Language Understanding

06 - Training Strategies

Hossein Zeinali

Training Strategies

- There are various types of training in Deep Learning
  - Standard Training (training from scratch)
  - Pre-Training and Fine-Tuning
    - Self-supervise learning
  - Few-Shot Learning
  - One-Shot Learning
  - Zero-Shot Learning

- There are no clear advantages for each strategy, and the best one should be selected based on the task.

Standard Training

- Train a model from scratch
  - We have to do training time to time
- It needs considerable labeled training data for each task
  - Data collection process is time-consuming and expensive
- Most of the conventional NLP systems were trained using this strategy
- Usually, the training process is slow and needs lots of computation cost, especially for DNN based methods.

Pre-Training and Fine-Tuning

- As mentioned before, collecting an abundant labeled training data is expensive
- Idea: pre-train a model somehow using unlabeled data and fine-tune the model to the specific task using small training data
- **Fine-Tuning (FT)** has been the most common approach in recent years
  - The weights of a pre-trained model are updated by training on a supervised dataset.
  - Typically, thousands to hundreds of thousands of labeled examples are used.
- The main advantage of fine-tuning is strong performance on many benchmarks.
- The main disadvantages are the need for a new large dataset for every task, the potential for poor generalization out-of-distribution.

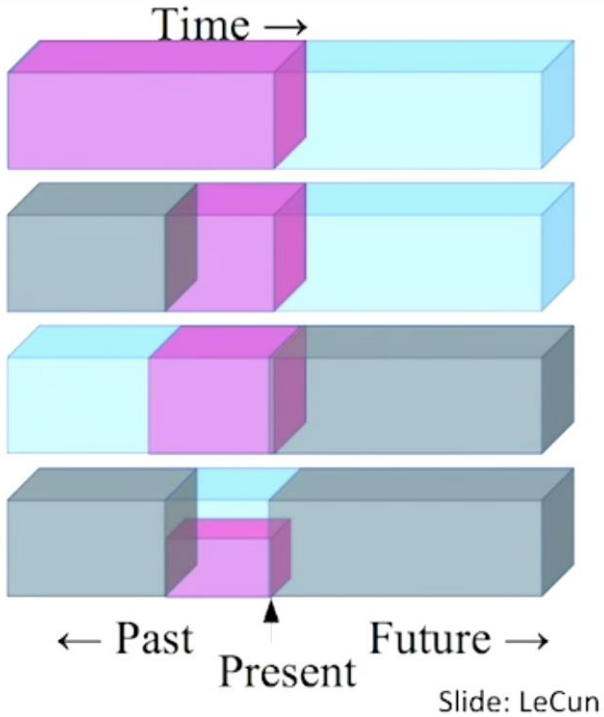
Pre-Training and Fine-Tuning

Self Supervision

- Supervised training:
  - Need a labeled training data
- Unsupervised training
  - Does not any labeled training data
  - Is useful in the generative scenario
- Self supervision: train a supervised model using unlabeled data
  - NLP: language modeling using next-word prediction of text data
  - Speech: acoustic modeling using next-frame prediction of speech data
  - Image: find the relation between patches or recovered the distorted images

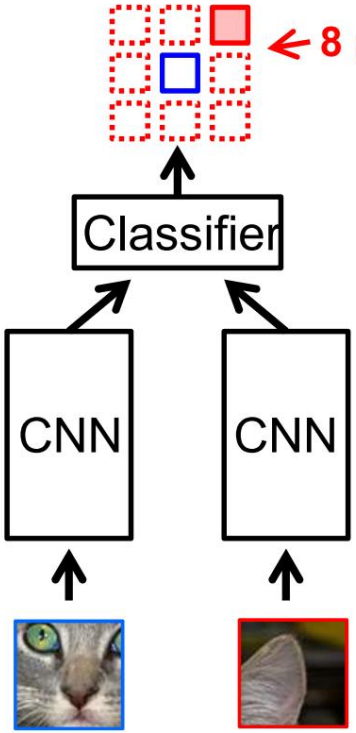
Self Supervision

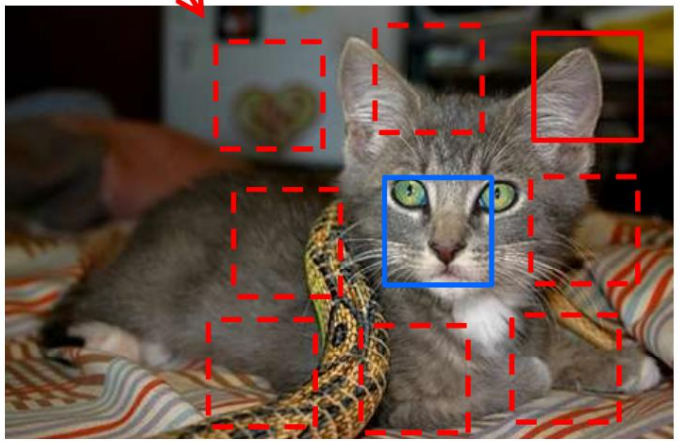
- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Slide: LeCun

Self Supervision

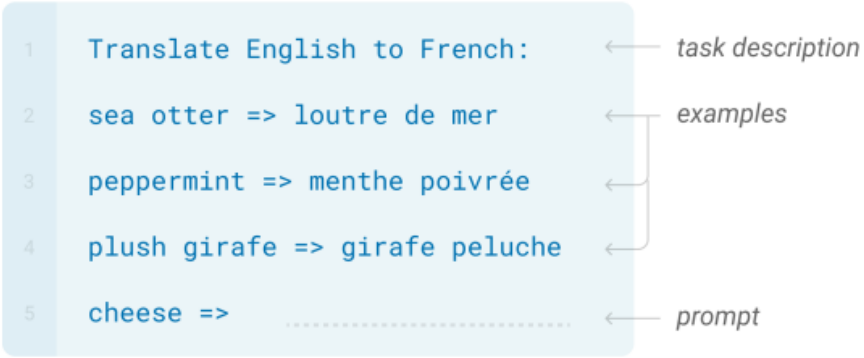




Source: Unsupervised visual representation learning by context prediction

Few-Shot Learning

- The model is given a few demonstrations of the task at inference time as conditioning, but no weight updates are allowed.
- Few-shot works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion. K is typically set in the range of 10 to 100 based on the model's context window.



One-Shot Learning

- Is the same as few-shot except that only one demonstration is allowed, in addition to a natural language description of the task.
- This condition most closely matches the way in which some tasks are communicated to humans.

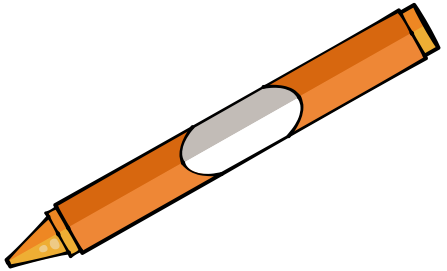
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Zero-Shot Learning

- Is the same as one-shot except that no demonstrations are allowed, and the model is only given a natural language instruction describing the task.
- This method provides maximum convenience, potential for robustness, and avoidance of spurious correlations, but is also the most challenging setting.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

Thanks for your attention



References and IP Notice

- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).