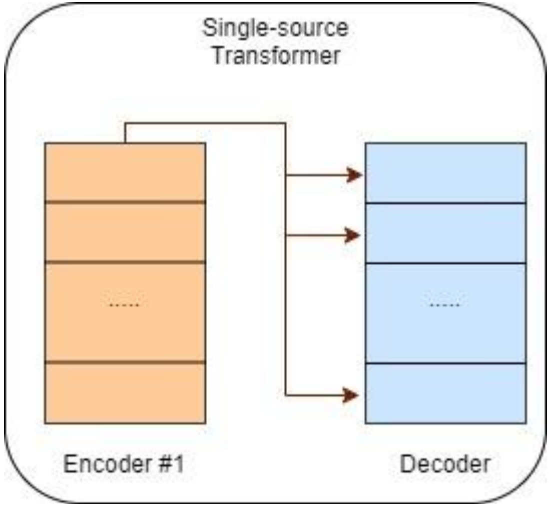
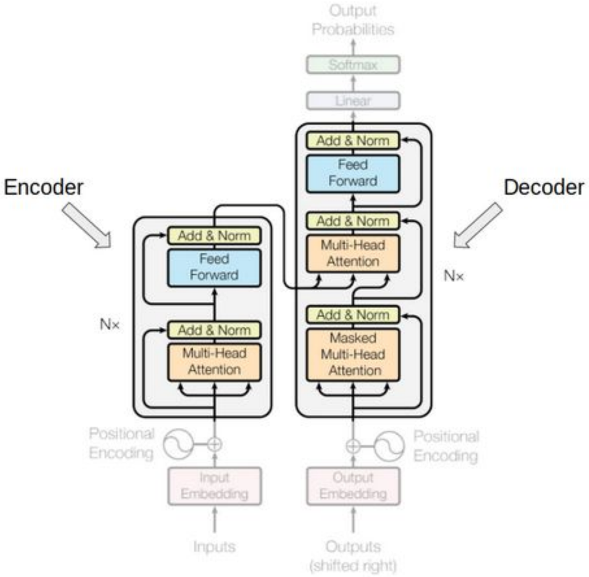
Input Combination Strategies for Multi-Source Transformer Decoder

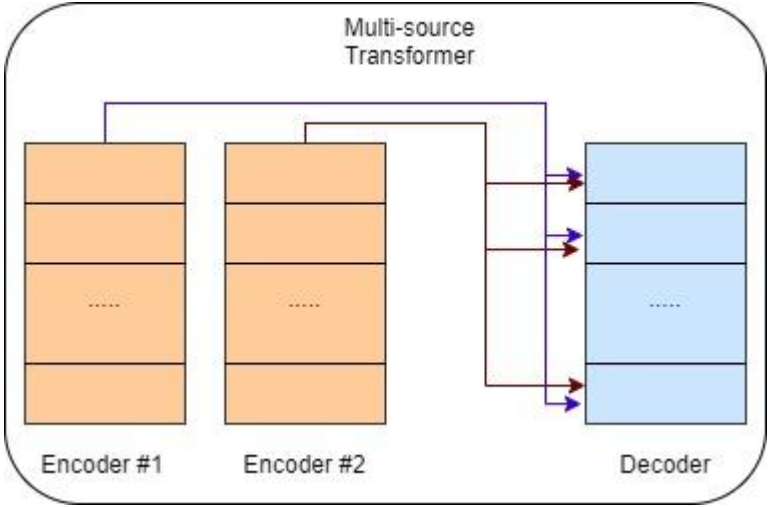
Single-Source Transformer

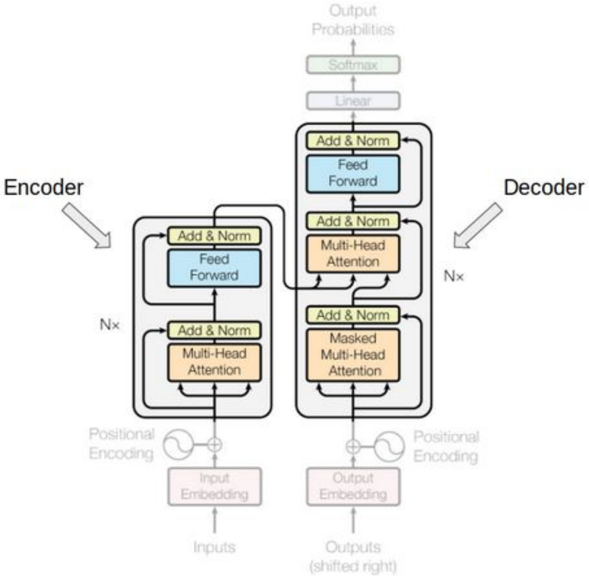




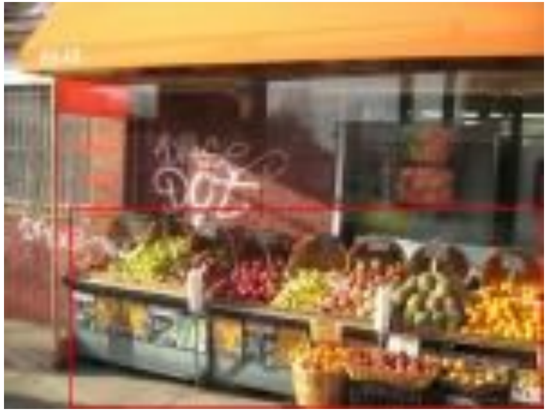
$$\mathcal{A}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Multi-Source Transformer





Multimodal translation



बाजार के बाहर फल स्टैंड

Fruit stand outside market

Multi-Source MT

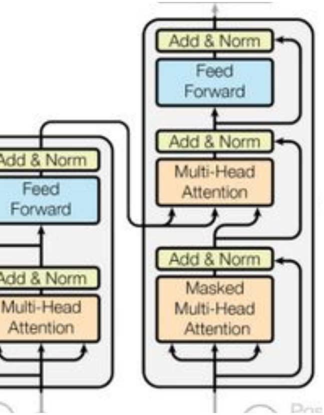
$$(s_1, s_2, \ldots, s_n) \longrightarrow t$$

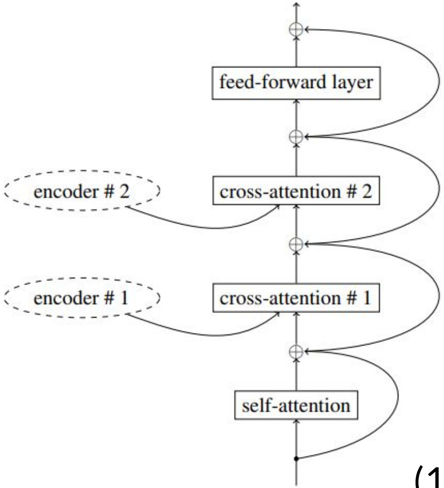
(वह खेल रहा है, he is playing) → Tō khēḷata āhē

(Hindi, English) → Marathi

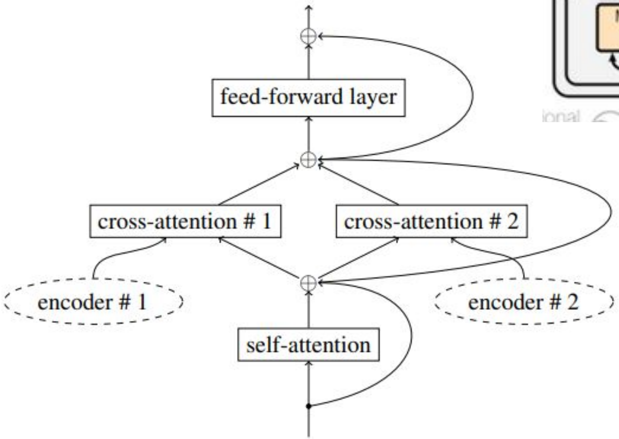
- Parallel sentences from one language s_1 to t exists. To improve the score, we can use languages related to t with which s_1 has parallel sentence corpus.

Proposed Strategies



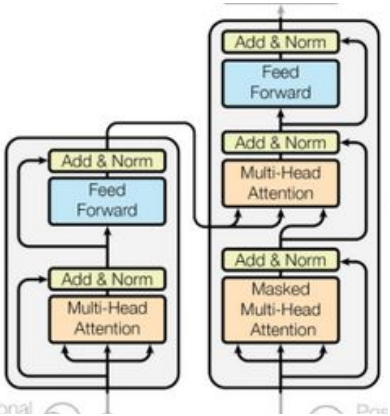


(1) Serial

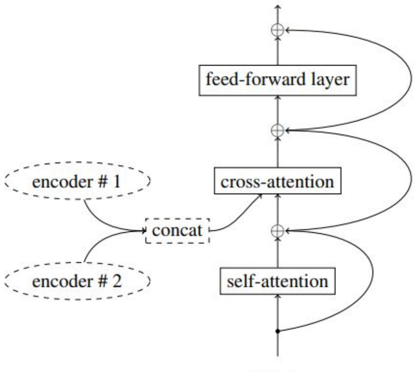


(2) Parallel

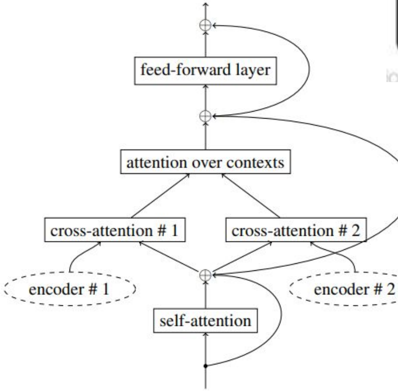
$$\mathcal{A}^h_{para}(Q, K_{1:n}, V_{1:n}) = \sum_{i=1}^{n} \mathcal{A}^h(Q, K_i, V_i)$$



Proposed Strategies



(3) Flat



(4) Hierarchical

$$K_{flat} = V_{flat} = \text{concat}_i(K_i)$$

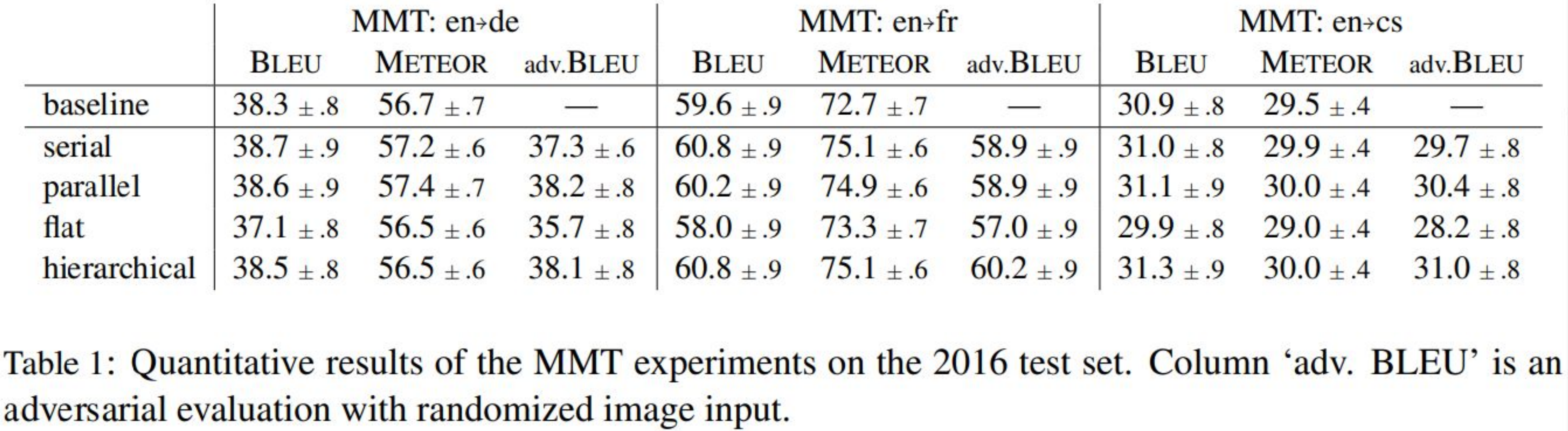
$$\mathcal{A}^h_{flat}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{flat}, V_{flat})$$

$$K_{hier} = V_{hier} = \text{concat}_i(\mathcal{A}^h(Q, K_i, V_i))$$

$$\mathcal{A}^h_{hier}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{hier}, V_{hier})$$

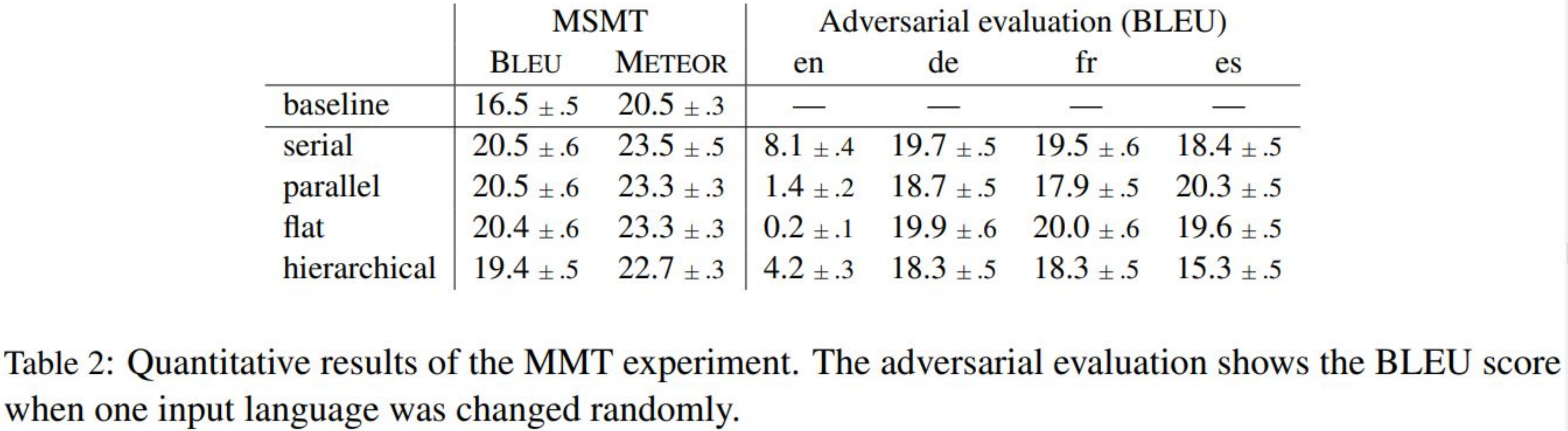
Hyper-parameters and datasets

- For Multimodal translation:
  - Multi30k dataset: contains triplets of images, English captions and their English translations into German, French and Czech. The dataset contains 29k triplets for training, 1,014 for validation and a test set of 1,000.
  - For getting image representation, linear projection into 512 dimensions on last convolutional layer of ResNet50 is applied.
  - 6 layers of encoder and decoder with 512 model dimension.
- For Multi-source MT:
  - Europarl corpus: Source languages ~ Spanish, French, German, and English; target languages ~ Czech. Dataset contains 511k 5-tuples of sentences for training, 1k for validation and another 1k for testing



	MMT: en→de			MMT: en→fr			MMT: en→cs		
	BLEU	METEOR	adv.BLEU	BLEU	METEOR	adv.BLEU	BLEU	METEOR	adv.BLEU
baseline	38.3 ± .8	56.7 ± .7	—	59.6 ± .9	72.7 ± .7	—	30.9 ± .8	29.5 ± .4	—
serial	38.7 ± .9	57.2 ± .6	37.3 ± .6	60.8 ± .9	75.1 ± .6	58.9 ± .9	31.0 ± .8	29.9 ± .4	29.7 ± .8
parallel	38.6 ± .9	57.4 ± .7	38.2 ± .8	60.2 ± .9	74.9 ± .6	58.9 ± .9	31.1 ± .9	30.0 ± .4	30.4 ± .8
flat	37.1 ± .8	56.5 ± .6	35.7 ± .8	58.0 ± .9	73.3 ± .7	57.0 ± .9	29.9 ± .8	29.0 ± .4	28.2 ± .8
hierarchical	38.5 ± .8	56.5 ± .6	38.1 ± .8	60.8 ± .9	75.1 ± .6	60.2 ± .9	31.3 ± .9	30.0 ± .4	31.0 ± .8

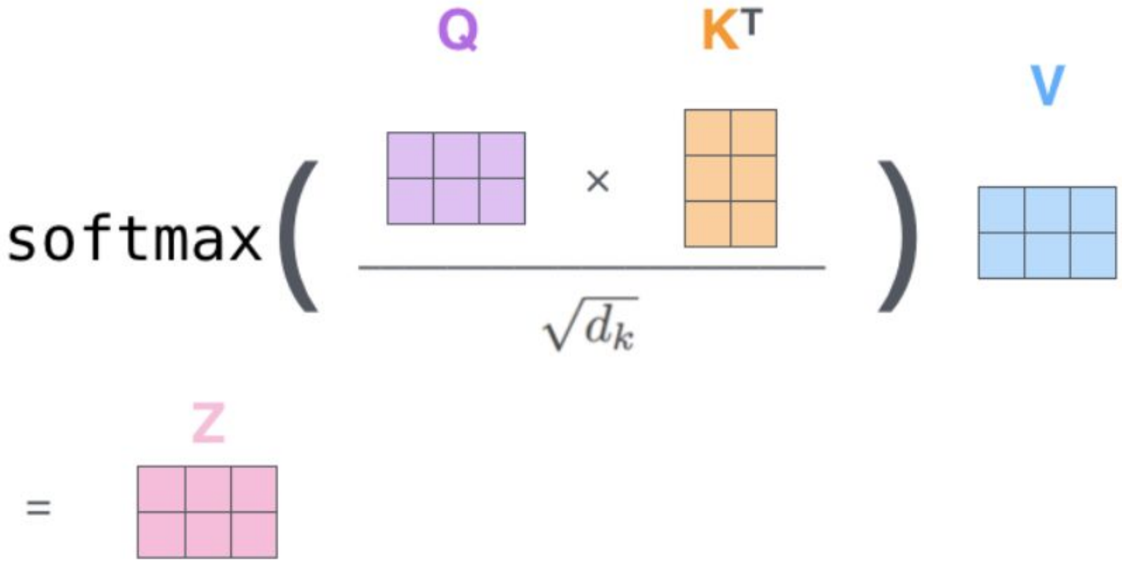
Table 1: Quantitative results of the MMT experiments on the 2016 test set. Column 'adv. BLEU' is an adversarial evaluation with randomized image input.



	MSMT		Adversarial evaluation (BLEU)			
	BLEU	METEOR	en	de	fr	es
baseline	16.5 ± .5	20.5 ± .3	—	—	—	—
serial	20.5 ± .6	23.5 ± .5	8.1 ± .4	19.7 ± .5	19.5 ± .6	18.4 ± .5
parallel	20.5 ± .6	23.3 ± .3	1.4 ± .2	18.7 ± .5	17.9 ± .5	20.3 ± .5
flat	20.4 ± .6	23.3 ± .3	0.2 ± .1	19.9 ± .6	20.0 ± .6	19.6 ± .5
hierarchical	19.4 ± .5	22.7 ± .3	4.2 ± .3	18.3 ± .5	18.3 ± .5	15.3 ± .5

Table 2: Quantitative results of the MMT experiment. The adversarial evaluation shows the BLEU score when one input language was changed randomly.

Attention


softmax
Q
×
K^T
$\sqrt{d_k}$
V
=
Z