Improved Techniques for Training Adaptive Deep Networks
Hao Li[1]*, Hong Zhang[2]*, Xiaojuan Qi[3], Ruigang Yang[2], Gao Huang[3]
* Equal Contribution

Motivation
- Adaptive Inference
  - Adjust the network structure dynamically based on inputs
  - Improve computational efficiency at test time
  - Use small models for “easy” inputs while big models for “hard” inputs

Method
- Forward knowledge transfer
- Backward knowledge transfer
- Student classifiers
- Teacher classifiers
- Regular forward
- Gradient scaling
- Conv modules

Results

Resolve gradient conflicts among classifiers

Gradient Equilibrium (GE)
- Rescale the magnitude of gradients along its backward propagation path.
  \[ R(x; s) = x; \nabla_R(x; s) = s \]

Encourage collaboration of classifiers

Inline Subnetwork Collaboration (ISC)
- Prediction of previous stage serves as a prior to facilitate learning of classifiers.

One-for-all Knowledge Distillation (OFA)
- The last classifier serves as a teacher model whose knowledge could be distilled into earlier exits.
  \[ L_i = \alpha CE_i + (1 - \alpha) KLD_i \]

Training adaptive inference networks effectively is difficult:

- How to resolve the conflicts among classifiers
- How to encourage the collaboration of classifiers

Studies

Ablation Studies

PyTorch Implementation: https://github.com/kalviny/IMTA

Links

PyTorch Implementation: https://github.com/kalviny/IMTA