ZAI 汉字检测器建模指南

字体数据源准备工作

从下列网站耐心收集需要检测的中文字体
http://www.ziticq.com/
http://www.tuyiyi.com/t-7913-1.html
http://www.cyhd.net/html/2015/fonts_0128/264.html
http://www.cyhd.net/html/2014/fonts_0815/19.html
http://www.zhaozi.cn/

中文语料数据准备工作

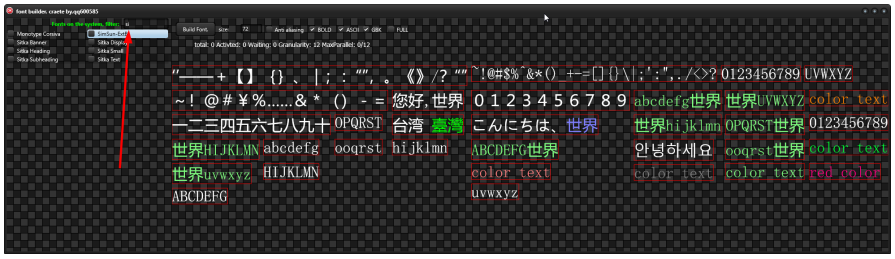
通过搜索引擎寻找“中文语料库”关键字，耐心提取
zChinese 项目内置了 600 万中文短句和词汇库通过 git 即可下载
https://github.com/PassByYou888/zChinese

硬件准备工作

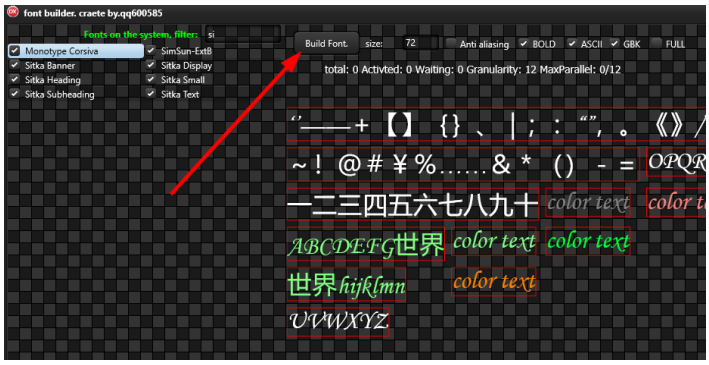
建议 CPU 朝向 intel I9 或则同等级的 AMD，8 核心以上
内存最少 64G
GPU 建议 tesla 系列 24G 显存级/titan RTX 24G，双卡 nvlink 上 p2p 配置是最佳组合

使用.zFont 字体训练自然场景文字检测器模型

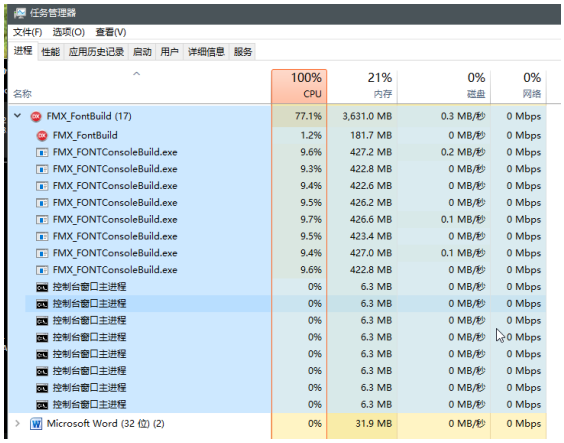
先将外部字体都准备好，安装到操作系统中
通过重启 FMX_FontBuild.exe，然后搜索安装的字体名可以确定该字体是否有效
选择字体后，会看见预览



将需要构建的目标字体勾上，点”Build font”，会启动并行化的构建程序
参数说明
Size：字符尺度参数，建议 56 以上，不要太小
Anti aliasing：反锯齿，如果勾选该参数，构建时会将原字符放大 4 倍，做高斯，再缩成 0.25 倍尺度，达到抗锯齿效果，该参数会非常消耗内存
BOLD：字符加粗
ASCII：$2E 到$7E 之间的可见字符，不包含$7F 到$FF 之间的字符
GBK：包含简繁港三种字体的中文
FULL：包含日韩含简繁港五种字体的中文
一般来说，如下图即可

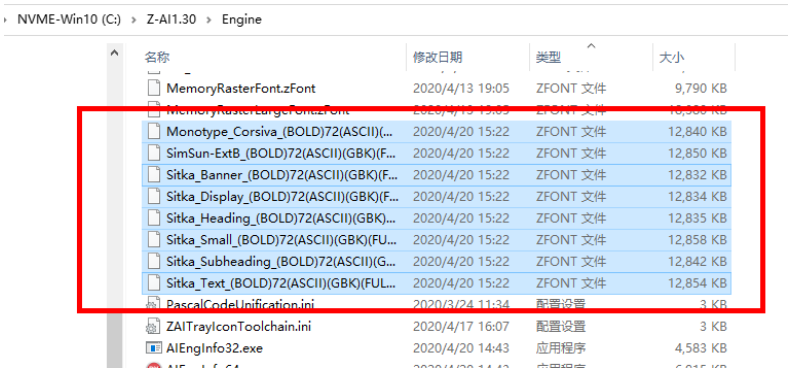


Build Font 的过程时并行化的，在生成过程会看到很多小进程

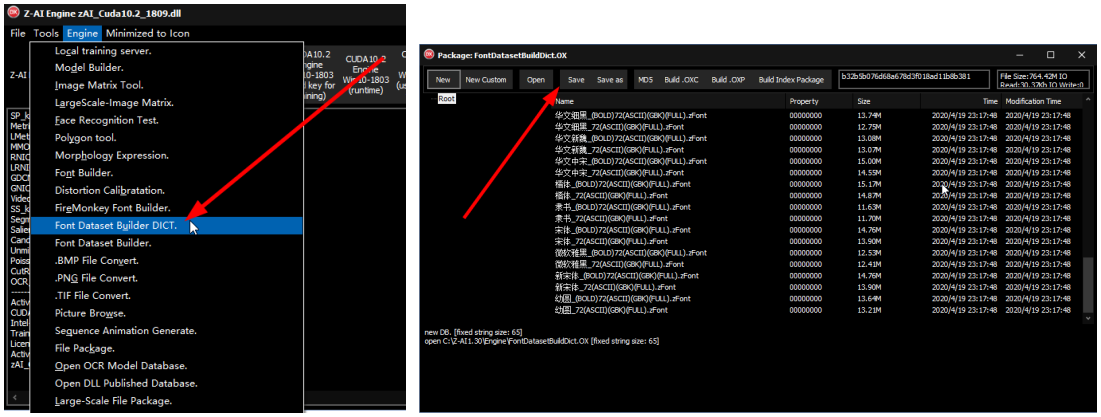


在 build Font 过程中，构建工具会看到一行线程状态提示，只要不是 0，就表示构建程序正在运行中，耐心等待一会，几分钟即可完成。

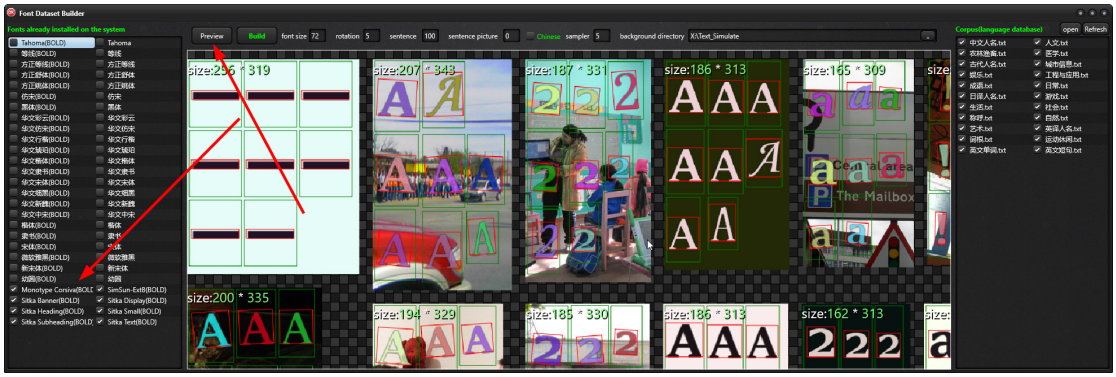
当字体构建完成后，在 FMX_FontBuild.exe 的当前目录中，可以找到我们构建的.zFont 光栅库。



打开 Font Dataset Builder 工具对应的语料库，将.zFont 导入进来，完成后 Save

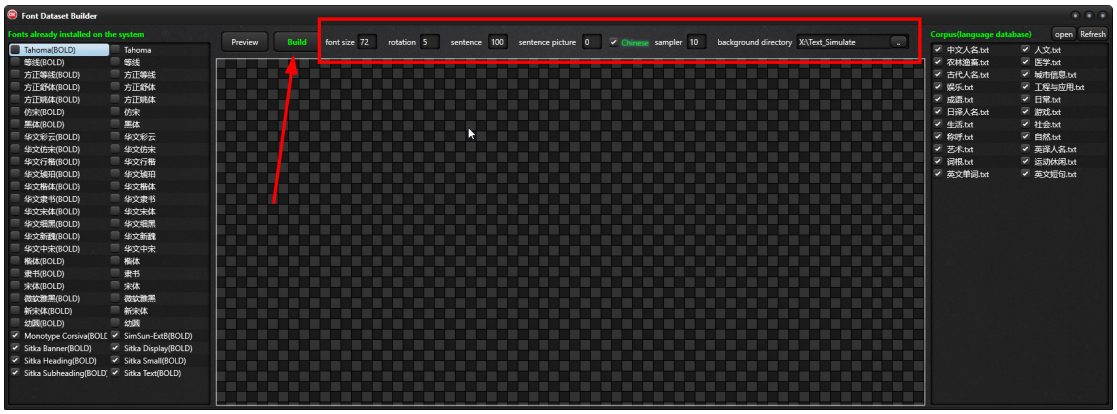


打开 Font Dataset Builder 工具，将我们刚才导入的字体勾上，点预览，确定正确吗？



如果正确，进入下一步，注意红框中的参数，确定无误以后点”Build”，开始生成过程会次序若干小时，生成的文件名有要求，需要放下以下位置，便于训练

LicensedDemo\Binary\OCR_DemoDataset.ImgMat



这时候在 LicensedDemo 中提供了以下两个训练程序，最少要求 10G 显存，如果显存大，根据备注把数据量给高，开始训练，这两个程序运训需要 30 个小时+足够的 NVME 硬盘空间，前面的步骤：如果样本库太大了，例如 100G，那么训练 60 个小时也很正常。

OCR_EndToEnd_Detector_Training_3L

OCR_EndToEnd_Detector_Training_6L

检测器出来以后，自己写程序训练分类器，分类器可以 GDCNIC,Metric,GNIC,Resnet54，或则直接使用以下程序训练分类器

OCR_EndToEnd_Metric_Training

基础不好请严格按文本指引操作，或则随时给我来信息留言。

By.qq600585

2020-4