Need to distinguish our work with pure model compression and knowledge distillation.

Our key: Input specialization leads to opportunities of using a specialized, compressed, more efficient model without loss of accuracy, sometimes, even better accuracy. We conduct the first comprehensive analysis on the mapping between input specialization and model compression. Our insights turns into an automatic framework, which equips a pool of specialized models distilled effectively from a general, complex model. It would select the best special model for a given set of inputs at runtime.

a) Define input specialization? (Question: do we want to only focus on reduction of #classes? e.g., 10 classes out 100 classes. How about other kind specializations? (e.g., type of classes; input distributions in each of the class, angle variation, i.e., photos taken from a camera with a fixed setting; background variation; image size...; and their combinations.)

Think about how input specialization could happen in real scenario.

other types of specialization.

b) How would input specializations lead to such efficient model? (Question: do we want to only focus on CNN? how about other ML algorithms?)

multi-class SVM?

Take CNN as an example:

* Convolutional layer is to learn various level of abstractions. Less classes -> less features in each layer + possibly no need of some high-level feature abstraction -> possible reduction in in #filters+ #layers.

* Fully-connected layer is to learn the final classification. Less classes -> less neurons in the final layer, and possibly less layers.

Any idea or analytical model that predict what kind of compression can be induced for a specialized set of inputs?

新的压缩模型的方法，所以压缩率更高.