA).

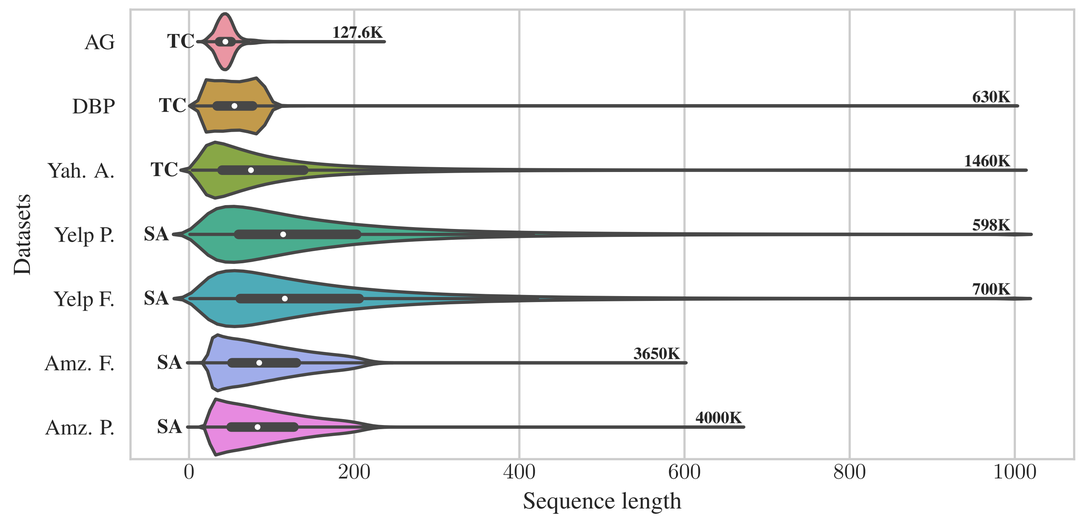


Figure 2: Visualization of sequences length distributions.

Results

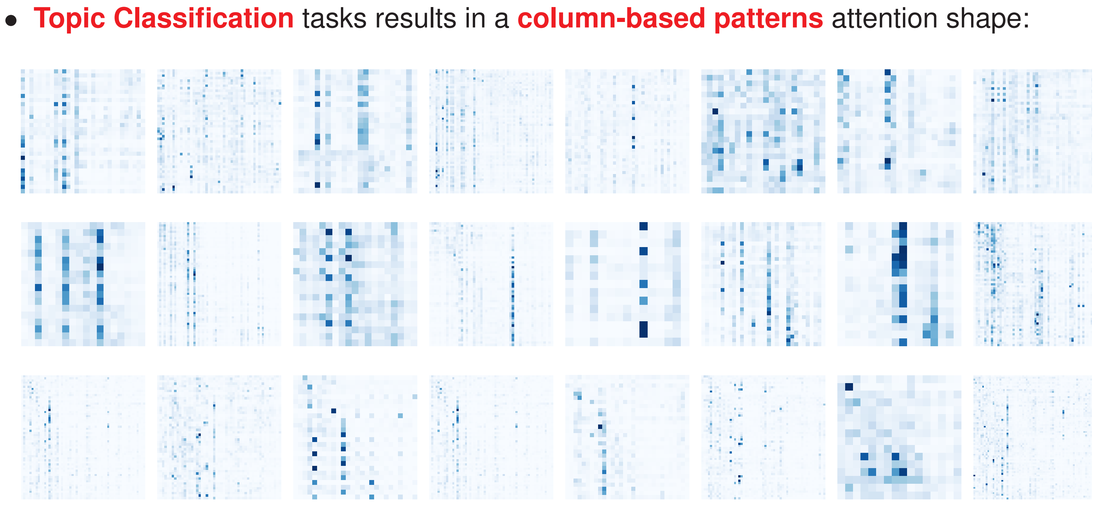
- Increase in **depth** and **representation size** in the big model is beneficial, compared to the simpler base model.
- Sentiment analysis tasks show an improvement of around 2% when using self-attention compared to a baseline without attention, while topic classification shows no gain.

Table 1: Test error rates (%) for text classification. In **bold**, our best model and stars (*) indicate attention mechanisms.

Model	Topic Classification			Sentiment Analysis			
	AG	DBP.	Yah. A.	Yelp P.	Yelp F.	Amz. F.	Amz. P.
Baseline (base model)	7.34	1.30	26.87	6.39	39.98	41.80	6.38
SANet* (base model)	7.86	1.27	26.99	6.26	38.16	40.08	5.55
Baseline (big)	**7.20**	**1.25**	25.90	6.42	38.92	40.58	5.82
SANet* (big)	7.42	1.28	**25.88**	**4.77**	**36.03**	**38.67**	**4.52**

Attention Behavior

- **Topic Classification** tasks results in a **column-based patterns** attention shape:



- **Sentiment Analysis** tasks results in a **diagonal band matrix** attention shape:

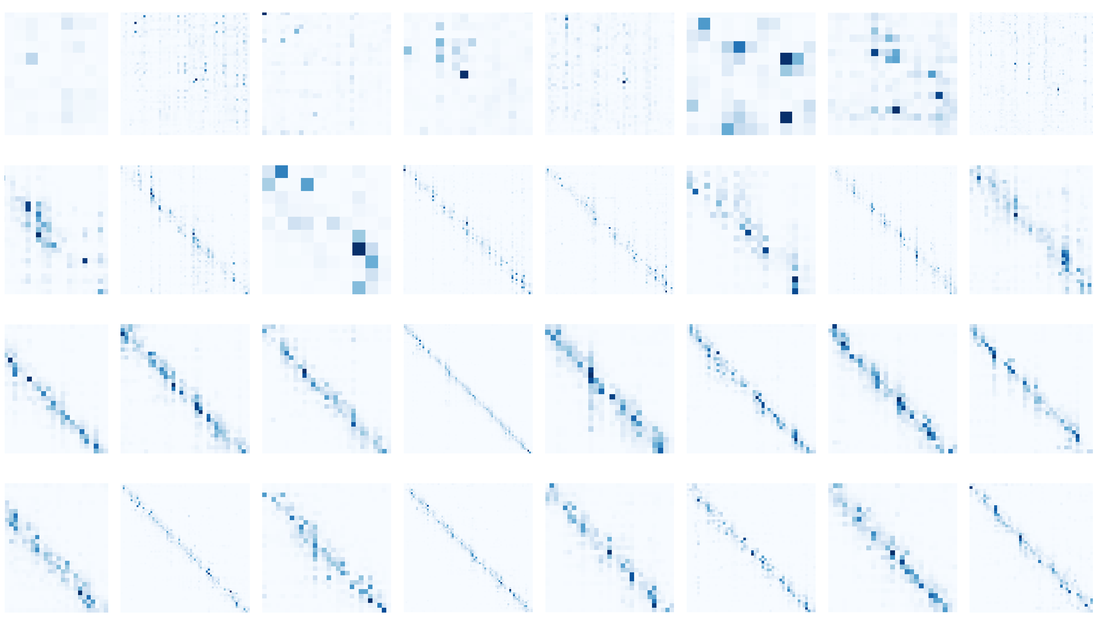


Figure 3: Randomly selected attention matrices for topic classification and sentiment analysis tasks. Each row corresponds to a different dataset (AG, DB, YA, YP, YF, AF, AP).

Quantitative Analysis

Table 2: Quantitative statistics of the self-attention mechanism behavior for the two text classification tasks.

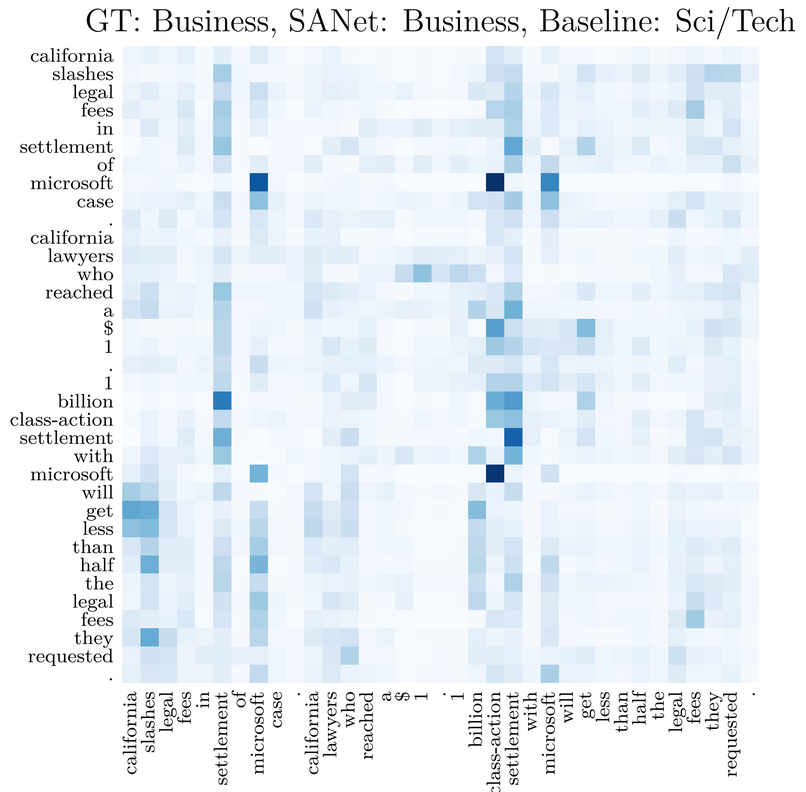
Metric	Topic Classification			Sentiment Analysis			
	AG	DBP.	Yah. A.	Yelp P.	Yelp F.	Amz. F.	Amz. P.
Gini coef.	55.31	67.94	67.45	65.16	84.18	89.50	87.76
Diag. (b=1)	7.44	8.49	6.34	5.02	23.54	41.77	40.01
Diag. (b=2)	11.86	13.80	9.83	7.89	36.89	62.35	60.34
Diag. (b=3)	16.21	18.88	13.28	10.62	45.49	73.53	71.43
Diag. (b=4)	20.42	23.74	16.59	13.19	50.90	79.49	77.21
Diag. (b=5)	24.48	28.25	19.65	15.62	54.54	83.09	80.56

- **Gini coefficient** measures the inequality in the attention weights distribution.
- **Diagonality** computes the proportion of attention weights which occur inside the band diagonal of a given bandwidth b.
- **Both** metrics results support our qualitative observations and strengthen the difference in attention behavior.

Attention Interpretability

- Attention on **Topic Classification** tasks looks for **presence** of interactions between important **concepts**, without considering relative distance, similarly to a **bag-of-word** approach.

GT: Business, SANet: Business, Baseline: Sci/Tech



- Attention on **Sentiment Analysis** tasks has strong focus on **neighboring relation**, with an interest concentrated around the diagonal which essentially consists of **skip-bigram** features with relatively small gaps.

GT: 4, SANet: 4, Baseline: 4

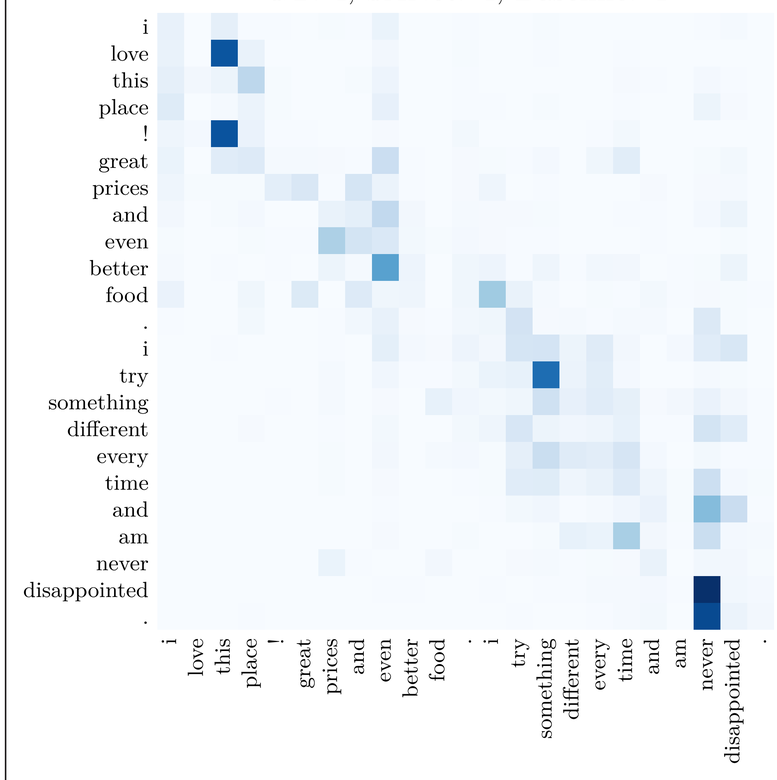


Figure 4: Self-attention different behavior for each text classification task.

Conclusion

- Interpretability through attention visualization allowed us to *discover* and *understand* the model's task-dependent behavior.
- Insights on the importance of modeling interaction between neighboring words in order to accurately extract sentiment.
- Possibility to use the global max pooling layer as a complementary tool for interpretability similarly to Class Activation Mapping (CAM).