Compiler Provenance

Jaemin Kim, SoftSec Lab KAIST
29th August 2022

Contents

- Simple compiler classifier with various features for different architectures
  - x86, arm, mips(eb)
  - Both 32 & 64-bits
- Features include:
  - Section presence & absence
  - Function prologue & epilogue
- Trained with little data, performing over 95% accuracy
  - Simple section listing works for most architectures
  - Use prologue & epilogue features otherwise

Feature 1: Section Listing

- Many configs show clear differences in ELF sections
- Removed section names containing ‘debug’

Section	clang	gcc
.bss	3760	4700
.comment	3760	4700
.data	3760	4700
.data.rel.ro	468	610
.dynamic	3760	4700
.dynstr	3760	4700
.dynsym	3760	4700
.eh_frame	3760	4700
.eh_frame_hdr	0	940
.fini	3760	4700
.fini_array	3760	4700
.gnu.hash	3760	4700
.gnu.version	3760	4700
.gnu.version_d	48	60
.gnu.version_r	3760	4700
.got	3760	4700
.got.plt	3760	4700
.hash	0	4700
.init	3760	4700
.init_array	3760	4700
.interp	3440	4300
.jcr	3760	2820
.note.ABI-tag	3440	4300
.plt	3760	4700
.rela.dyn	3760	4700
.rela.plt	3760	4700
.rodata	3760	4700
.shstrtab	3760	4700
.strtab	3760	4700
.symtab	3760	4700
.tbss	32	40
.text	3760	4700
NULL	3760	4700

normal_arm64

Section	clang	gcc
.MIPS.abiflags	3760	4700
.MIPS.stubs	3760	4700
.bss	3760	4700
.comment	3760	4700
.ctors	3760	4700
.data	3712	4640
.data.rel.ro	468	4075
.dtors	3760	4700
.dynamic	3760	4700
.dynstr	3760	4700
.dynsym	3760	4700
.eh_frame	3760	4700
.eh_frame_hdr	1596	1869
.fini	3760	4700
.fini_array	16	20
.gnu.attributes	3760	4700
.gnu.version	3760	4700
.gnu.version_d	48	60
.gnu.version_r	3760	4700
.got	3760	4700
.got.plt	3424	0
.hash	3760	4700
.init	3760	4700
.init_array	32	40
.interp	3440	4300
.jcr	3760	2820
.note.ABI-tag	3440	4300
.pdr	3760	4700
.plt	3424	0
.reginfo	3760	4700
.rel.dyn	3744	580
.rel.plt	3424	0
.rld_map	3440	4300
.rodata	3760	4700
.sbss	3024	3780
.sdata	3760	4700
.shstrtab	3760	4700
.strtab	3760	4700
.symtab	3760	4700
.tbss	32	40
.text	3760	4700
NULL	3760	4700

normal_mips32

Section	clang	gcc	icc
.bss	1446	1464	1464
.comment	1446	1464	1464
.data	1446	1464	1464
.data.rel.ro	720	726	1341
.dynamic	1446	1464	1464
.dynstr	1446	1464	1464
.dynsym	1446	1464	1464
.eh_frame	1446	1464	1464
.eh_frame_hdr	1446	1464	1464
.fini	1446	1464	1464
.fini_array	1446	1464	1464
.gnu.hash	1446	1464	1464
.gnu.version	1446	1464	1464
.gnu.version_r	1446	1464	1464
.got	1446	1464	1464
.got.plt	1446	732	1464
.hash	0	0	1464
.init	1446	1464	1464
.init_array	1446	1464	1464
.interp	1446	1464	1464
.note.ABI-tag	1446	1464	1464
.note.gnu.build-i	0	1464	1464
.plt	1446	1464	1464
.plt.got	720	732	1464
.rela.dyn	1446	1464	1464
.rela.plt	1446	1464	1464
.rodata	1446	1464	1464
.shstrtab	1446	1464	1464
.text	1446	1464	1464
.tm_clone_table	0	732	0
NULL	1446	1464	1464

binutils + coreutils, x86_64

Feature 1: Section Listing

- Train : test ratio set to 5 : 95
- Works on 7 out of 9 configs

```
Saved model at models/sections/bincore_x64_sections
5/5 [==============================] - 0s 7ms/step - loss: 0.0000e+00 - accuracy: 1.0000
232 examples in training, 4142 examples for testing.
loss: 0.0000
accuracy: 1.0000
```

```
Saved model at models/sections/normal_arm32_sections
9/9 [==============================] - 0s 8ms/step - loss: 0.0000e+00 - accuracy: 1.0000
420 examples in training, 8040 examples for testing.
loss: 0.0000
accuracy: 1.0000
```

```
Saved model at models/sections/normal_arm64_sections
9/9 [==============================] - 0s 9ms/step - loss: 0.0000e+00 - accuracy: 1.0000
433 examples in training, 8027 examples for testing.
loss: 0.0000
accuracy: 1.0000
```

Feature 1: Section Listing

```
Saved model at models/sections/normal_mips32_sections
8/8 [==============================] - 0s 12ms/step - loss: 0.0000e+00 - accuracy: 0.9678
460 examples in training, 8000 examples for testing.
loss: 0.0000
accuracy: 0.9678
```

```
Saved model at models/sections/normal_mipseb32_sections
9/9 [==============================] - 0s 11ms/step - loss: 0.0000e+00 - accuracy: 0.9690
419 examples in training, 8041 examples for testing.
loss: 0.0000
accuracy: 0.9690
```

```
Saved model at models/sections/normal_x86_32_sections
9/9 [==============================] - 0s 9ms/step - loss: 0.0000e+00 - accuracy: 1.0000
402 examples in training, 8058 examples for testing.
loss: 0.0000
accuracy: 1.0000
```

```
Saved model at models/sections/normal_x86_64_sections
9/9 [==============================] - 0s 9ms/step - loss: 0.0000e+00 - accuracy: 1.0000
410 examples in training, 8050 examples for testing.
loss: 0.0000
accuracy: 1.0000
```

Feature 2: Feature Prologue & Epilogue

- mips 64-bit archs don’t fit well with just sections

```
Saved model at models/sections/normal_mips64_sections
9/9 [==============================] - 0s 9ms/step - loss: 0.0000e+00 - accuracy: 0.6629
424 examples in training, 8036 examples for testing.
loss: 0.0000
accuracy: 0.6629
```

```
Saved model at models/sections/normal_mipseb64_sections
9/9 [==============================] - 0s 10ms/step - loss: 0.0000e+00 - accuracy: 0.6535
423 examples in training, 8037 examples for testing.
loss: 0.0000
accuracy: 0.6535
```

Feature 2: Feature Prologue & Epilogue

rkspace/dataset/normal_dataset/elf/mipseb_64

Section	clang	gcc	icc
.MIPS.abiflags	3760	4700	0
.MIPS.options	3760	4700	0
.MIPS.stubs	3760	4700	0
.bss	3760	4700	0
.comment	3760	4700	0
.ctors	3760	4700	0
.data	3712	4640	0
.data.rel.ro	2636	4075	0
.dtors	3760	4700	0
.dynamic	3760	4700	0
.dynstr	3760	4700	0
.dynsym	3760	4700	0
.eh_frame	3760	4700	0
.fini	3760	4700	0
.fini_array	16	20	0
.gnu.attributes	3760	4700	0
.gnu.version	3760	4700	0
.gnu.version_d	48	60	0
.gnu.version_r	3760	4700	0
.got	3760	4700	0
.hash	3760	4700	0
.init	3760	4700	0
.init_array	32	40	0
.interp	3440	4300	0
.jcr	3760	2820	0
.note.ABI-tag	3440	4300	0
.pdr	3760	4700	0
.rel.dyn	464	580	0
.rld_map	3440	4300	0
.rodata	3760	4700	0
.sbss	3008	3760	0
.sdata	3760	4700	0
.shstrtab	3760	4700	0
.strtab	3760	4700	0
.symtab	3760	4700	0
.tbss	32	40	0
.text	3760	4700	0
NULL	3760	4700	0

rkspace/dataset/normal_dataset/elf/mips_64

Section	clang	gcc	icc
.MIPS.abiflags	3760	4700	0
.MIPS.options	3760	4700	0
.MIPS.stubs	3760	4700	0
.bss	3760	4700	0
.comment	3760	4700	0
.ctors	3760	4700	0
.data	3712	4640	0
.data.rel.ro	2636	4075	0
.dtors	3760	4700	0
.dynamic	3760	4700	0
.dynstr	3760	4700	0
.dynsym	3760	4700	0
.eh_frame	3760	4700	0
.fini	3760	4700	0
.fini_array	16	20	0
.gnu.attributes	3760	4700	0
.gnu.version	3760	4700	0
.gnu.version_d	48	60	0
.gnu.version_r	3760	4700	0
.got	3760	4700	0
.hash	3760	4700	0
.init	3760	4700	0
.init_array	32	40	0
.interp	3440	4300	0
.jcr	3760	2820	0
.note.ABI-tag	3440	4300	0
.pdr	3760	4700	0
.rel.dyn	464	580	0
.rld_map	3440	4300	0
.rodata	3760	4700	0
.sbss	3008	3760	0
.sdata	3760	4700	0
.shstrtab	3760	4700	0
.strtab	3760	4700	0
.symtab	3760	4700	0
.tbss	32	40	0
.text	3760	4700	0
NULL	3760	4700	0

ARM 64-bit: clang vs. gcc

```
00000000004024e4 <yy_scan_bytes>:
  4024e4: a9bd57f6  stp x22, x21, [sp, #-48]!
  4024e8: a9014ff4  stp x20, x19, [sp, #16]
  4024ec: a9027bfd  stp x29, x30, [sp, #32]
```

```
00000000004099a0 <guess>:
  4099a0: a9be7bfd  stp x29, x30, [sp, #-32]!
  4099a4: 910003fd  mov x29, sp
```

```
 402548: a9427bfd  ldp x29, x30, [sp, #32]
 40254c: a9414ff4  ldp x20, x19, [sp, #16]
 402550: a8c357f6  ldp x22, x21, [sp], #48
 402554: d65f03c0  ret
```

```
 409a0c: a8c27bfd  ldp x29, x30, [sp], #32
 409a10: d65f03c0  ret
 409a14: 00000000  udf #0
```

```
Saved model at models/sections/normal_arm64_proepi
9/9 [==============================] - 0s 1ms/step - loss: 0.0000e+00 - accuracy: 0.9968
412 examples in training, 8048 examples for testing.
loss: 0.0000
accuracy: 0.9968
```

MIPS 64-bit: clang vs. gcc

```
00000001200ad158 <mips_elf_count_got_entry>:
   1200ad158:  67bdffc0    daddiu   sp,sp,-64
   1200ad15c:  ffbf0038    sd ra,56(sp)
   1200ad160:  ffbc0030    sd gp,48(sp)
   1200ad164:  ffb40028    sd s4,40(sp)
   1200ad168:  ffb30020    sd s3,32(sp)
   1200ad16c:  ffb20018    sd s2,24(sp)
   1200ad170:  ffb10010    sd s1,16(sp)
   1200ad174:  ffb00008    sd s0,8(sp)
```

```
1200ad364:  dfb10010    ld s1,16(sp)
1200ad368:  dfb20018    ld s2,24(sp)
1200ad36c:  dfb30020    ld s3,32(sp)
1200ad370:  dfb40028    ld s4,40(sp)
1200ad374:  dfbc0030    ld gp,48(sp)
1200ad378:  dfbf0038    ld ra,56(sp)
```

```
0000000120009384 <highlight_level_to_string>:
   120009384:  67bdffd0    daddiu   sp,sp,-48
   120009388:  ffbf0028    sd ra,40(sp)
   12000938c:  ffbe0020    sd s8,32(sp)
   120009390:  ffbc0018    sd gp,24(sp)
   120009394:  03a0f025    move  s8,sp
```

```
12000944c:  dfbf0028    ld ra,40(sp)
120009450:  dfbe0020    ld s8,32(sp)
120009454:  dfbc0018    ld gp,24(sp)
120009458:  67bd0030    daddiu   sp,sp,48
12000945c:  03e00008    jr ra
120009460:  00000000    nop
```

Feature 2: Feature Prologue & Epilogue

1. Check first & last n lines of functions
2. find sd / ld instructions
3. Determine whether the instructions are FILO / FIFO

Feature 2: Feature Prologue & Epilogue

label	sp_alloc_fifo	sp_alloc_filo
clang	4.0	44.0
clang	3.0	28.0
gcc	7.0	1.0
gcc	8.0	2.0
gcc	14.0	3.0
clang	2.0	34.0
clang	3.0	27.0
gcc	8.0	1.0
clang	2.0	1.0
clang	2.0	38.0
clang	2.0	49.0
gcc	4.0	1.0
gcc	7.0	3.0
gcc	23.0	0.0
clang	2.0	33.0
gcc	8.0	2.0
clang	2.0	85.0
gcc	196.0	9.0
clang	2.0	3.0
gcc	8.0	2.0
gcc	99.0	0.0
clang	2.0	60.0
gcc	9.0	1.0
gcc	83.0	0.0
clang	6.0	632.0
gcc	47.0	0.0
gcc	14.0	3.0
gcc	20.0	2.0
gcc	12.0	0.0
gcc	294.0	0.0

```
Saved model at models/sections/normal_mips64_proepi
9/9 [==============================] - 0s 2ms/step - loss: 0.0000e+00 - accuracy: 0.9793
435 examples in training, 8025 examples for testing.
loss: 0.0000
accuracy: 0.9793
```

```
Saved model at models/sections/normal_mipseb64_proepi
9/9 [==============================] - 0s 2ms/step - loss: 0.0000e+00 - accuracy: 0.9734
443 examples in training, 8017 examples for testing.
loss: 0.0000
accuracy: 0.9734
```

- Limitation
  - Confidence level is low for short programs with very few functions