Development of C programs for
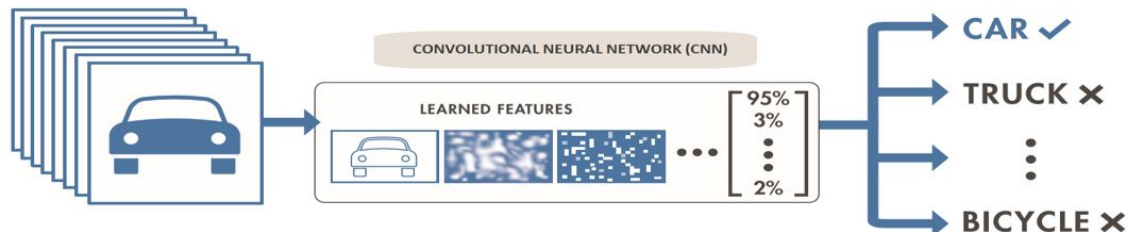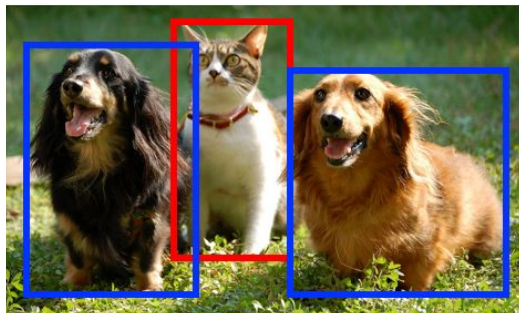# Convolutional Neural Network Accelerators

Sujin Kang

Prof. Nikil Dutt, Hamid Nejatollahi, Kenshu Seto

Dept. of Computer Engineering in Hanyang University
University of California Irvine
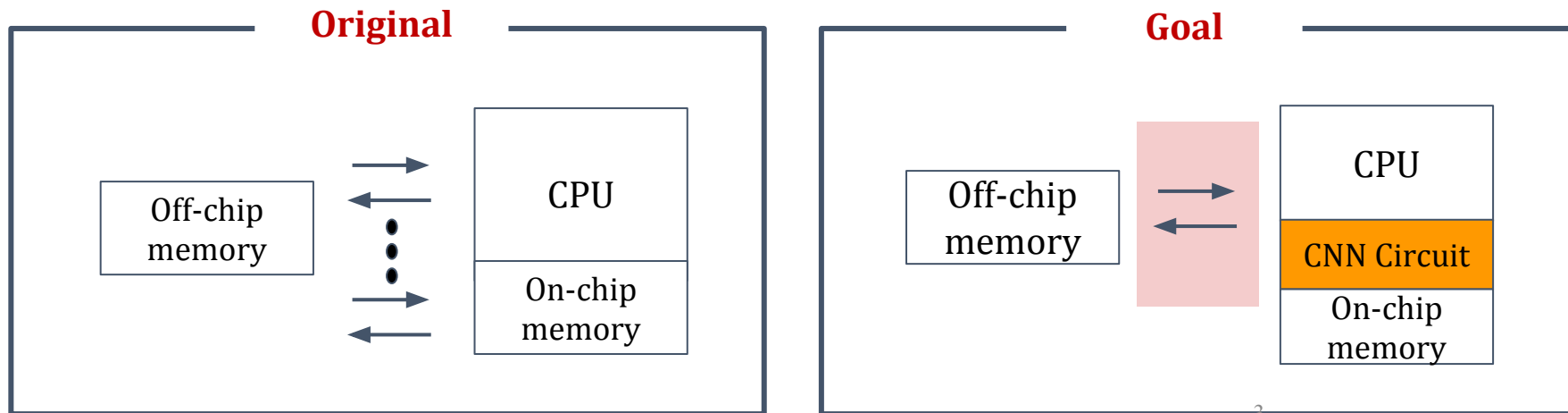University in Tokyo

Convolution Neural Network


```
Layer (type)                        Output Shape          Param #      Connected to
=================================================================================
maxpooling2d_24 (MaxPooling2D)      (None, 512, 7, 7)     0            maxpooling2d_input_4[0][0]

batchnormalization_10 (BatchNorm    (None, 512, 7, 7)     1024         maxpooling2d_24[0][0]

flatten_8 (Flatten)                 (None, 25088)         0            batchnormalization_10[0][0]

dense_22 (Dense)                    (None, 4096)          102764544    flatten_8[0][0]

dropout_15 (Dropout)                (None, 4096)          0            dense_22[0][0]

batchnormalization_11 (BatchNorm    (None, 4096)          8192         dropout_15[0][0]

dense_23 (Dense)                    (None, 4096)          16781312     batchnormalization_11[0][0]

dropout_16 (Dropout)                (None, 4096)          0            dense_23[0][0]

batchnormalization_12 (BatchNorm    (None, 4096)          8192         dropout_16[0][0]

dense_24 (Dense)                    (None, 2)             8194         batchnormalization_12[0][0]
=================================================================================
Total params: 119571458
```
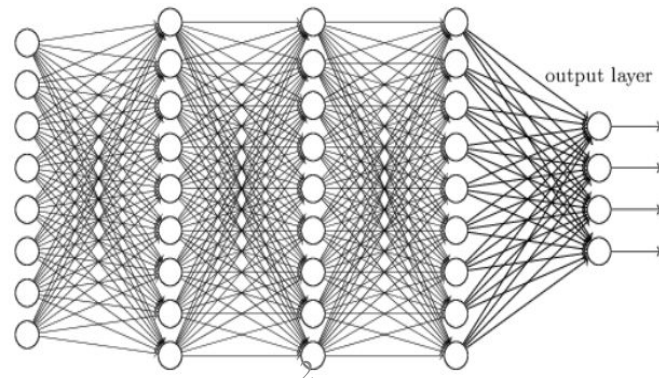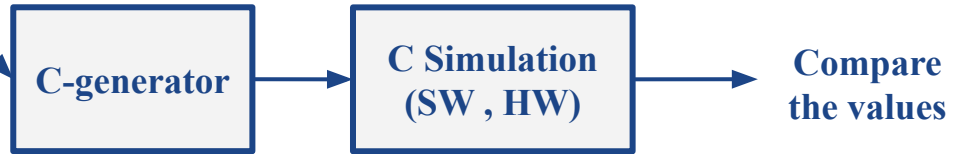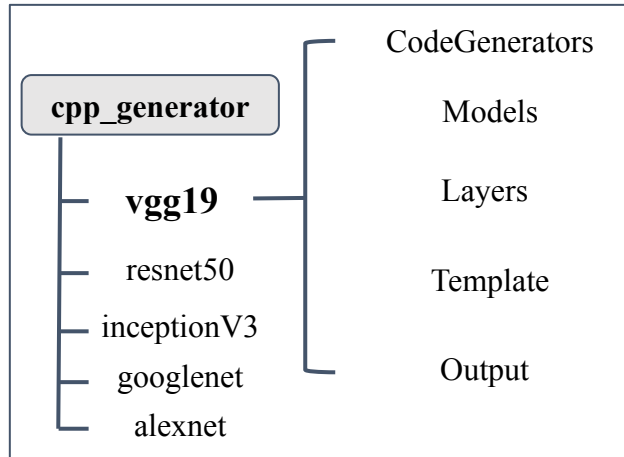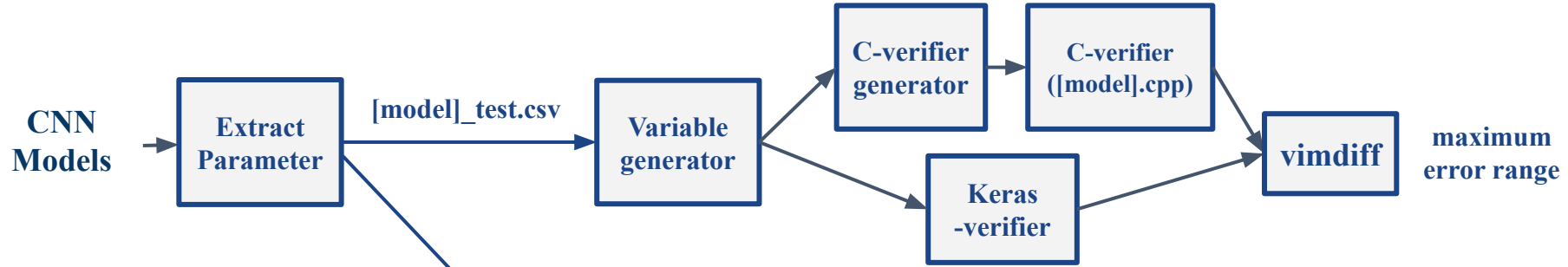
Research task & Goal

Acceleration of Convolution Neural Networks
to embed in light machine

Focus on Reducing Memory access time

Original

Off-chip
memory

CPU

On-chip
memory

Goal

Off-chip
memory

CPU

CNN Circuit

On-chip
memory

3

Work process

Result

Print Diff & Maximum error range

C_verifier_code save generated C code in various models

test_file [model_name] _test file (csv file, contain layer information)

cpp_generator

- **CodeGenerators**
  - C_verifier.py : generate 1) software function code by layers, 2) static variables, 3) output variables, 4) software function call
  - CodeGenerators.py : generate 1) print function code, 2) initialization code
- **Layers** save data and layer information
  - Conv2D.py
  - Data.py
  - InputLayer.py
  - Layers.py
- **Models** construct Models and save output values
  - hw
  - Models.py
- **Template** store template files in 4 directories
  - Function
  - Init
  - Main
  - Print
- **Output** store the files of outputs

extract-network-params extract network params in each model

keras-verification using keras library to compare values between keras and c code

- Keras_Verifier.py
- verifier.sh

variable_generator generate value of variables in weight, bias, input array (randomly)


```
Convolution2D : [[[[3428 3542 3721 3759 2827 2844 3570 3237]
   [3079 3679 3789 3859 3249 2889 2926 3300]
   [3318 3542 3467 3522 3155 2820 2562 3521]
   [2783 2933 3661 3818 3437 2833 3332 3546]
   [2159 2861 3279 3297 2945 3018 3529 3758]
   [3157 3610 3429 3103 3313 2761 3319 4211]
   [2610 3004 3469 3512 3224 3061 3686 3620]
   [2746 3398 3917 3770 3378 3416 3533 3549]]

  [[2815 3419 3223 3203 2731 2819 2837 2572]
   [2726 2907 3173 3244 2733 2543 2331 2895]
   [2795 2593 2848 3171 2468 1895 2718 2747]
   [2097 2691 3212 3091 2640 2830 3032 2941]
   [2537 2888 2977 2612 2660 2384 2545 3163]
   [2383 2615 3061 2947 2504 2500 3141 3426]
   [2156 2616 2872 2751 2242 2391 2832 3449]
   [2614 3338 3138 3187 3036 2965 2824 3295]]

  [[3643 4003 4188 4214 3544 2944 3783 3894]
   [3557 3803 4433 4318 3331 2941 3637 4168]
   [3592 4275 4153 3891 3333 3113 3308 3704]
   [3416 3503 3904 3893 3999 3120 3598 4074]
   [2707 3046 3870 3855 3400 3046 3624 3881]
   [2746 3545 3931 3618 3283 3250 4202 4530]
   [3034 3762 3927 3810 3526 3324 4025 4451]
   [3252 3876 4325 4446 3940 3966 3908 3945]]

  [[3452 3814 4295 4258 3326 2917 3666 3340]
   [3186 3591 3769 3769 2831 2885 3324 3522]
   [3202 3703 3768 3548 3257 3453 3007 3358]
   [3181 3465 3683 3683 3631 2938 3409 3987]
   [2485 2992 3762 3512 2948 2946 3473 3839]
   [2646 3472 3631 3080 3094 3092 3937 4281]
   [3220 3679 3866 3410 3260 3060 3476 4120]
   [2835 3635 4104 4407 3779 3325 3827 3640]]]]
BatchNormalization : [[[[3426 3540 3719 3757 2825 2842 3568 3235]
   [3077 3677 3787 3857 3247 2887 2924 3298]
Output/keras_output.txt
```

```
Convolution2D : [[[[3428 3542 3721 3759 2827 2844 3570 3237]
   [3079 3679 3789 3859 3249 2889 2926 3300]
   [3318 3542 3467 3522 3155 2820 2562 3521]
   [2783 2933 3661 3818 3437 2833 3332 3546]
   [2159 2861 3279 3297 2945 3018 3529 3758]
   [3157 3610 3429 3103 3313 2761 3319 4211]
   [2610 3004 3469 3512 3224 3061 3686 3620]
   [2746 3398 3917 3770 3378 3416 3533 3549]]

  [[2815 3419 3223 3203 2731 2819 2837 2572]
   [2726 2907 3173 3244 2733 2543 2331 2895]
   [2795 2593 2848 3171 2468 1895 2718 2747]
   [2097 2691 3212 3091 2640 2830 3032 2941]
   [2537 2888 2977 2612 2660 2384 2545 3163]
   [2383 2615 3061 2947 2504 2500 3141 3426]
   [2156 2616 2872 2751 2242 2391 2832 3449]
   [2614 3338 3138 3187 3036 2965 2824 3295]]

  [[3643 4003 4188 4214 3544 2944 3783 3894]
   [3557 3803 4433 4318 3331 2941 3637 4168]
   [3592 4275 4153 3891 3333 3113 3308 3704]
   [3416 3503 3904 3893 3999 3120 3598 4074]
   [2707 3046 3870 3855 3400 3046 3624 3881]
   [2746 3545 3931 3618 3283 3250 4202 4530]
   [3034 3762 3927 3810 3526 3324 4025 4451]
   [3252 3876 4325 4446 3940 3966 3908 3945]]

  [[3452 3814 4295 4258 3326 2917 3666 3340]
   [3186 3591 3769 3769 2831 2885 3324 3522]
   [3202 3703 3768 3548 3257 3453 3007 3358]
   [3181 3465 3683 3683 3631 2938 3409 3987]
   [2485 2992 3762 3512 2948 2946 3473 3839]
   [2646 3472 3631 3080 3094 3092 3937 4281]
   [3220 3679 3866 3410 3260 3060 3476 4120]
   [2835 3635 4104 4407 3779 3325 3827 3640]]]]
BatchNormalization : [[[[3426 3540 3719 3757 2825 2842 3568 3235]
   [3077 3677 3787 3857 3247 2887 2924 3298]
Output/C_output.txt
```
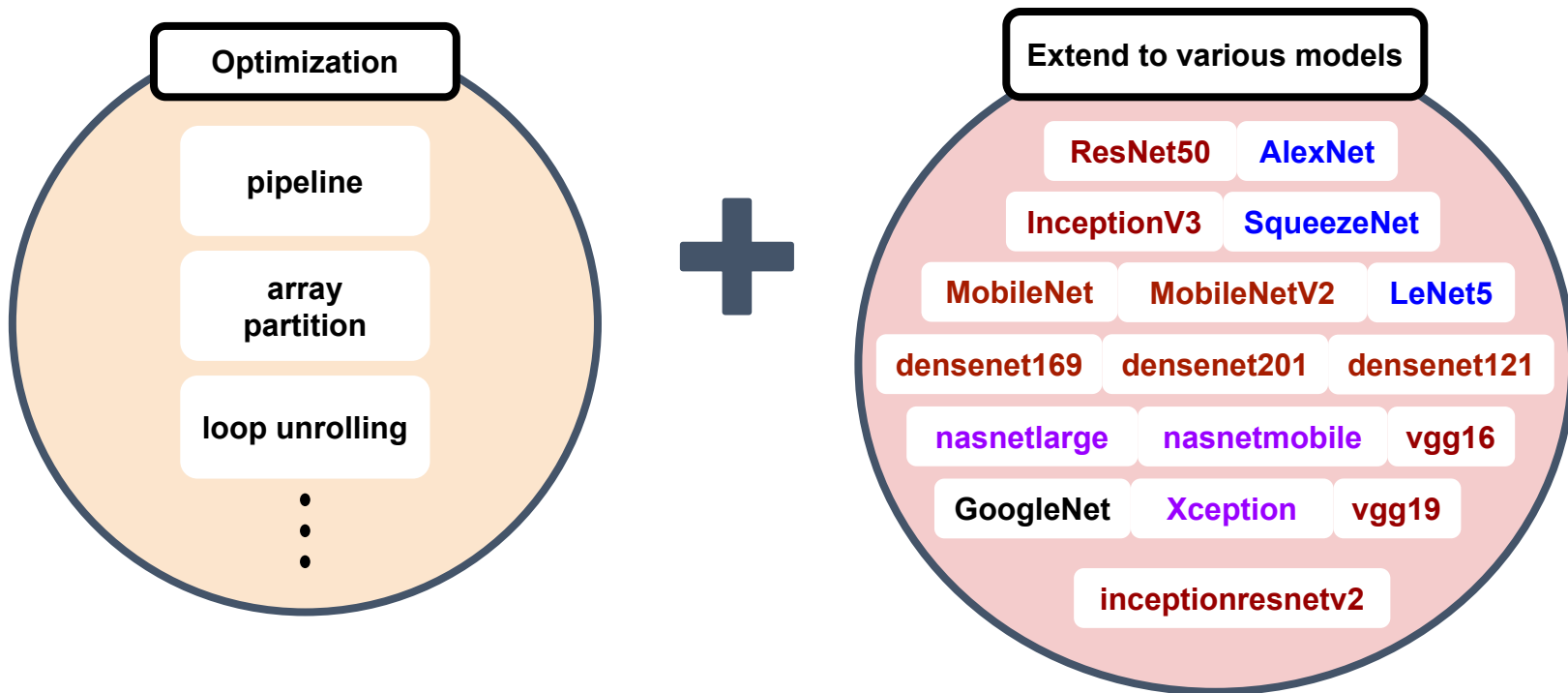
```
maximum error : 1.79645846179e-06 when c has an element of 82081644544.0 and keras has an element of82081497088.0
2 files to edit
```

5

Future Work

Optimization

- pipeline
- array partition
- loop unrolling
- ⋮

+

Extend to various models

ResNet50 AlexNet
InceptionV3 SqueezeNet
MobileNet MobileNetV2 LeNet5
densenet169 densenet201 densenet121
nasnetlarge nasnetmobile vgg16
GoogleNet Xception vgg19
inceptionresnetv2

THANK YOU