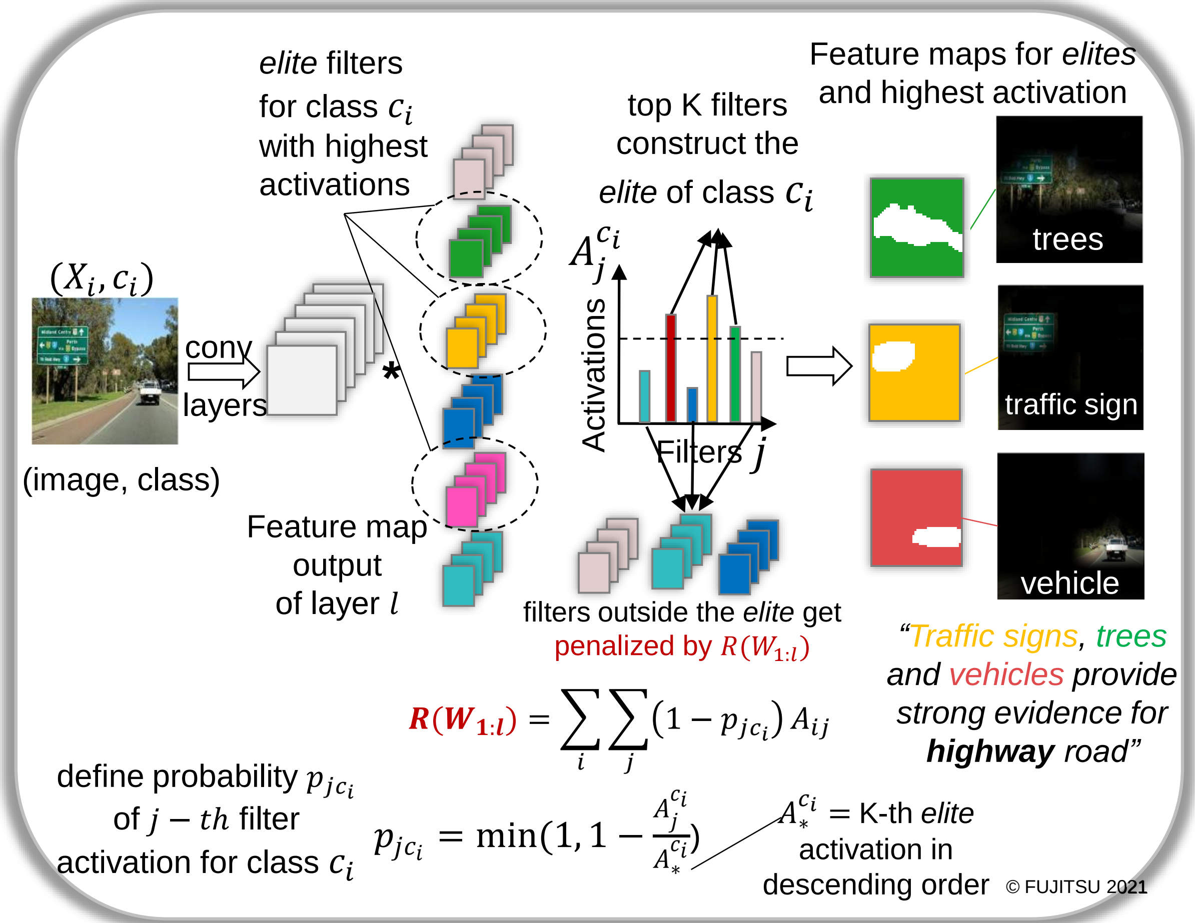
Elite BackProp: Training Sparse Interpretable Neurons

- Elite BackProp (EBP) is a method to train more interpretable neural networks by inducing **class-wise activation sparsity**.

- **EBP associates each class with a handful of *elite filters*** that activate rarely and have strong activation magnitude on images from that class.

- During training **filters outside the elite** of each class **will be penalized** by a factor proportional to the magnitude of their activation.

- **Experiments show that EBP enhances the performance of a rule extraction algorithm** that explains a CNN by inducing compact rules with high fidelity to the original model.



(X_i, c_i)
(image, class)
conv layers
*
elite filters for class c_i with highest activations
Feature map output of layer l
top K filters construct the *elite* of class c_i
$A_j^{c_i}$
Activations
Filters j
filters outside the *elite* get penalized by $R(W_{1:l})$
Feature maps for *elites* and highest activation
trees
traffic sign
vehicle
define probability p_{jc_i} of $j-th$ filter activation for class c_i
$A_*^{c_i}$ = K-th *elite* activation in descending order

$$R(W_{1:l}) = \sum_i \sum_j (1 - p_{jc_i})\, A_{ij}$$

$$p_{jc_i} = \min\left(1, 1 - \frac{A_j^{c_i}}{A_*^{c_i}}\right)$$

*"Traffic signs, trees and vehicles provide strong evidence for **highway** road"*