설명가능한 인공지능 (CAM)

Prof. Seungchul Lee

Industrial AI Lab.

Issues on CNN (or Deep Learning)

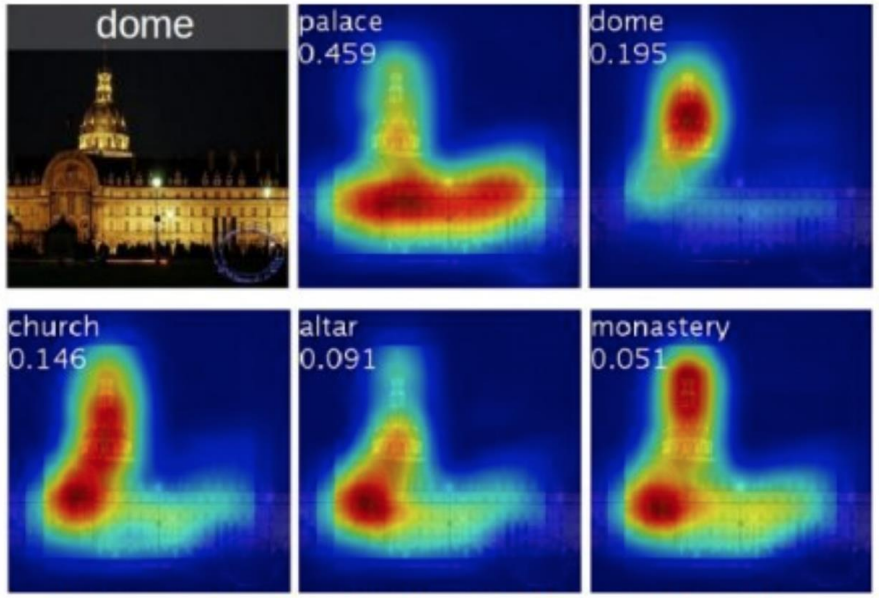
- Deep learning performs well comparing with any other existing algorithms
- But works as a black box
  - A classification result is simply returned without knowing how the classification results are derived → little interpretability

- When we visually identify images, we do not look at the whole image
- Instead, we intuitively focus on the most important parts of the image
- When CNN weights are optimized, the more important parts are given higher weights

- Class activation map (CAM)
  - We can determine which parts of the image the model is focusing on, based on the learned weights
  - Highlighting the importance of the image region to the prediction

Visualizing Convolutional Neural Networks

- Class Activation Maps (CAMs)
- A class activation map (CAM) for a given class highlights the image regions used by the CNN to identify that class

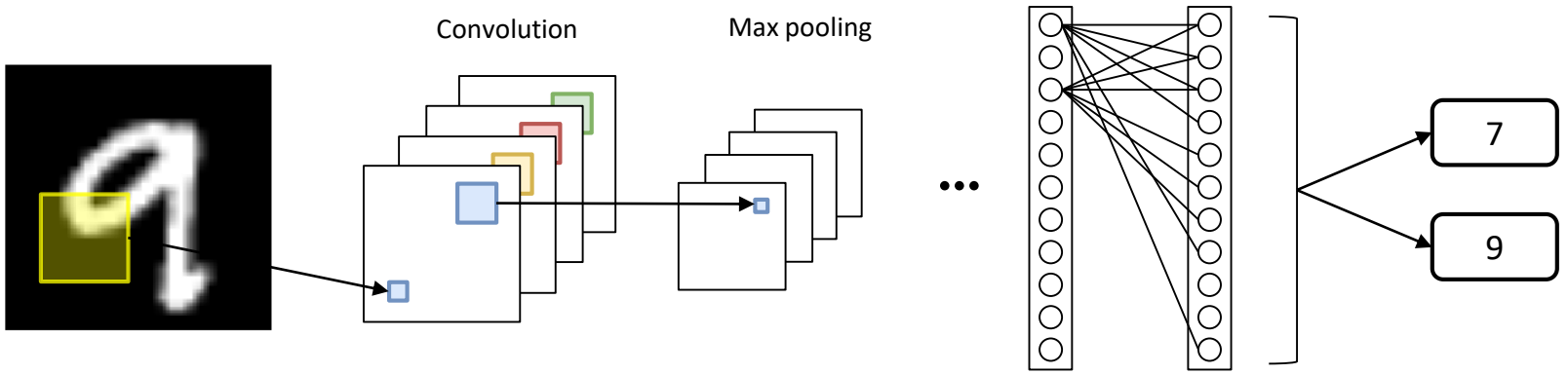


Class activation maps of top 5 predictions

Class activation maps for one object class

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.
Learning Deep Features for Discriminative Localization. CVPR'16

Fully Connected Layer

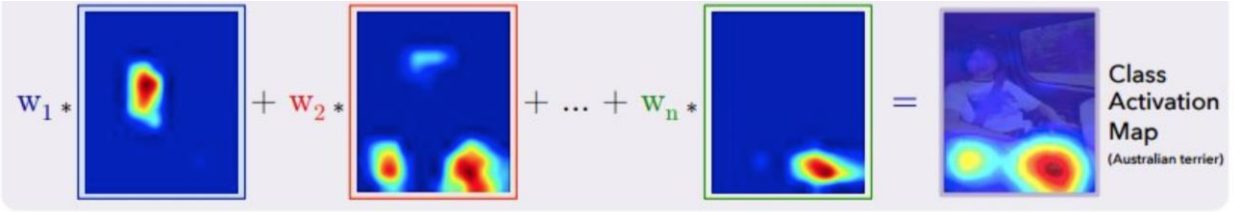


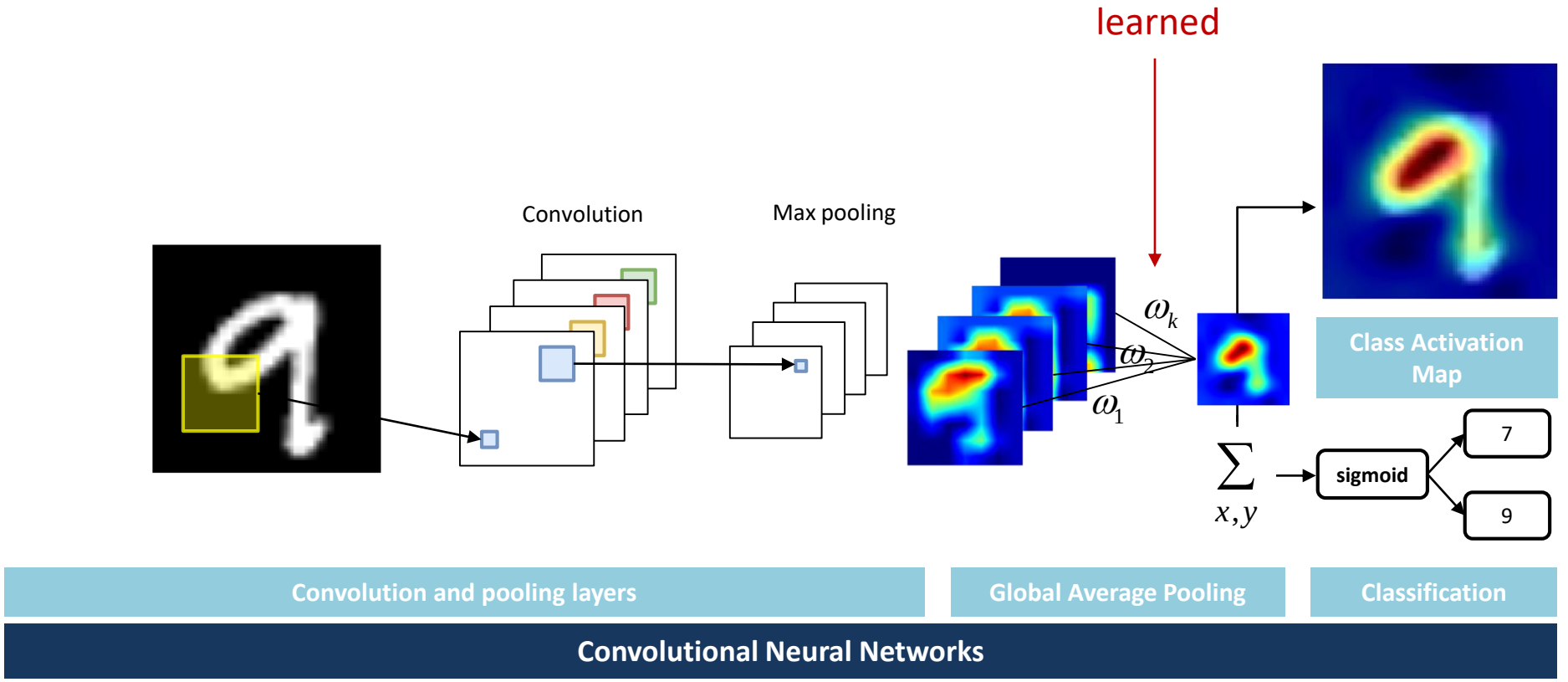
Convolution and pooling layers | Fully connected layer | Classification

Convolutional Neural Networks

Global Average Pooling

- Class Activation Map (CAM)
- (or Attention)





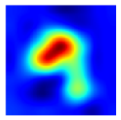
B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.
Learning Deep Features for Discriminative Localization. CVPR'16

Global Average Pooling Implementation

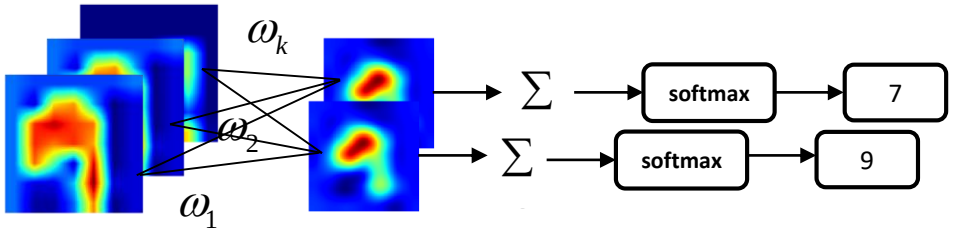
```
model = tf.keras.models.Sequential([
    tf.keras.layers.Conv2D(32,
                           (3,3),
                           activation = 'relu',
                           padding = 'SAME',
                           input_shape = (28, 28, 1)),
    tf.keras.layers.MaxPool2D((2,2)),
    tf.keras.layers.Conv2D(64,
                           (3,3),
                           activation = 'relu',
                           padding = 'SAME',
                           input_shape = (14, 14, 32)),
    tf.keras.layers.MaxPool2D((2,2)),
    tf.keras.layers.GlobalAveragePooling2D(),
    tf.keras.layers.Dense(2, activation = 'softmax', use_bias = False)
])
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 28, 28, 32)        320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 14, 14, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 14, 14, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 7, 7, 64)          0
_________________________________________________________________
global_average_pooling2d (Gl (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 2)                 128
=================================================================
```

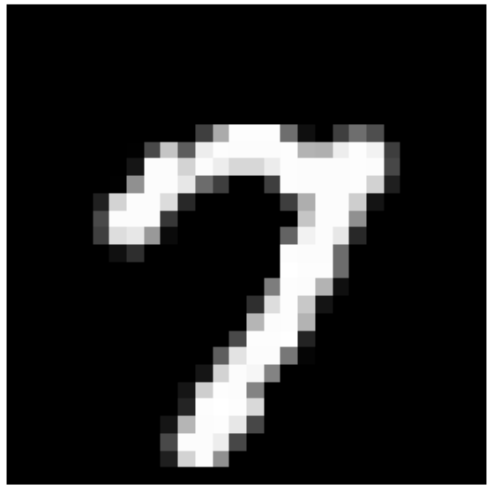
$$S_c = \sum_k \omega_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \underbrace{\sum_k \omega_k^c \; f_k(x,y)}$$

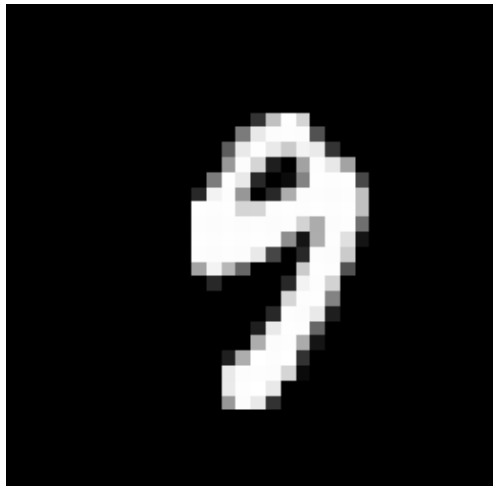


$$P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)}$$

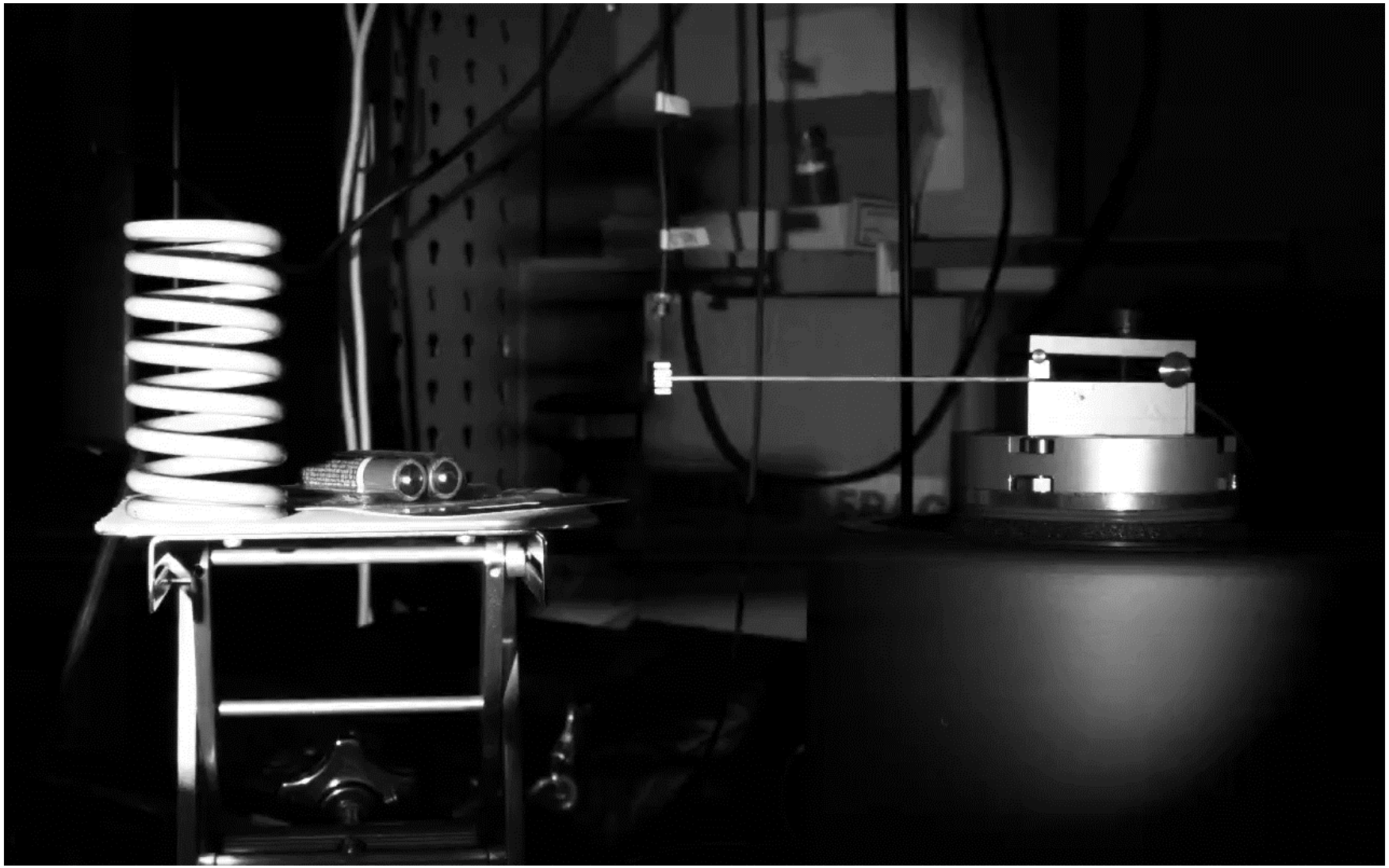


Example: MNIST



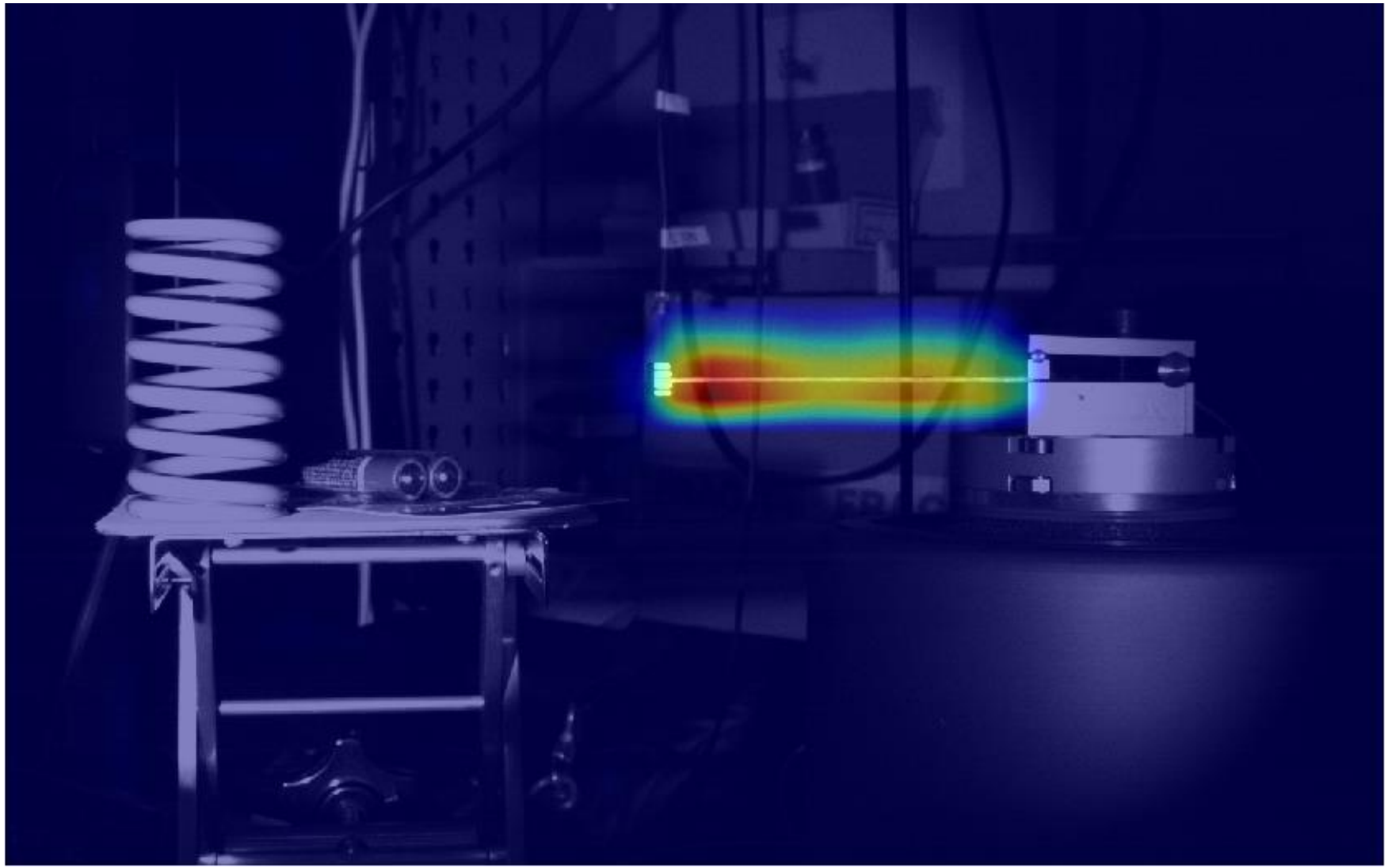


Cantilever



In collaboration with KIMM

Cantilever



In collaboration with KIMM