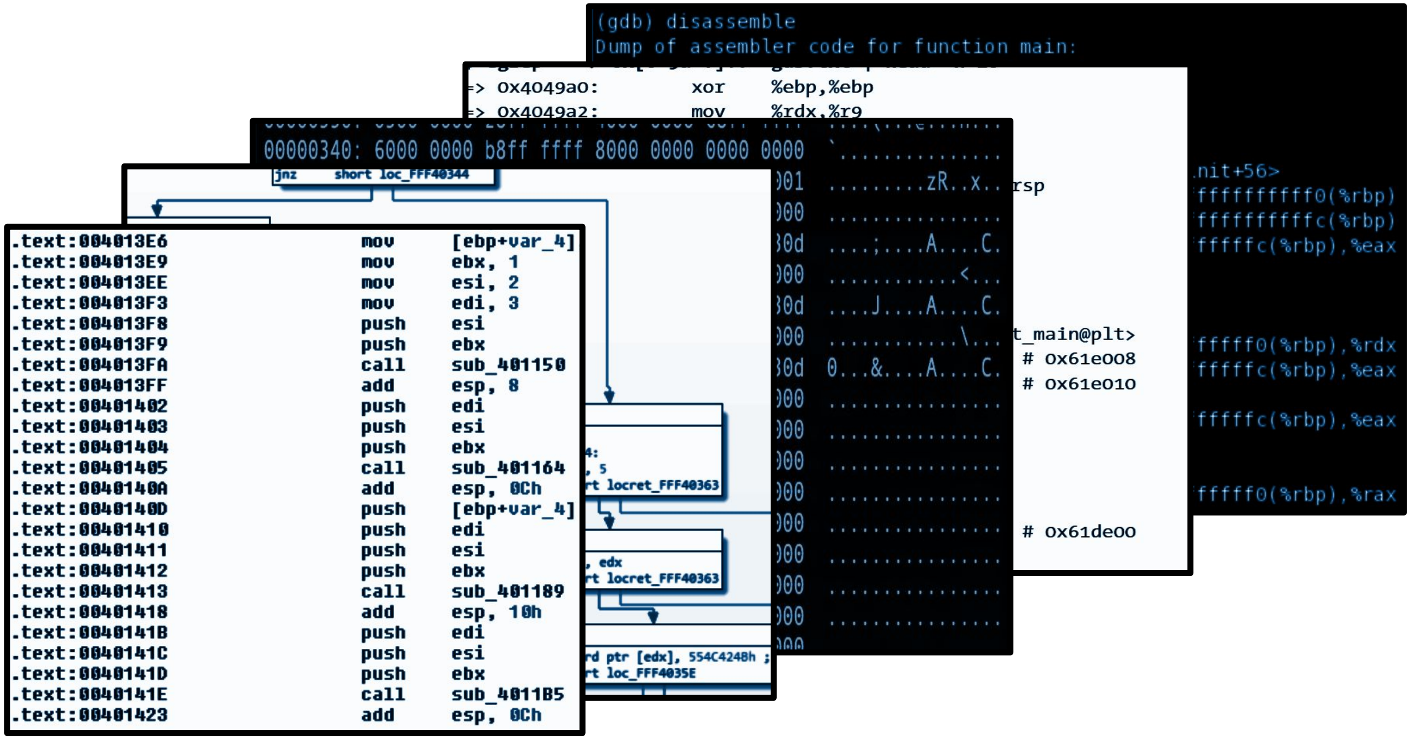
CNN-Based Malware Visualization and Explainability

Lara Dedic
Applied Machine Learning Researcher
ldedic@novetta.com
www.novetta.com

Malware Analysis

Challenges

- Complex malware is inspected manually
- Takes a long time and lots of effort for a single executable to be analyzed
- Extensive domain knowledge is required



```
(gdb) disassemble
Dump of assembler code for function main:
nit+56>
ffffffffff0(%rbp)
fffffffffc(%rbp)
fffffc(%rbp),%eax
ffffff0(%rbp),%rdx
fffffc(%rbp),%eax
fffffc(%rbp),%eax
ffffff0(%rbp),%rax
```

```
=> 0x4049a0:        xor    %ebp,%ebp
=> 0x4049a2:        mov    %rdx,%r9
rsp
t_main@plt>
 # 0x61e008
 # 0x61e010
 # 0x61de00
```

```
00000340: 6000 0000 b8ff ffff 8000 0000 0000 0000  `...............
001  .........zR..x..
000  ................
30d  ....;....A....C.
000  ............<...
30d  ....J....A....C.
000  ............\...
30d  0...&....A....C.
000  ................
```

```
jnz     short loc_FFF40344
4:
, 5
rt locret_FFF40363
, edx
rt locret_FFF40363
rd ptr [edx], 554C4248h ;
rt loc_FFF4035E
```

```
.text:004013E6                 mov     [ebp+var_4]
.text:004013E9                 mov     ebx, 1
.text:004013EE                 mov     esi, 2
.text:004013F3                 mov     edi, 3
.text:004013F8                 push    esi
.text:004013F9                 push    ebx
.text:004013FA                 call    sub_401150
.text:004013FF                 add     esp, 8
.text:00401402                 push    edi
.text:00401403                 push    esi
.text:00401404                 push    ebx
.text:00401405                 call    sub_401164
.text:0040140A                 add     esp, 0Ch
.text:0040140D                 push    [ebp+var_4]
.text:00401410                 push    edi
.text:00401411                 push    esi
.text:00401412                 push    ebx
.text:00401413                 call    sub_401189
.text:00401418                 add     esp, 10h
.text:0040141B                 push    edi
.text:0040141C                 push    esi
.text:0040141D                 push    ebx
.text:0040141E                 call    sub_4011B5
.text:00401423                 add     esp, 0Ch
```

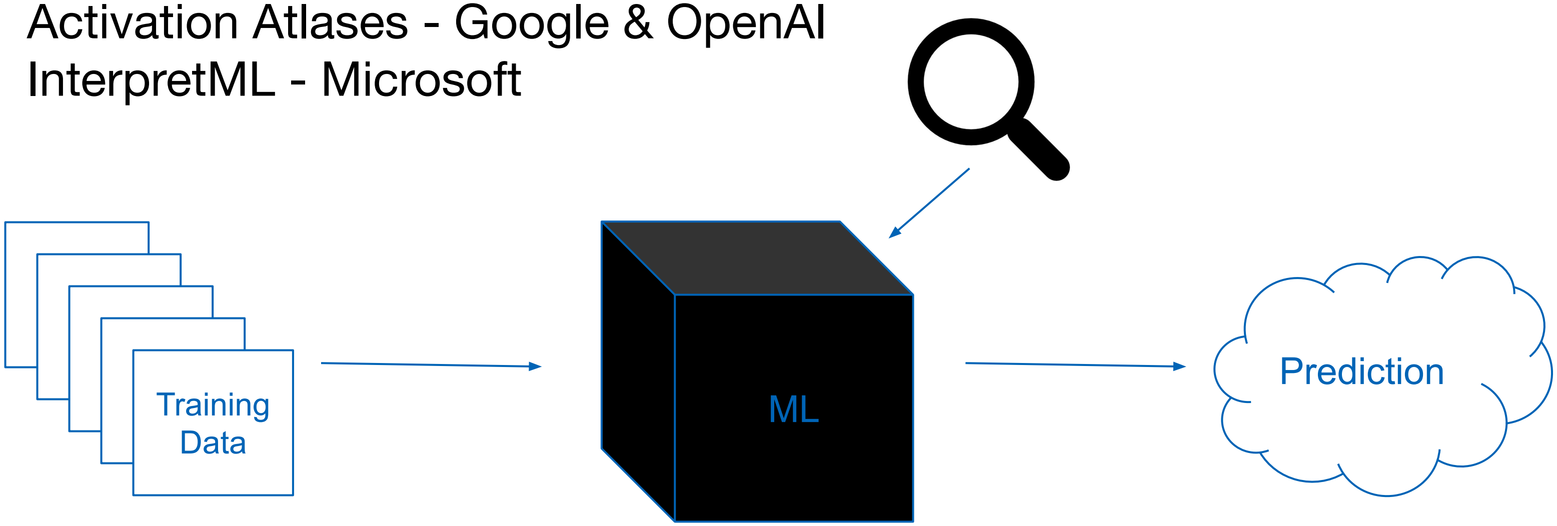
Machine Learning

Malware Classification

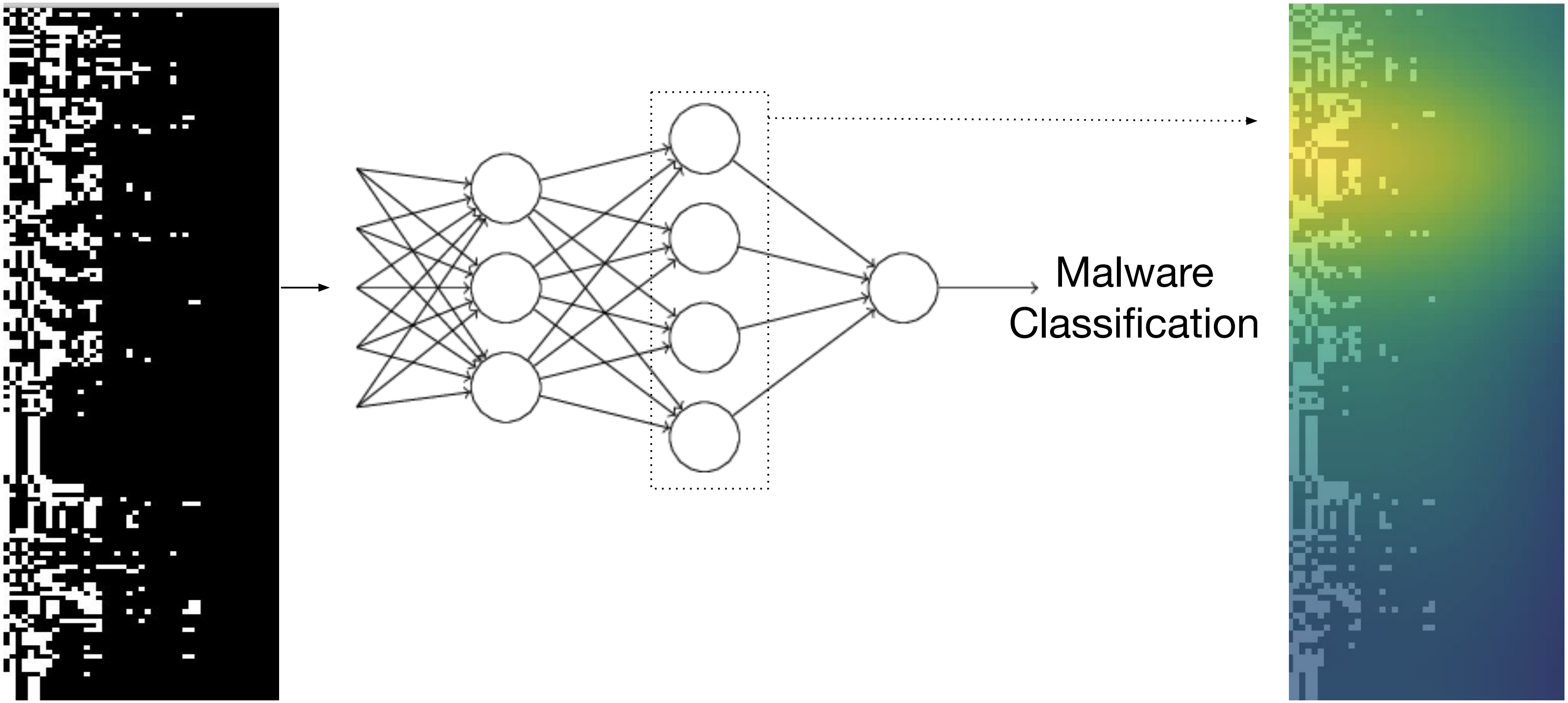
- “Deep Learning on Disassembly Data.” Davis & Wolff. Black Hat 2015
- “Activation Analysis of a Byte-based Deep Neural Network for Malware Classification” Coull. CAMLIS 2018
- “Malware Detection by Eating a Whole EXE” Raff, Barker, Sylvester, Brandon, Catanzaro & Nicholas. arXiv 2017

Explainable ML

- Activation Atlases - Google & OpenAI
- InterpretML - Microsoft



Proof Of Concept



Input
Executable → Image

CNN
ResNet

Output
Grad-CAM Heatmaps

Malware Classification

Convolutional Neural Network

fast.ai

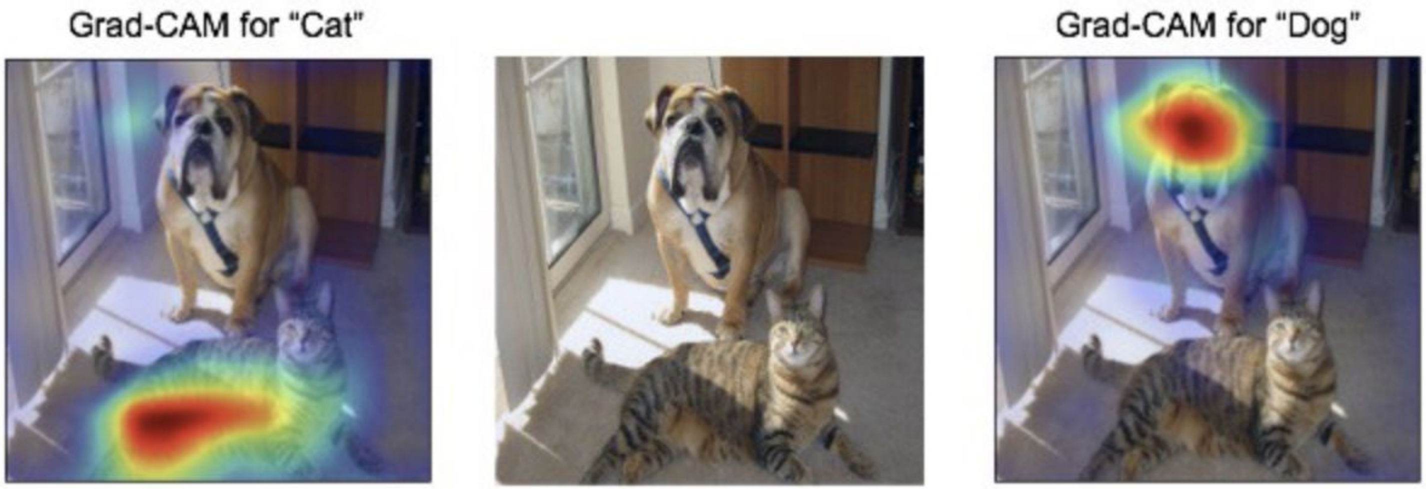
- Simplifies applying state-of-the-art deep learning models and techniques
  - Enables rapid prototyping
- Deep Learning Python Library & MOOC
  - Top Down Approach
  - PyTorch base

```
cnn_learner(data, models.resnet50, metrics=[accuracy])
```

Grad-CAM

Gradient-weighted Class Activation Mappings

- Generated from class-specific gradients passed to a convolutional layer of a CNN
  - Retain spatial information
  - Deeper layers capture higher level visual constructs
- No retraining required
- Guided Grad-CAM: Class discriminative and high resolution



Grad-CAM as Visual Explanations

Establish trust in our models

- Debug ML
- Detect Bias
- Explain unexpected predictions

Machine teaching

- Let models teach humans how to make better decisions about data

Leverage it in other domains

- Medicine
- Malware reverse engineering and analysis

Setup

Data

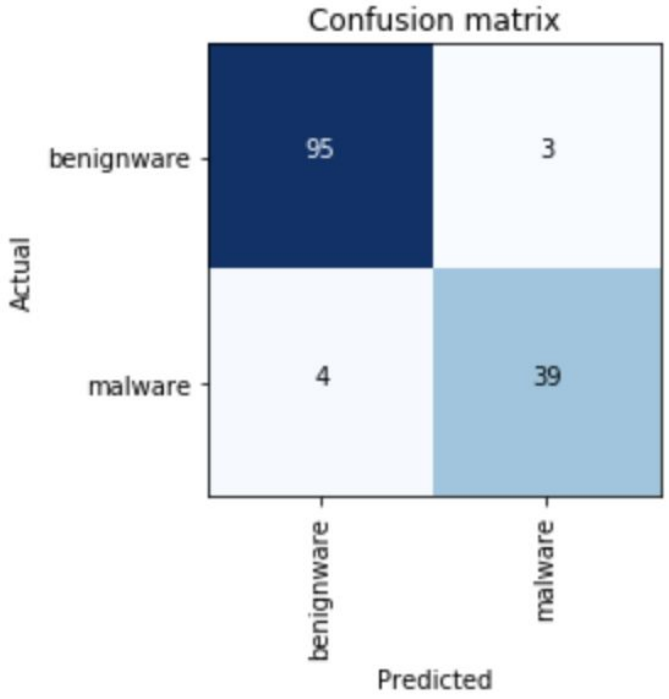
- ~1400 Windows PEs
- 70% benignware , 30% malware

Training

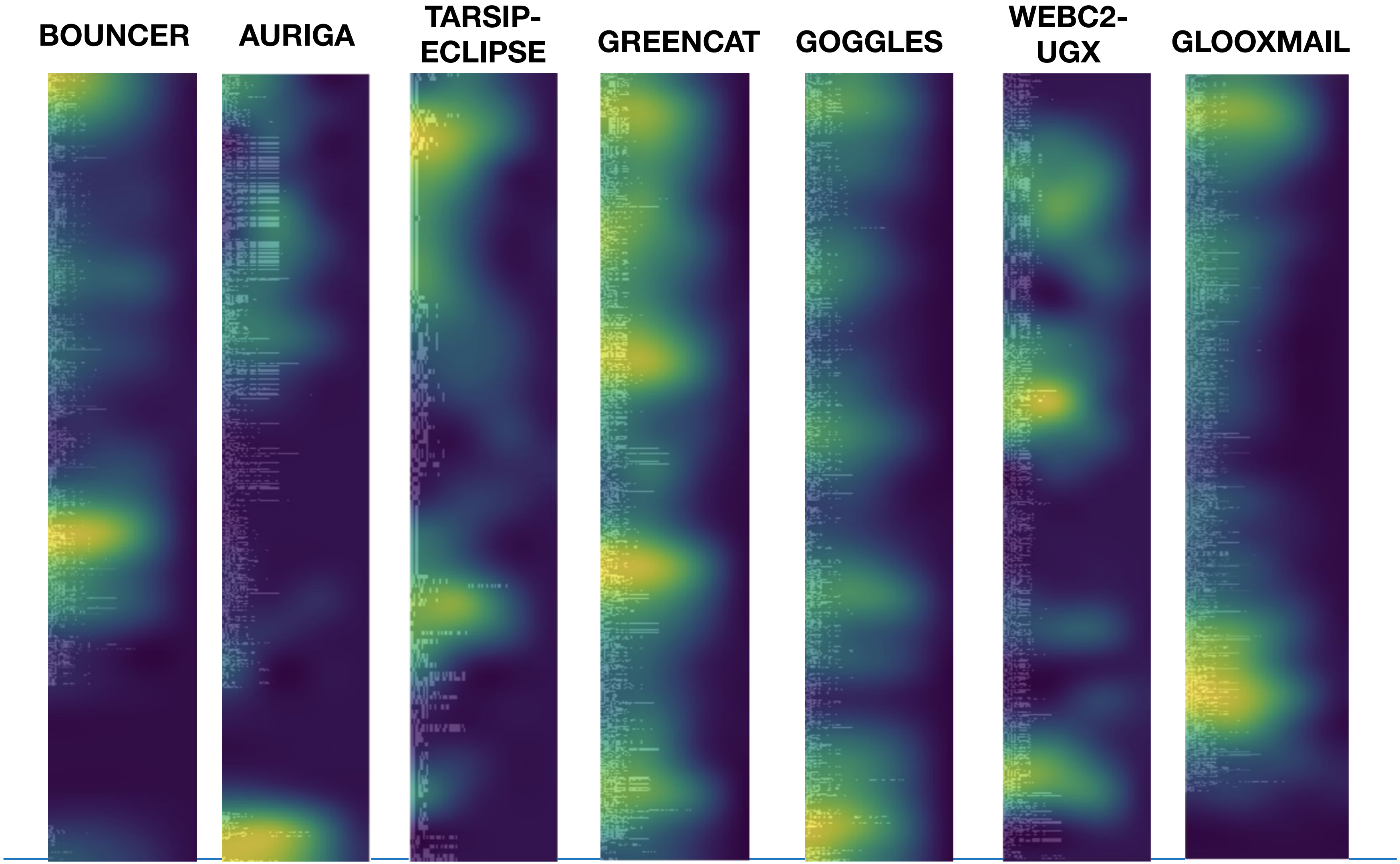
- Train ResNet-50 for malware vs benignware classification
- Grad-CAM to be generated from final Conv Layer
- AWS p2.xlarge

Results

- F1 Score: 96.4%



APT1 Heatmaps



Challenges

Image Representations of Malware

- Variable length images
- Code execution not represented well in this format

Evaluating Heat Maps

- Human judgement

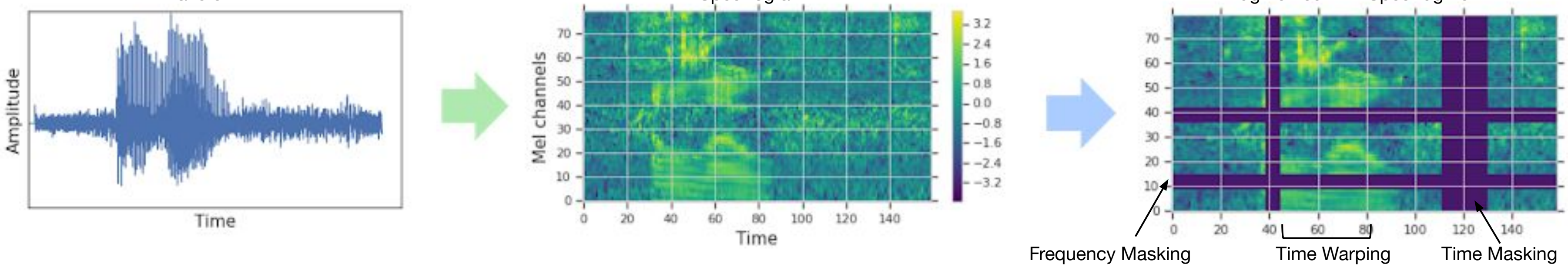
Small Dataset

- ~ 1400 samples from VirusTotal

SpecAugment

SpecAugment is a state of the art data augmentation technique created by Google Brain in April of 2019 for automatic speech recognition tasks.

Audio Classification



Waveform

Spectrogram

Augmented with SpecAugment

Frequency Masking

Time Warping

Time Masking

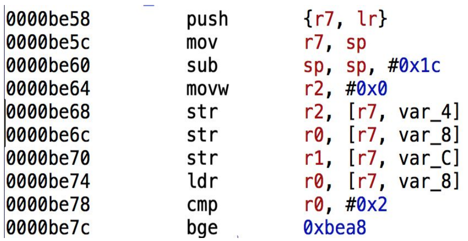
Take a dataset from any domain.

Make it an image classification task.

Use SpecAugment to increase the size of your dataset or prevent your model from overfitting.

Malware Classification

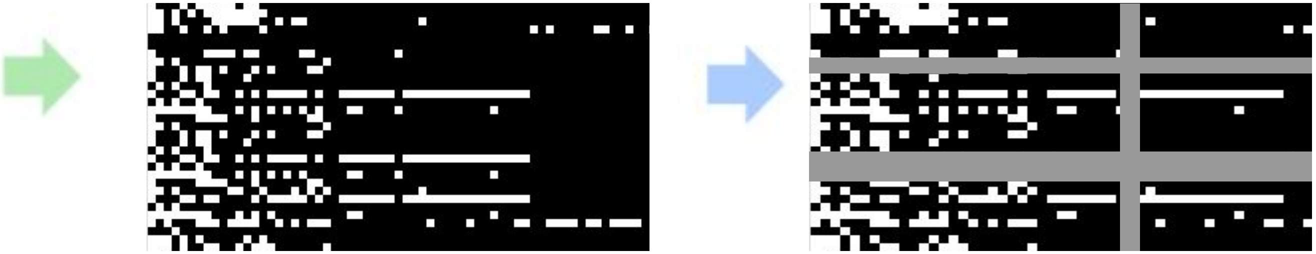
Assembly



```
0000be58        push        {r7, lr}
0000be5c        mov         r7, sp
0000be60        sub         sp, sp, #0x1c
0000be64        movw        r2, #0x0
0000be68        str         r2, [r7, var_4]
0000be6c        str         r0, [r7, var_8]
0000be70        str         r1, [r7, var_C]
0000be74        ldr         r0, [r7, var_8]
0000be78        cmp         r0, #0x2
0000be7c        bge         0xbea8
```

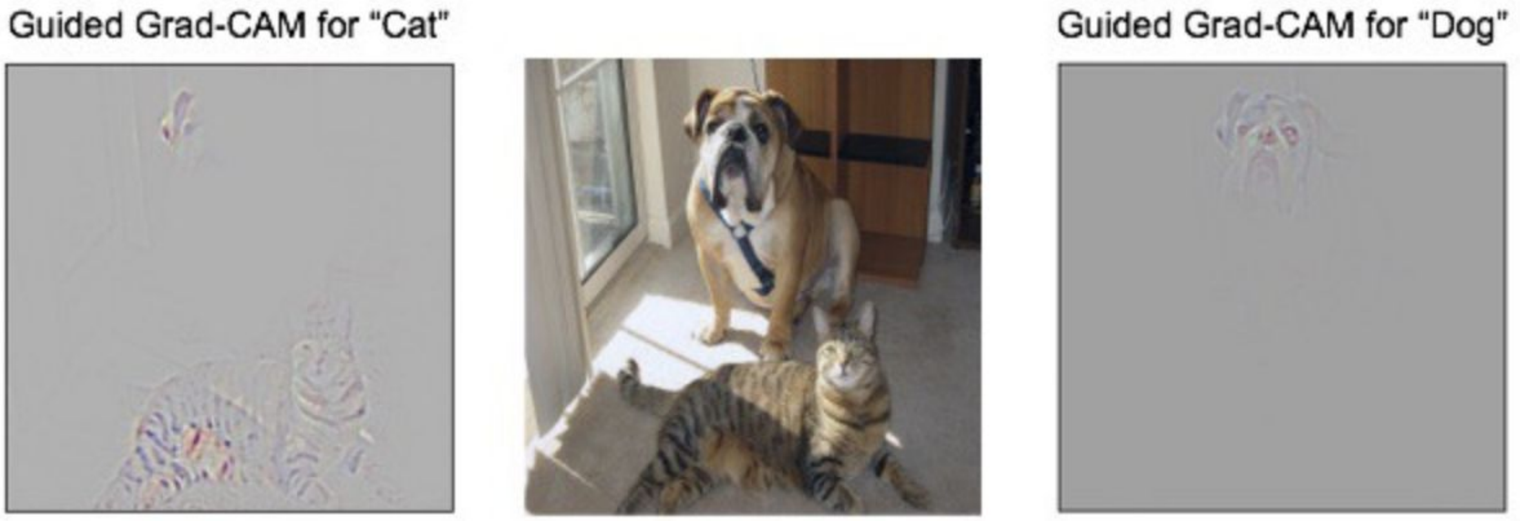
Binary Representation in Pixels

Augmented with SpecAugment



What’s Next

- **Malware Family Classification**
  - Use Guided Grad-CAM to help analysts discover characteristics of classes of malware
- **Larger Dataset**
- **Input Configuration**



Links

- Grad-CAM
- fast.ai
- SpecAugment
- InterpretML
- Activation Atlases