supplement of XYLayoutLM: Towards Layout-Aware Multimodal Networks For Visually-Rich Document Understanding

Zhangxuan Gu[1,2], Changhua Meng[2], Ke Wang[2], Jun Lan[2], Weiqiang Wang[2], Ming Gu[2], Liqing Zhang[1*]
[1]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, [2]Ant Group
zhangxgu@126.com
{changhua.mch,kaywang.wk,yelan.lj,weiqiang.wwq,guming.mg}@antgroup.com
zhang-lq@cs.sjtu.edu.cn

Method	Inference time/one doc	Device
LayoutReader	(10.3ms ± 34.1us)*1024	one V100
XY Cut	8.99ms ± 28.3 µs	CPU

Table 1. Inference time for detecting reading order.

1. Appendix

1.1. Pseudo-code

The pseudo-code is shown in Algorithm 1.

1.2. Inference time for LayoutReader and our method

We show the inferenece time for our XYLayoutLM and LayoutReader in Table 1. The $*1024$ means it need to decode 1024 times for one document by default.

*Corresponding author.

Algorithm 1 Augmented XY Cut Algorithm

Require: boxes: $B = \{b_i\}_{i=1}^{K}$, thresholds: $\lambda_x, \lambda_y, \theta$
Ensure: proper reading order: $O = \{s(i)\}_{i=1}^{K}$

function CUT($boxes, n, result, tmp, direction$)
  if $len(boxes) = 0$ **or** n **then**
    return $result$
  end if
  sort $boxes$ by $direction$ ▷ also **sort** tmp
  if $direction$ is Y-axis **then**
    $next \leftarrow$ X-axis
  else if $direction$ is X-axis **then**
    $next \leftarrow$ Y-axis
  end if
  $cur \leftarrow 0$
  $sets \leftarrow$ **project** $boxes$ to $direction$
  for i in range($len(boxes)$) **do**
    $set \leftarrow sets[i]$
    if $set \cap sets[i:] = \emptyset$ **then**
      $result$ += CUT($boxes[cur : i+1], i - cur, [\,], tmp[cur : i+1], next$)
      $cur \leftarrow i + 1$
    end if
  end for
  if $cur \neq i + 1$ **then**
    $result$ += $tmp[cur : i+1]$
  end if
  return $result$
end function
$tmp \leftarrow$ range(K)
for i in range($len(B)$) **do**
  $b \leftarrow B[i]$
  Random init $v_x \in N(-1, 1), v_y \in N(-1, 1)$
  if $|v_x| > \lambda_x$ **then**
    $b[0] \mathrel{+}= \theta \cdot v_x, b[2] \mathrel{+}= \theta \cdot v_x$
  end if
  if $|v_y| > \lambda_y$ **then**
    $b[1] \mathrel{+}= \theta \cdot v_y, b[3] \mathrel{+}= \theta \cdot v_y$
  end if
end for
$O \leftarrow$ CUT($B, K, [\,], tmp$,Y-axis)