Area Attention

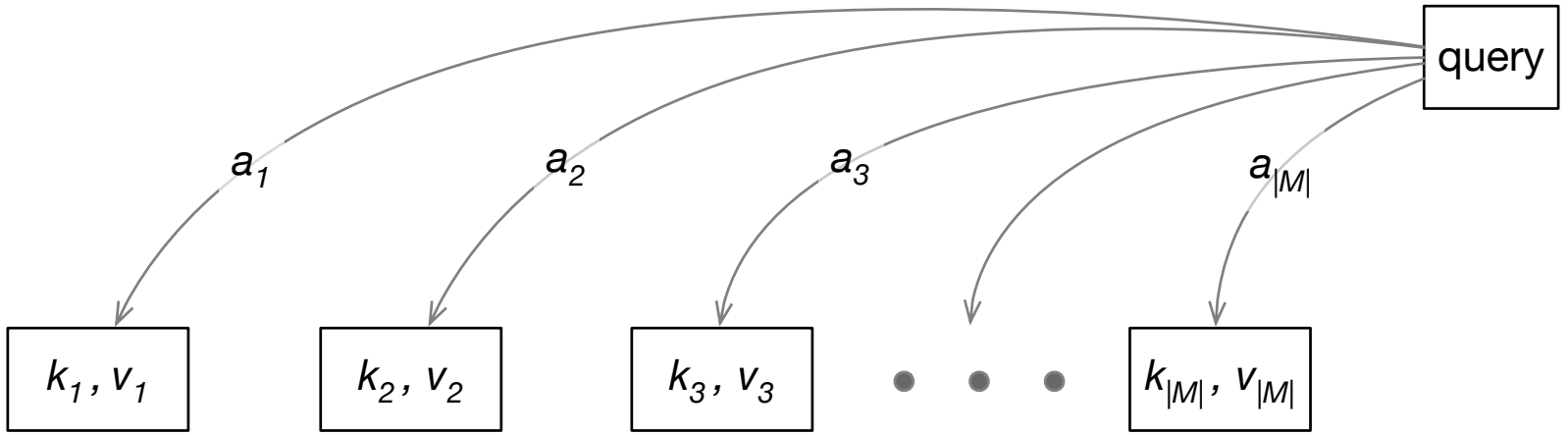
Yang Li, Lukasz Kaiser, Samy Bengio, Si Si

Google Research

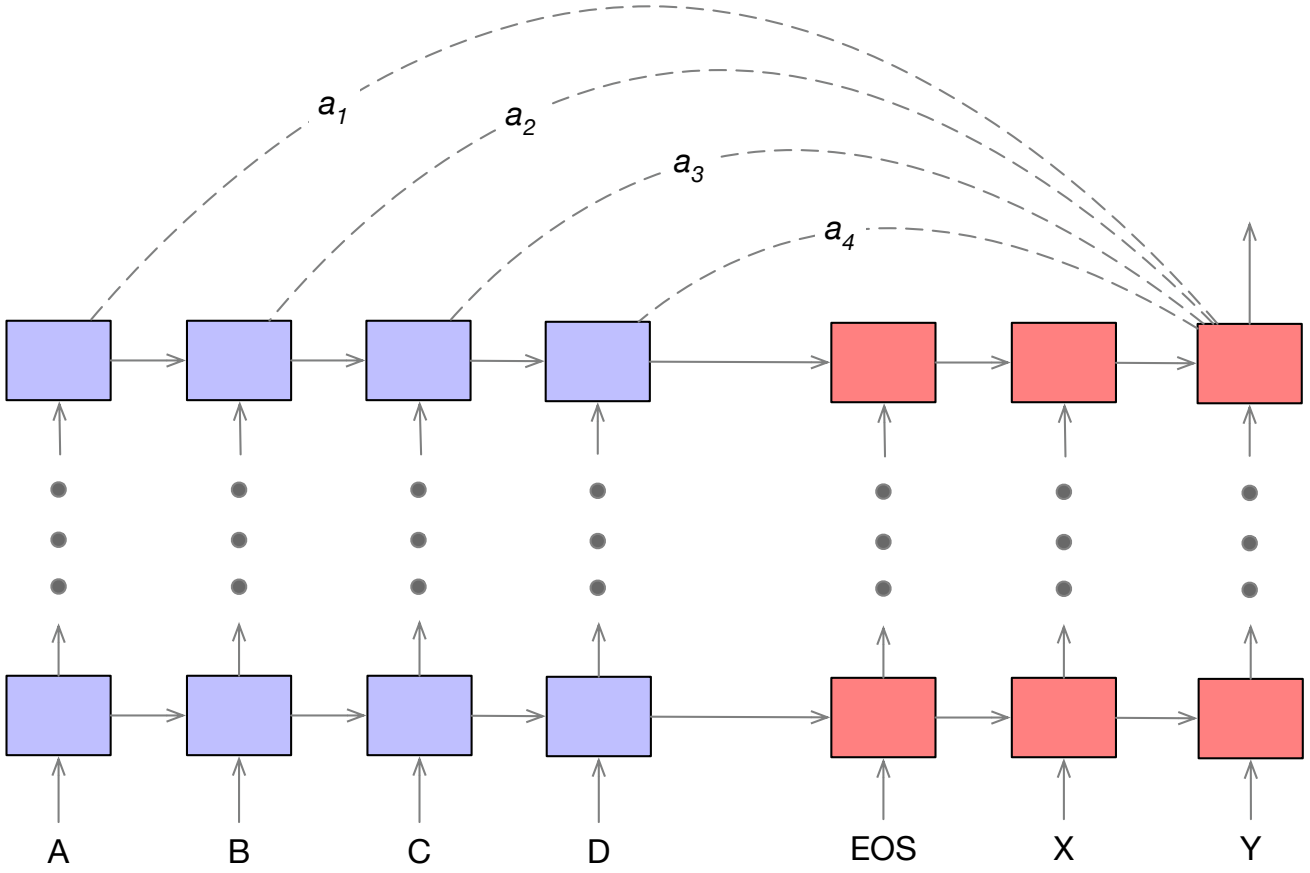
Neural Attentional Mechanisms

$$a_i = \frac{\exp(f_{att}(q, k_i))}{\sum_{j=1}^{|M|} \exp(f_{att}(q, k_j))}$$

$$O_q^M = \sum_{i=1}^{|M|} a_i v_i$$

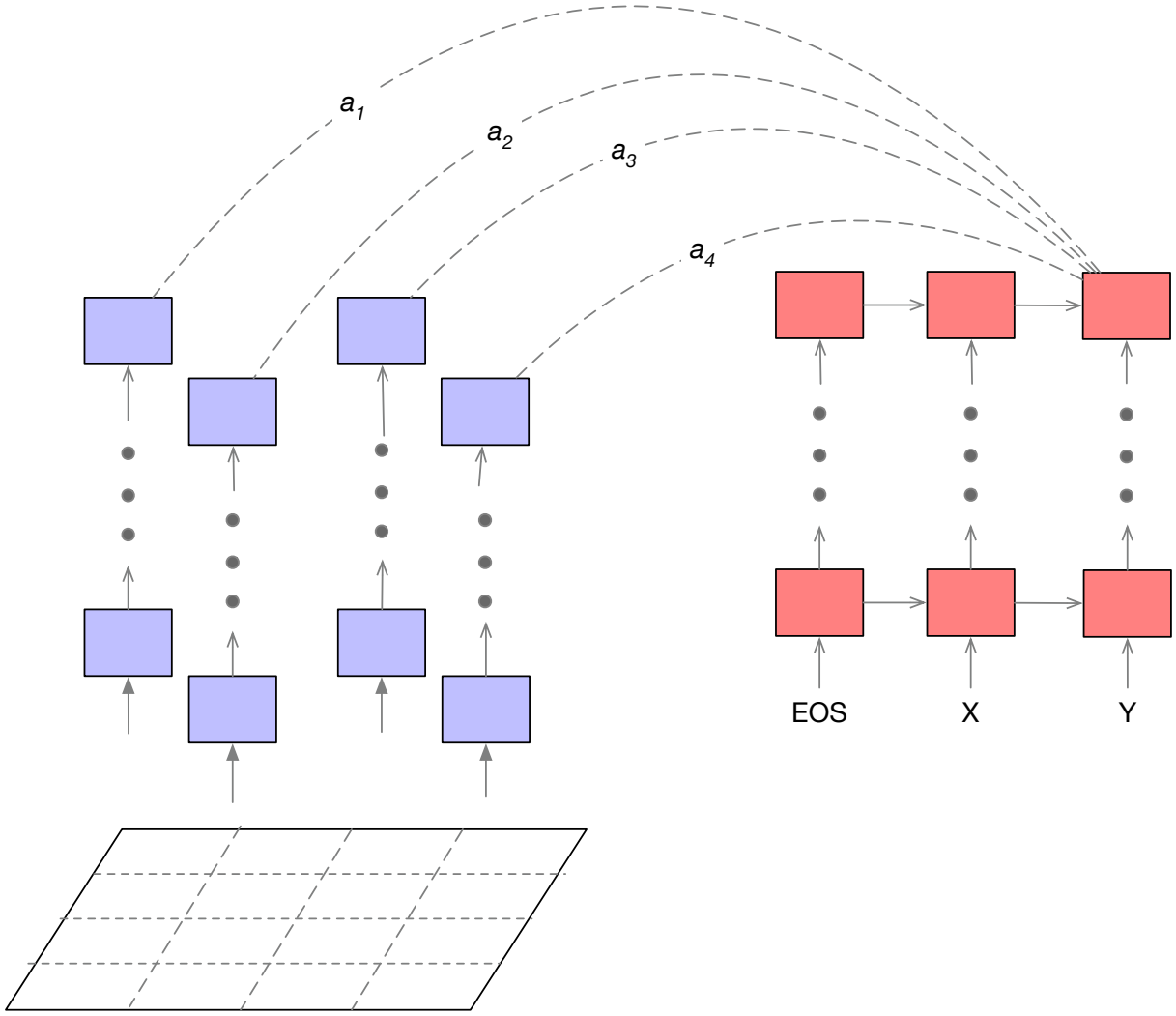


Neural Machine Translation



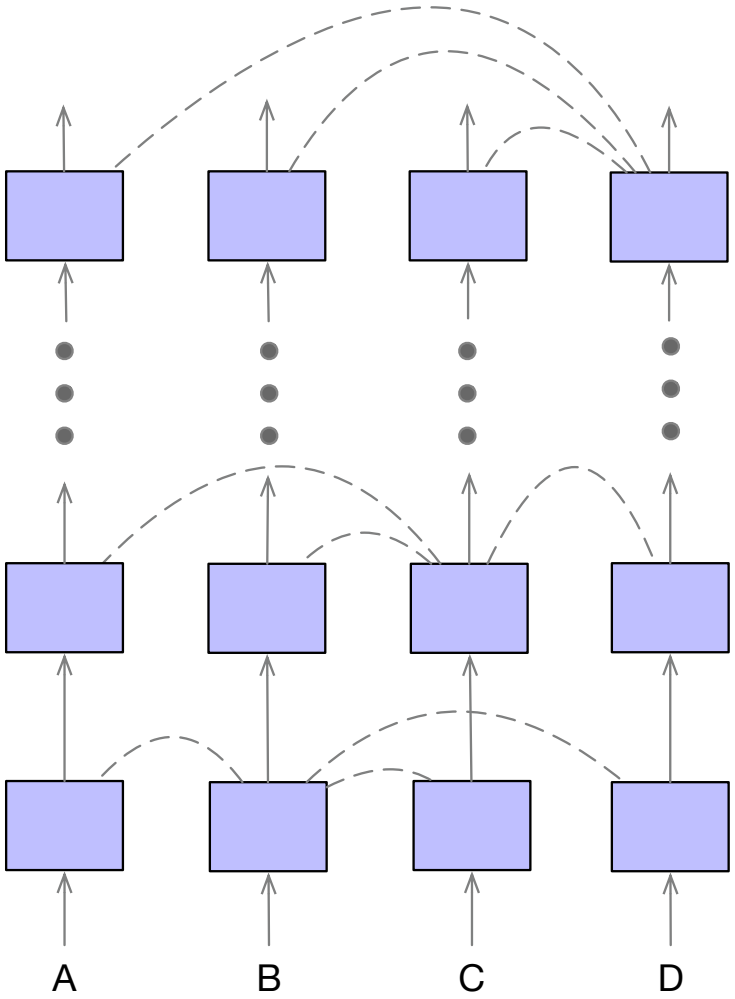
Bahdanau, Cho & Bengio, ICLR'15
Luong, Pham, & Manning, ACL'15

Image Captioning

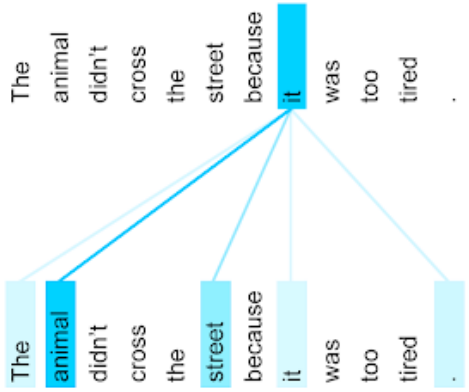


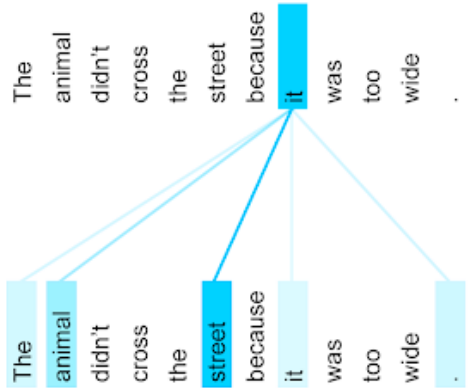
Xu, Ba, Kiros, Cho, Courville, Salakhutdinov, Zemel & Bengio, ICML'15
Sharma, Ding, Goodman & Soricut, ACL'18

Attention-Based Architectures



Transformer

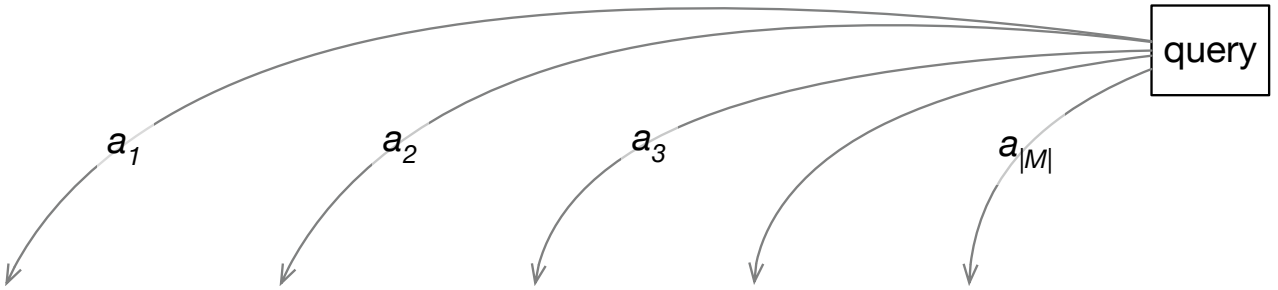




Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, NIPS'17

Limitations

The unit of attention is predetermined rather than learned.



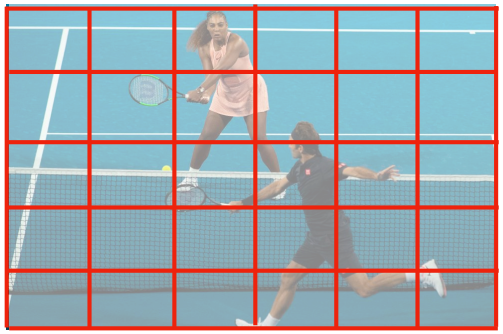
Word

Airlines	began	charging	for	the	first	and	second	checked	bags

Character

A	r	e		y	o	u		a	t		h	o	m	e	?

Image Grid Cell



Research Goal

Enable a model to attend to information at varying granularity. The unit of attention emerges from learning.

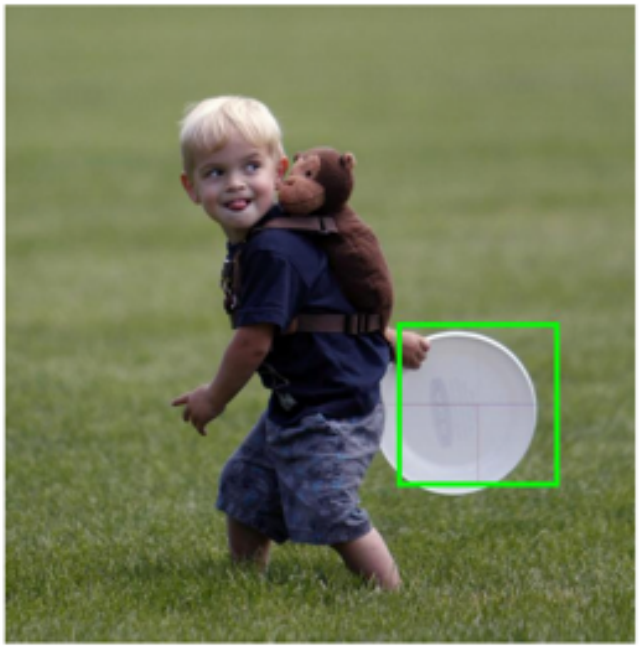
Characters ⇨ Words

A r e y o u a t h o m e ?

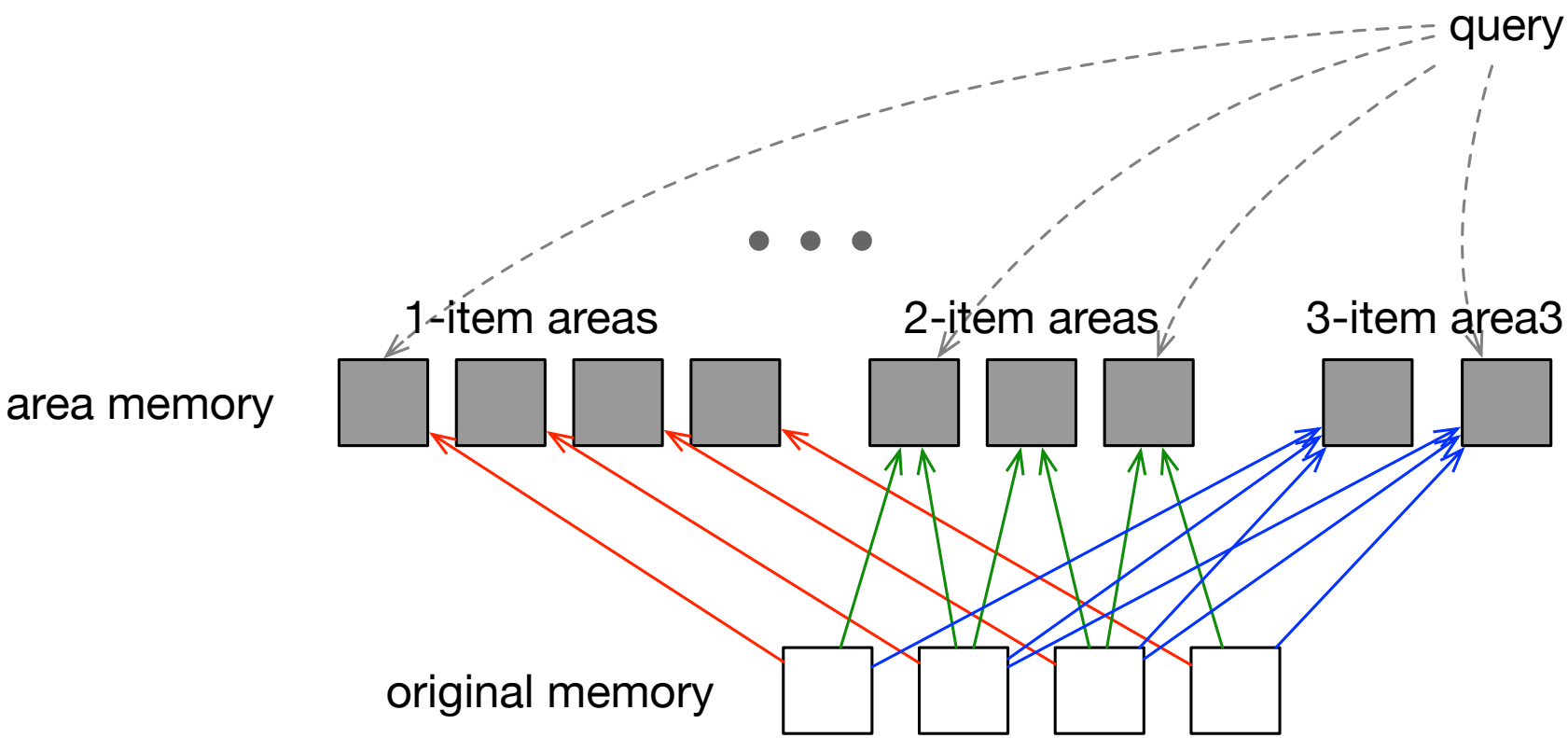
Words ⇨ Phrases

Airlines began charging for the first and second checked bags

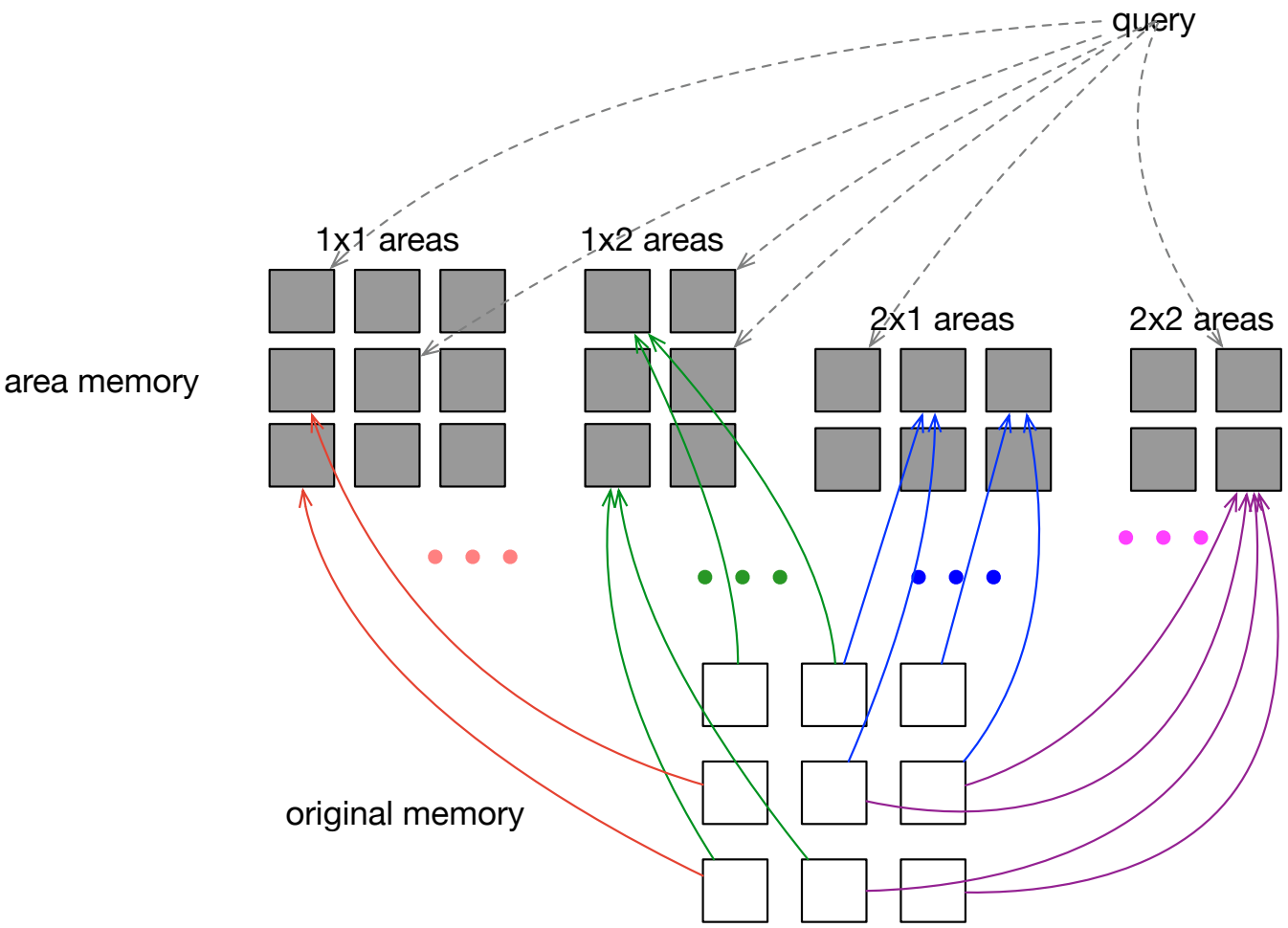
Grid cells ⇨ Objects



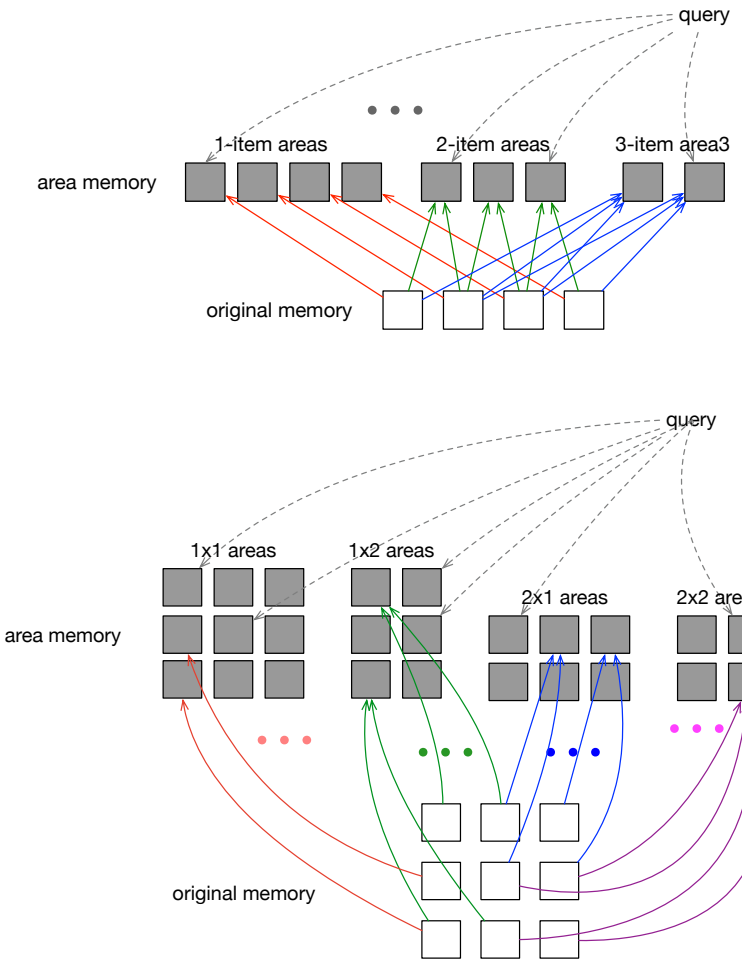
1D Area Attention



2D Area Attention



Features of Each Area



Area Features

Mean

Sum

Max

Standard deviation

Area shape, e.g., 2x2

Area Attention consistently Improves upon Transformer & LSTM

Transformer
Machine Translation

Model	Regular Attention		Area Attention (Eq.3 and 4)		Area Attention (Eq.9 and 4)	
	EN-DE	EN-FR	EN-DE	EN-FR	EN-DE	EN-FR
Tiny	18.58	27.03	19.13*	27.4*	**19.27***	**27.91****
Small	22.55	31.93	22.84	32.31*	**23.2****	**32.93****
Base	28.16	38.97	28.47	**39.27***	**28.52***	39.19
Big	29.26	41.0	29.49	41.18	**29.77***	**41.46***

LSTM
Machine Translation

#Cells	#Heads	Regular Attention		Area Attention (Eq.3,4)		Area Attention (Eq.9,4)	
		EN-DE	EN-FR	EN-DE	EN-FR	EN-DE	EN-FR
256	1	16.58	22.77	19.26*	29.35*	**19.46***	**29.79****
256	4	16.73	28.1	20.25*	**30.49***	**20.74****	30.2*
512	1	18.65	30.32	**21.82***	**32.80***	21.80*	32.73*
512	4	19.16	30.55	23.09*	33.75*	**23.41***	**34.09****
1024	1	19.4	31.99	**23.69***	34.65*	23.48*	**34.76***
1024	4	20.21	32.21	24.55*	35.95*	**24.85****	**35.97***

Transformer
Image Captioning

Model	COCO40		Flickr 1K	
	CIDEr	ROUGE-L	CIDEr	ROUGE-L
Benchmark (Sharma et al., 2018)	1.032	0.700	0.359	0.416
Benchmark Replicate	1.034	0.701	0.355	0.409
2 × 2 Eq.3 & 4	**1.060**	0.704	0.364	**0.420**
3 × 3 Eq.3 & 4	**1.060**	0.706	**0.377**	0.419
3 × 3 Eq.9 & 4	1.045	**0.707**	0.372	**0.420**

Area Attention

Yang Li, Lukasz Kaiser, Samy Bengio, Si Si
Google Research

Poster session
Tue Jun 11th 06:30 — 09:00 PM @ Pacific Ballroom #27

Source code
https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/layers/area_attention.py