DropKey for Vision Transformer

Bonan Li[1], Yinhan Hu[1], Xuecheng Nie[2], Congying Han[1], Xiangjian Jiang[3], Tiande Guo[1], Luoqi Liu[2]
University of Chinese Academy of Sciences[1] MT Lab, Meitu Inc.[2] University of Cambridge[3]

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Main contributions

In this paper, we propose to analyze and improve the dropout technique in self-attention layer, further pushing forward the frontier of ViTs for vision tasks in a general way. Specifically, we focus on three core aspects:

➢ ***What to drop in self-attention layer?***

Different from dropping attention weights as in the vanilla design, we propose to set the Key as the dropout unit, which is essential input of self-attention layer and significantly affects the output. We theoretically verify this property via implicitly introducing an adaptive smoothing coefficient for the attention operator from the perspective of gradient optimization by formulating a Lagrange function.

➢ ***How to schedule the drop ratio?***

Different from dropping attention weights as in the vanilla design, we propose to set the Key as the dropout unit, which is essential input of self-attention layer and significantly affects the output. We theoretically verify this property via implicitly introducing an adaptive smoothing coefficient for the attention operator from the perspective of gradient optimization by formulating a Lagrange function.

➢ ***Whether need to perform structured drop?***

We implement two structured versions of the dropout operation for ViTs: the block-version dropout that drops keys corresponding to contiguous patches in images or feature maps; the cross-version dropout that drops keys corresponding to patches in horizontal and vertical stripes. We conduct thorough experiments to validate their efficacy and find that the structure trick useful for CNN is not essential for ViT, due to the powerful capability of ViT to grasp contextual information in full image range.

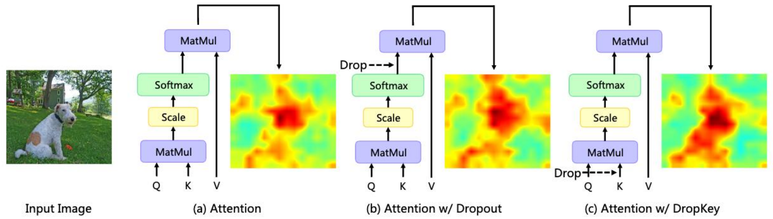

Figure 1.Comparison between the proposed DropKey and existing vanilla dropout techniques in self-attention layers for ViTs. (a) Self-attention without dropout, which suffers overfittings problem to local patch; (b) Self-attention with vanilla dropout, which regularizes the attention weights but still overfits specific patterns; (c) Self-attention with our DropKey, which overcomes prior problems and improves the model to capture vital information in a global manner.

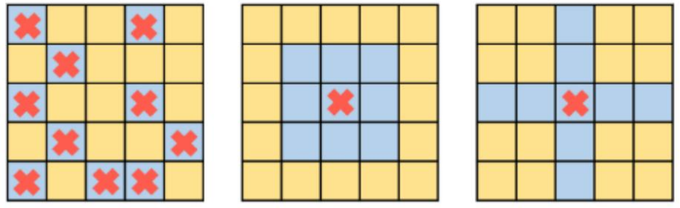

Figure 2.Mask sampling in DropKey, DropKey-Block and DropKey-Cross. The yellow patches are used as attend key to interact with a query and blue patches are dropped. Red symbol denotes the valid seed and the window size of DropKey-Block and DropKey-Cross is 3 and 1, respectively.

Algorithm 1 Attention with DropKey code

```
# N: token number, D: token dim
# Q: query (N, D), K: key (N, D), V: value (N, D)
# use_DropKey: whether use DropKey
# mask_ratio: ratio to mask

def Attention(Q, K, V, use_DropKey, mask_ratio)
   attn = (Q * (Q.shape[1] ** -0.5)) @ K.transpose(-2, -1)

   # use DropKey as regularizer
   if use_DropKey == True:
      m_r = torch.ones_like(attn) * mask_ratio
      attn = attn + torch.bernoulli(m_r) * -1e12

   attn = attn.softmax(dim=-1)
   x = attn @ V
   return x
```

Experiments

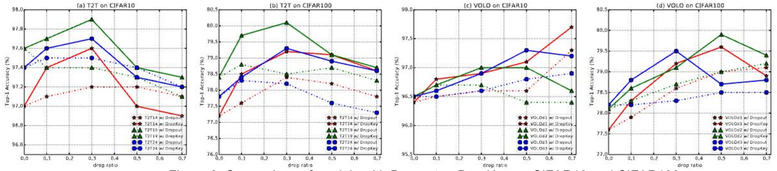

Figure 3. Comparison of models with Dropout or DropKey on CIFAR10 and CIFAR100.

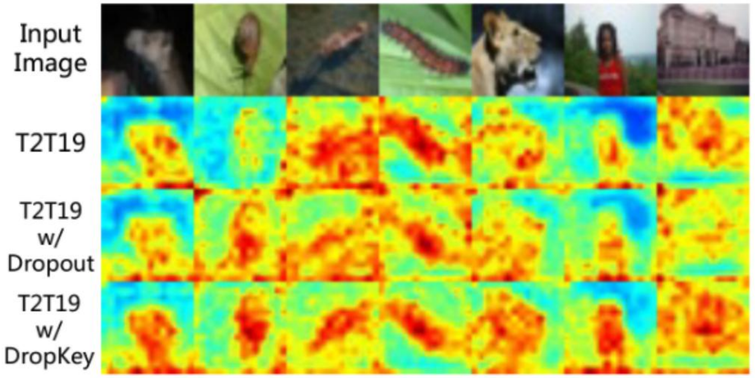

Figure 4. Visualization on CIFAR100.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR	41.9	62.1	44.2	20.6	45.7	60.7
DETR + Dropout	42.2	62.3	44.3	20.8	45.9	61.1
DETR+ DropKey	**42.9**	**63.4**	**44.7**	**21.1**	**46.7**	**61.8**

Figure 5. Comparison of DETR with Dropout and DropKey on COCO validation set.

Conclusion

- We propose to set Key as the drop unit, which yields a novel dropout-before-softmax scheme.
- We present a new decreasing schedule for drop ratio, which stabilizes the training phase by avoiding overfittings in low-level features and maintaining sufficient high-level features.
- We also experimentally show that structured dropout is not necessary for ViT