Context-Based Network with Transformer for Image2latex

Nuo Pang, Chun Yang*, Xiaobin Zhu, Jixuan Li and Xu-Cheng Yin

image2latex

$$= 1 + Q_1 \sum_{n=-\infty}^{\infty} \frac{1}{|\vec{y} - 2\pi n a \hat{z}|^6},$$

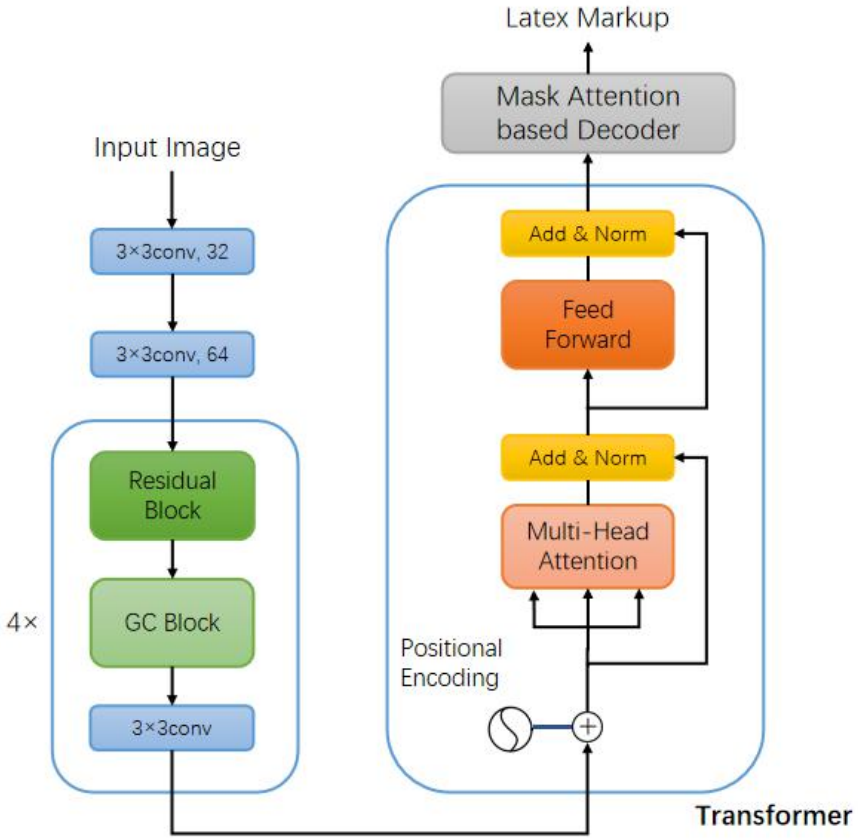
```
=1 +Q_1
\sum_{n = -\infty}^\infty \frac{1}{\mid \vec{y} -
2\pi n a \hat{z} \mid^6},
```

- *the complex two-dimensional structure,*
- *variant scales of Input*
- *very long representation sequence*

$$_p = E_{p\sigma}(x)\omega_\sigma = N(n)e^{i(p_0 x^0 + p_2 x^2 + p_3 x^3)} D_n(\rho)\omega_\sigma$$

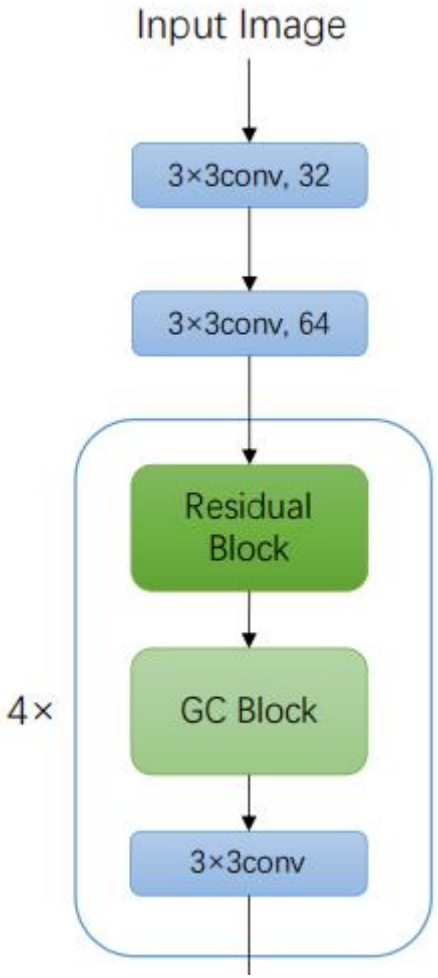
```
p=E_{p \sigma}(x) \omega_{\sigma}=N(n) e^{i\left(p_{0}
x^{0}+p_{2} x^{2}+p_{3} x^{3}\right)} D_{n}(\rho)
\omega_{\sigma}
```

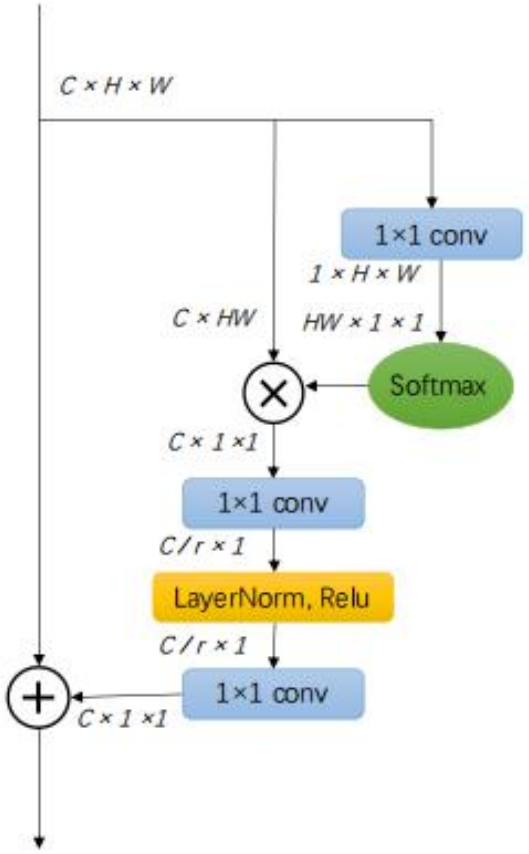
our model



- *Feature Extractor: ResNet + Global context block*
- **A transformer-based encoder is inserted between the feature extractor and decoder**
- **A mask attention-based decoder**

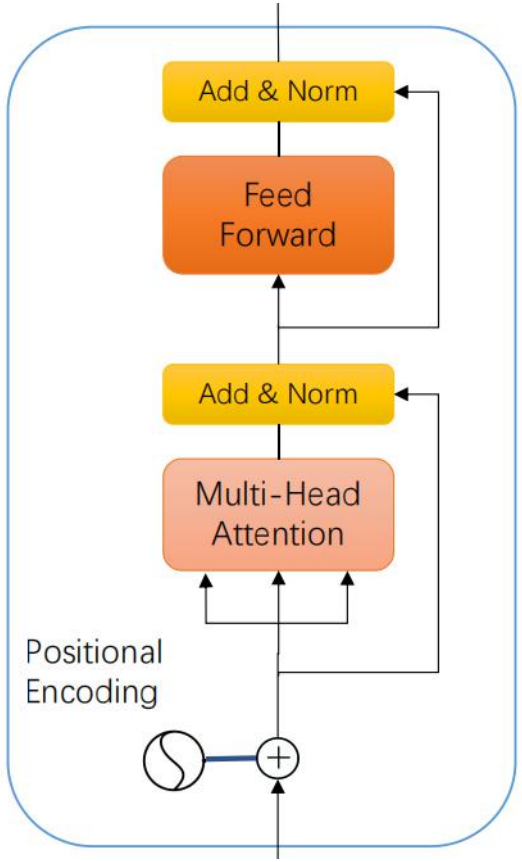
Feature Extractor





Global context block

Transformer-based Encoder



- *Positional encoding*
- **Scaled Multi-Head Dot-Product Attention**
- *Position-wise Feed-Forward Network*

Mask Attention-based Decoder

$$e_{tij} = V_a^T tanh(W_a h_{t-1} + U_a f_{ij})$$

$$a_{tij} = \frac{\exp(e_{tij})}{\sum_{k=1}^{H} \sum_{l=1}^{W} \exp(e_{tkl})}$$

$$m_{tij} = max(m_{ij} - a_{(t-1)ij}, 0)$$

$$c_t = \sum_{i=1}^{H} \sum_{j=1}^{W} a_{tij} f_{ij} \longrightarrow c_t = \sum_{i=1}^{H} \sum_{j=1}^{W} a_{tij} f_{ij} m_{ij}$$

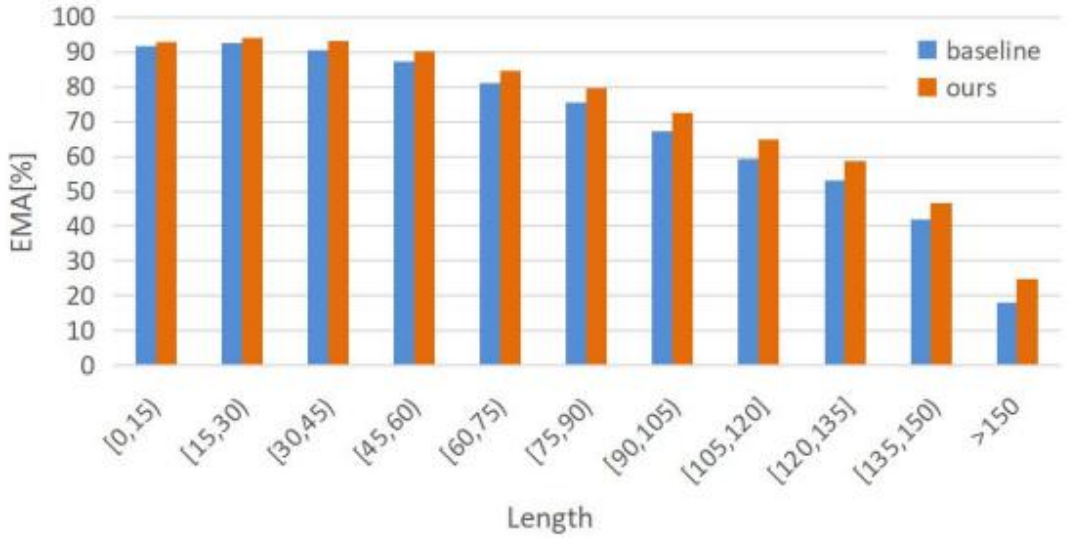
$$h_t = g(h_{t-1}, y_{t-1}, c_t)$$

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{m} \text{softmax}\left(f\left(y_{t-1}, h_t, c_t\right)\right)$$

Results

COMPARISON WITH DIFFERENT MODELS ON THE IM2LATEX-100K

MODEL	BLEU	EDA	EMA
INFTY [1]	66.65	53.82	26.66
WYGIWYS [22]	87.73	87.60	79.88
Coarse-to-Fine Attention [23]	87.07	-	78.10
Double Attention [29]	88.42	88.57	-
our model	**89.72**	**90.07**	**82.13**



ABLATION EXPERIMENTS FOR OUR THREE KEY MODULES

MODEL	BLEU	EDA	EMA
Baseline	86.93	86.48	78.37
Baseline + GC	87.71	87.43	79.24
Baseline + Transformer	88.37	88.24	79.91
Baseline + Mask Attention	87.85	87.61	79.46