Paper Reading: Area Attention

Yang Li, Lukasz Kaiser, Samy Bengio, Si Si

Google Research

Motivation

- Existing attention mechanisms, are mostly item-based in that a model is designed to attend to a single item in a collection of items (the memory).
- An area in the memory that may contain multiple items can be worth attending to.
- Area attention: a way to attend to an area of the memory.

Area-Based Attention Mechanisms

- 1-dimensional case
- 2-dimensional case

1-dimensional case

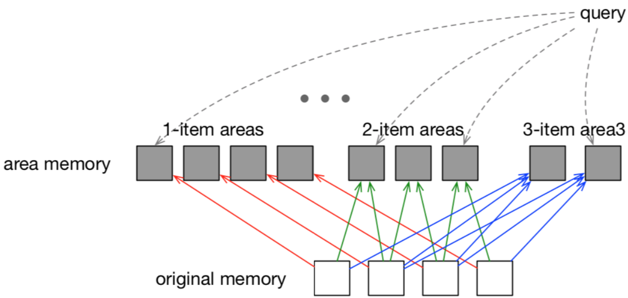

Figure 1: An illustration of area attention for the 1-dimensional case. In this example, the memory is a 4-item sequence and the maximum size of an area allowed is 3.

The number of areas: $|R| = (L - S)S + (S + 1)S/2$. Here, S is the maximum size of an area and L is the length of the sequence.

2-dimensional case

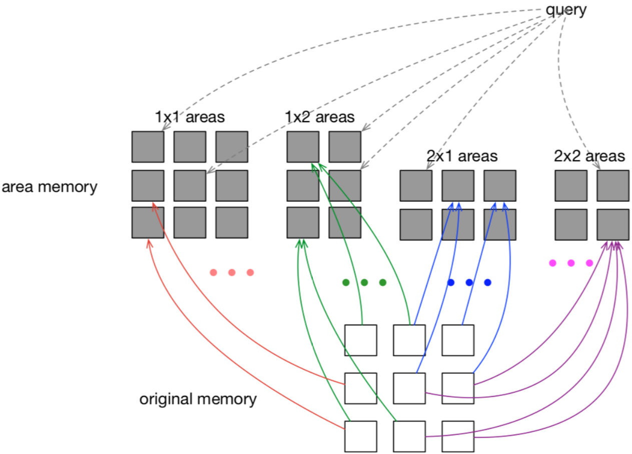

Figure 2: An illustration of area attention for the 2-dimensional case. In this example, the memory is a 3x3 grid and the dimension allowed for an area is 2x2.

Define Key and Value for each area

Mean as Key:

$$\mu_i = \frac{1}{|r_i|} \sum_{j=1}^{|r_i|} k_{i,j}$$

Sum as Value:

$$v_i^{r_i} = \sum_{j=1}^{|r_i|} v_{i,j}$$

Where $|r_i|$ is the size of the area r_i.

Combining Area Features

Standard deviation:

$$\sigma_i = \sqrt{\frac{1}{|r_i|}\sum_{l=1}^{|r_i|}(k_{i,l} - \mu_i)^2}$$

Height and width of each area:

$$e_i^h = 1(h_i)E^h,\ e_i^w = 1(w_i)E^w,\ e_i = [e_i^h, e_i^w]\ .$$

Combination:

$$k_i^r = \phi(\mu_i W_\mu + \sigma_i W_\sigma + e_i W_e)W_d \tag{9}$$

where ϕ is a nonlinear transformation such as ReLU, and $W_\mu \in \mathbb{R}^{D\times D}$, $W_\sigma \in \mathbb{R}^{D\times D}$, $W_e \in \mathbb{R}^{2S\times D}$ and $W_d \in \mathbb{R}^{D\times D}$. W_μ, W_σ, W_e and W_d are trainable parameters.

Experiments on Machine Translation

Set up:

Configuration	#Hidden Layers	Hidden Size	Filter Size	#Attention Heads
Tiny	2	128	512	4
Small	2	256	1024	4
Base	6	512	2048	8

Character-level translation tasks:

Table 1: The BLEU scores on character-level translation tasks for the Transformer-based architecture with varying model capacities.

Model Configuration	Regular Attention		Area Attention (Eq.3 and 4)	
	EN-DE	EN-FR	EN-DE	EN-FR
Tiny	6.97	9.47	**7.39**	**11.79**
Small	12.18	18.75	**13.44**	**21.24**
Base	24.65	32.80	**25.03**	**33.69**

Token-level translation tasks:

Table 2: The BLEU scores on token-level translation tasks for the variations of the Transformer-based architecture.

Model Configuration	Regular Attention		Area Attention (Eq.3 and 4)	
	EN-DE	EN-FR	EN-DE	EN-FR
Tiny	18.60	27.07	**18.80**	**27.29**
Small	**22.80**	31.91	**22.80**	**32.28**
Base	27.96	39.10	**28.17**	**39.22**

Experiments on Image caption

Table 5: Test accuracy of image captioning models that are trained on COCO and tested on Flickr. See the previous results of the benchmark model at the row "T2T8x8 COCO" in Table 7 of (Sharma et al., 2018).

Self & Enc-Dec Attention on Image	Self-Attention on Caption	ROUGE-L	CIDEr
Regular	Regular	0.409	0.355
2×2 Eq. 3	Regular	0.410	0.359
$3 \times 3^*$ Eq. 9	Regular	0.419	**0.367**
$3 \times 3^*$ Eq. 9	2^* Eq. 9	**0.421**	0.365

Reviewer Comments

- Some important related studies are missing.
- The experimental setting for NMT looks unnormal: Character-level translation
- Motivation: Why we need to attend multiple (adjacent) items to boost the performance?
- Influence on different size of area?

Area Attention and Phrase-Based Attention

- Both: attend n grams span.
- Both: do not have solid experimental results.
- ...

- Area: 1-D, 2-D; Phrase-based: 1-D
- Area: mean as key, sum as value; Phrased-based: convolution
- ...

Open question: How to evaluate the performance of phrase attention?

Conclusion

- A new way for calculating attention by attending to whole areas.
- Interpretability: For example, whether the relative improvement from phrasal attention grows/shrinks as a function of the encoder's depth?