Applying AI in telco domain

Deep learning on the field: from labs to real-world environment

Kovács Ferenc

Nokia Bell Labs

Topics

- **Introduction**
- Domain Adaptation
- Knowledge Distillation
- Watermarking

Using AI in networks

- What can AI do for networks?
  - Help implement specialized use cases (embedded in network functions or services)
  - Provide better insights (data analytics)
  - Enabler for automation (closed-loop management)
- AI brings new operational and management requirements
  - AI model encapsulation for deployment and model specific processing
  - Management of the AI itself (data and control interfaces, training cycles, LCM)
  - Continuous performance monitoring and assurance to contain model errors and prevent failure cascading (trust, verification and dependability issues)
- Need for new learning approaches
  - Unsupervised learning methods (without labeled data)
  - Inclusion of domain knowledge (interpretability)

Telco and traditional AI use cases

	Human centric AI	Industrial/telco AI
Training	Most results created by supervised learning on complete and fully labeled data sets.	No labeled data may exist, retrofitting labels on existing data is challenging or impossible.
Data quality and quantity	Huge quantity of high quality data is available (even if sometimes not publicly accessible).	Data quality and quantity are highly varying, lots of erroneous or missing data, poor or missing metadata.
Data diversity	Single-domain and uniform datasets (given size of images, speech samples, text, etc.).	Data is much more diverse (complexity of technology, service variety, vendors, etc.).
Data dynamicity	Mostly stationary datasets (e.g., the dataset of 14M images on ImageNet) as baseline training data.	Data is not from a stationary distribution but changes (with user behavior, network and OTT technology, SW updates, etc.)
Benchmarking and interpretation	Human performance as benchmark (base error), results can easily be interpreted by human perception.	No established base error, results on structured data are not intuitive for human perception, hard to validate.

Data modeling challenges

Missing important parts of network state and domain knowledge

- Human centric AI
  - The full input state is represented by the data (e.g., image, speech sample, sentence)
  - Expected output is known and can be inferred from the input (no hidden variables)
- Network measurements
  - Data is aggregated over time, NF instances, measurement groups, interfaces, etc.
  - Measurement coverage may be low considering the full system and e2e services
  - Only partial states are observed (unknown number of hidden variables; system state changes cannot be inferred from observed state)
  - Data may have unclear semantics (e.g., collected from 3rd party system)

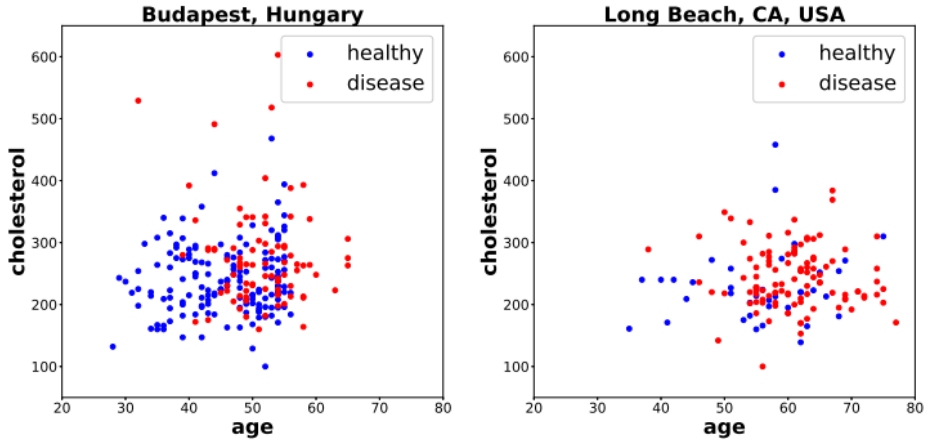
Model universality and reusability

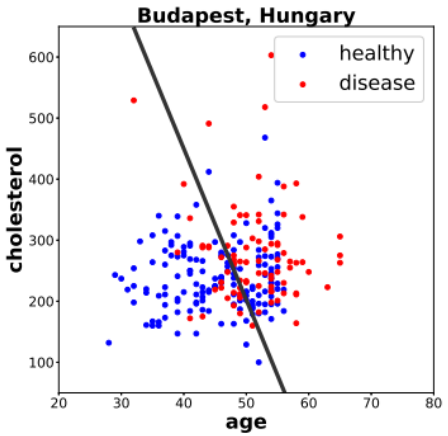
- Human centric AI
  - Focus is on solving similar perception or control problems based on universal datasets (e.g., MNIST, ImageNet)
  - Transfer learning between the models or extending an existing model may be possible
- AI for networks
  - Models trained on private organizational data are specific to the data and problem *(recall from course 1: NN architecture reflects input and output, weights reflect unique data distribution)*
  - Creating universally applicable and shareable models is challenging
  - Models are IPR
  - Limited resources (e.g., RAN, edge)

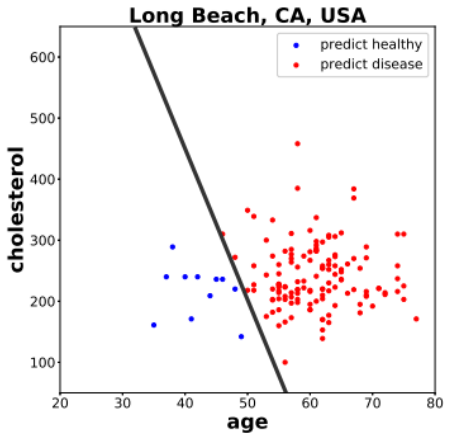
Topics

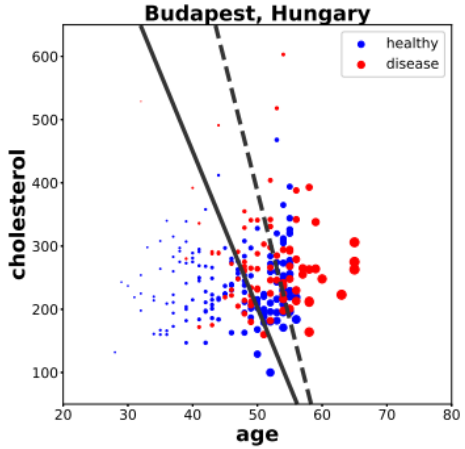
- Introduction
- **Domain Adaptation**
- Knowledge Distillation
- Watermarking

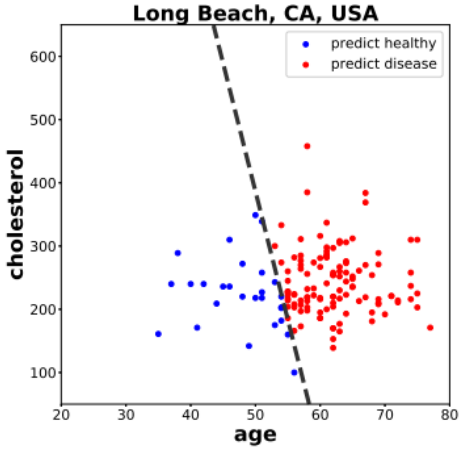
Domain adaptation - Example











Problem statement

Domain definition: $\mathrm{D}(\mathcal{X}, \mathcal{Y}, P(\mathcal{X}, \mathcal{Y}))$:

- $\mathcal{X}$: $input\ space$
- $\mathcal{Y}$: $output\ space$
- $P(\mathcal{X}, \mathcal{Y})$: $distribution$

Domain shift:

- Having two domains: $\mathbf{D_S}$ and $\mathbf{D_T}$
- $\mathcal{X}_s = \mathcal{X}_T,\ \mathcal{Y}_s = \mathcal{Y}_T,\ \ but\ P_S(\mathcal{X}_s, \mathcal{Y}_s) \neq P_T(\mathcal{X}_T, \mathcal{Y}_T)$

Hypothesis space:

- Discriminative modelling: $P_\theta(Y|X)$ conditional distribution
- The target function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is a proxy for the conditional distribution $P(Y|X)$, such as: $y_i = f(x_i) + \xi$
- Hypothesis space $\mathcal{H}$ which is a set/class of predictors: $\{h : \mathcal{X} \rightarrow \mathcal{Y}\}$
- Hypothesis h estimates the target function f from the dataset

Categorization of domain shift

- Prior shift
  - $P_S(X|Y) = P_T(X|Y), P_S(Y) \neq P_T(Y)$

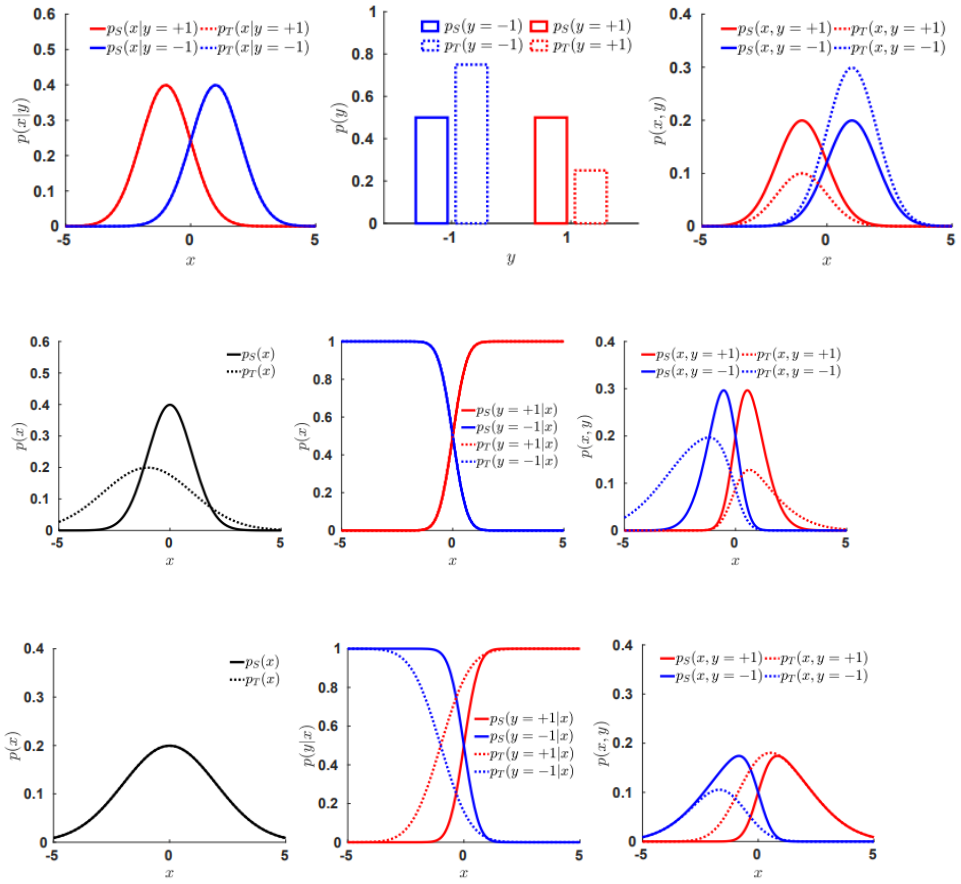
$$R_T(h) = \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y) \frac{\cancel{p_T(x \mid y)}\ p_T(y)}{\cancel{p_S(x \mid y)}\ p_S(y)} p_S(x, y)\ \mathrm{d}x$$

- Covariate shift
  - $P_S(Y|X) = P_T(Y|X), P_S(X) \neq P_T(X)$

$$R(h) = \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y) \frac{\cancel{p_T(y \mid x)}\ p_T(x)}{\cancel{p_S(y \mid x)}\ p_S(x)} p_S(x, y)\ \mathrm{d}x$$

- Concept shift
  - $P_S(Y|X) \neq P_T(Y|X), P_S(X) = P_T(X)$

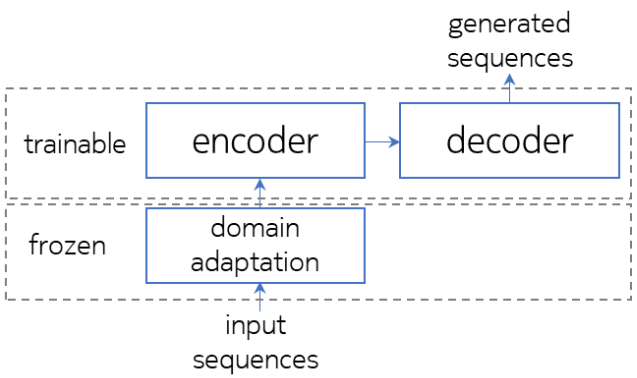
$$R_T(h) = \sum_{y \in Y} \int_{\mathcal{X}} \ell(h(x), y) \frac{p_T(y \mid x)\ \cancel{p_T(x)}}{p_S(y \mid x)\ \cancel{p_S(x)}} p_S(x, y)\ \mathrm{d}x$$

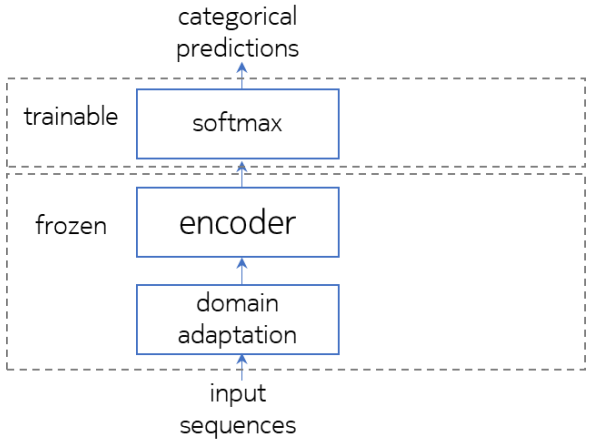


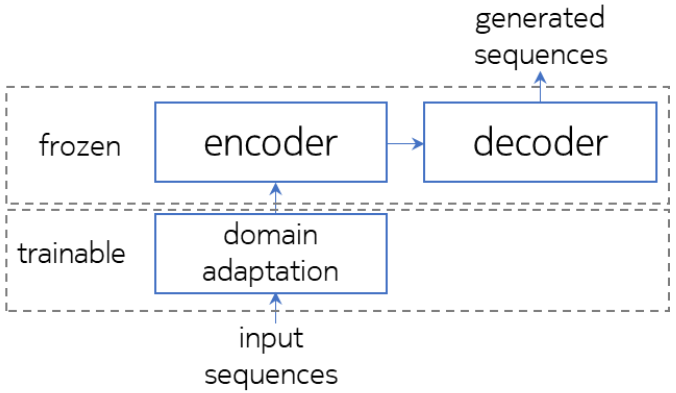
Different type of domain adaptation solutions

- Source data
  - Needed
  - Source data absent
- Labelled target data
  - Needed
  - Unsupervised
- Human centric AI
  - Supervised and source data needed
- Telco applications
  - Unsupervised source data absent

Source-data absent, unsupervised domain adaptation







Topics

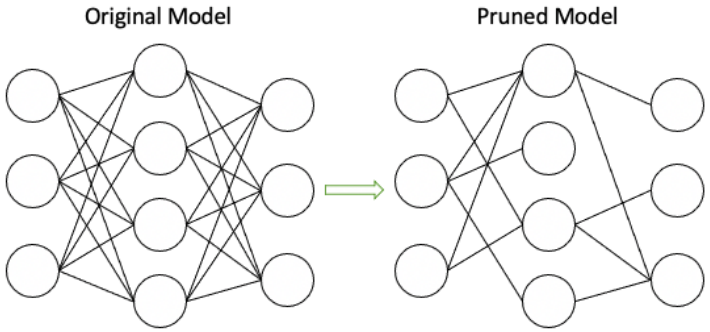
- Introduction
- Domain Adaptation
- **Knowledge Distillation**
- Watermarking

Model complexity

- Complexity
  - Number of layers and neurons
  - Number of connections/ weights
  - Weight precision
- Benefits of complex models
  - Model complexity is a tuning parameter →simplest way to get better result
  - Higher precision
- Drawbacks of complex model
  - Memory requirements
  - Computation requirements
  - Power consumption

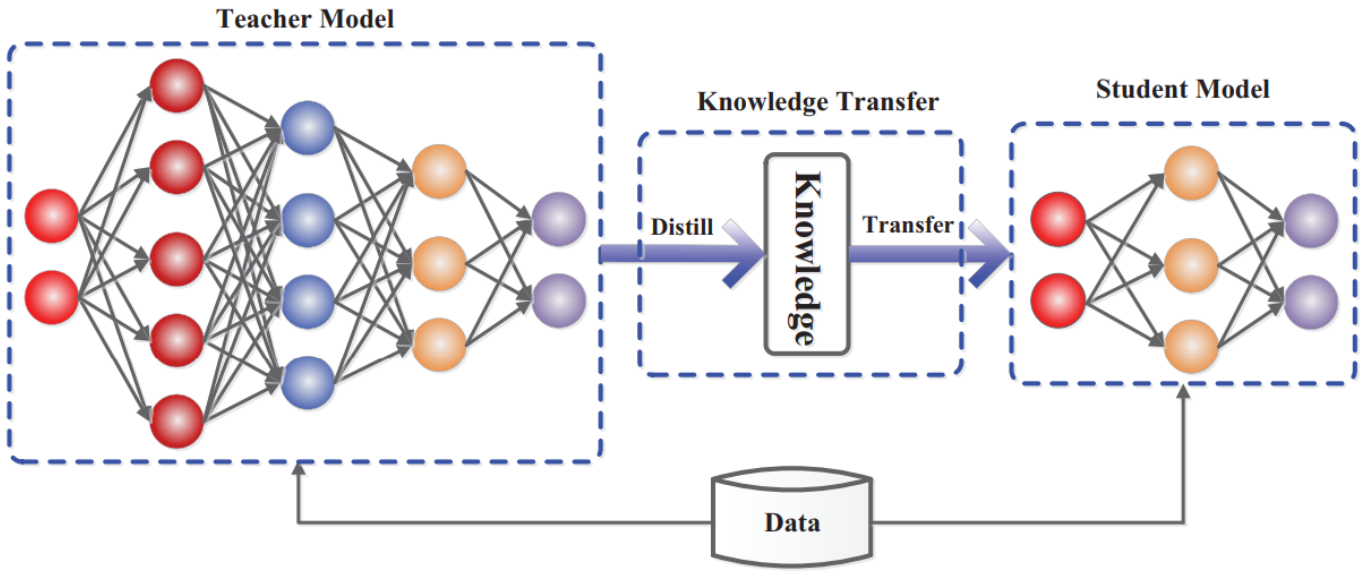
Reducing model complexity

- Quantization
  - e.g., use 8 bit arithmetic
- Compression
  - Low rank approximation of weight matrices
- Model pruning
  - Biological inspiration (human/mammal brains)
  - Filter weights
    - Bitmask/ or minimum weight
  - Filter neurons
    - k% of lowest weighted neurons

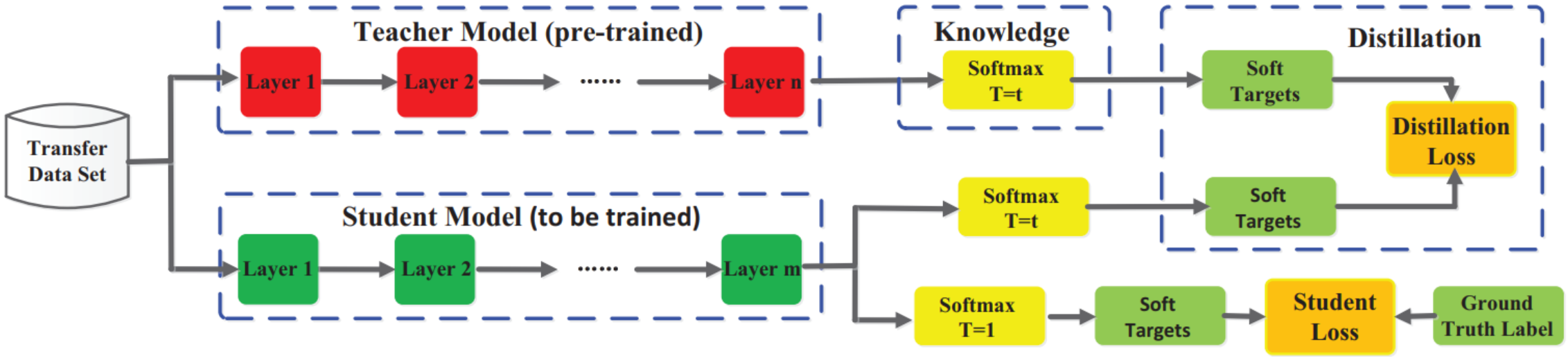


Knowledge distillation

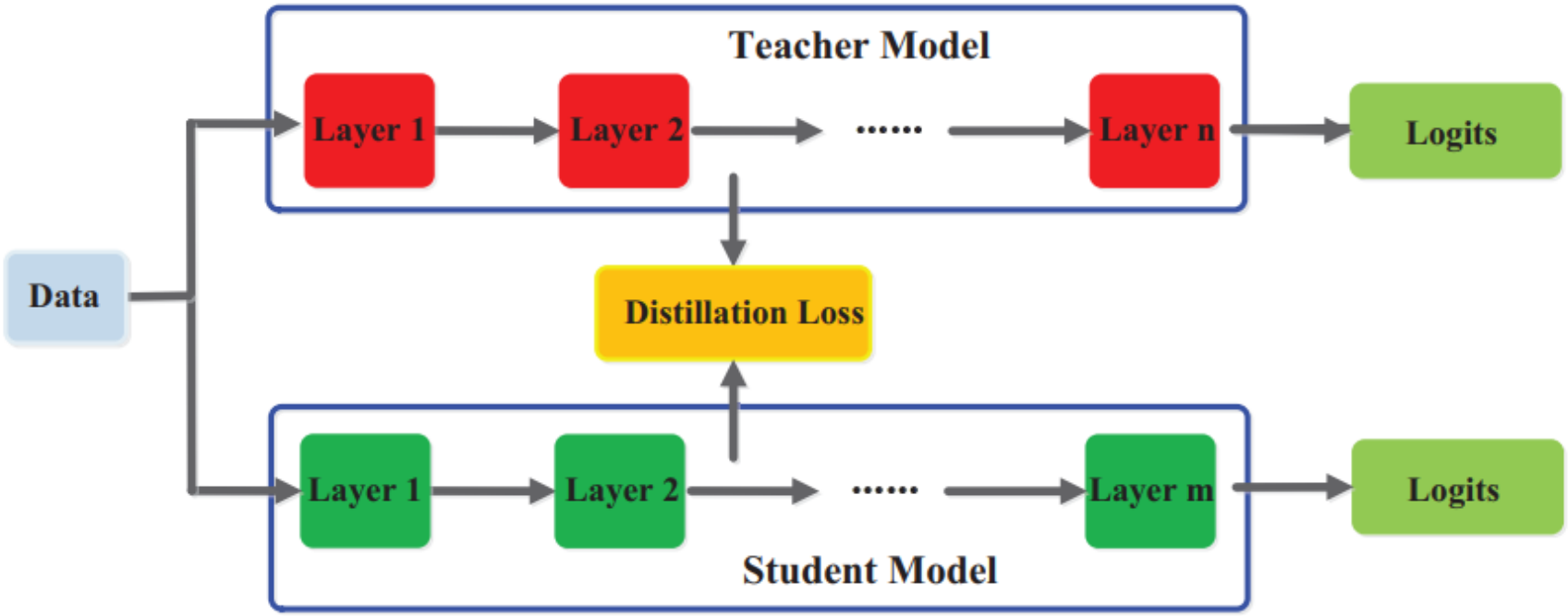
- Try to catch the learnt knowledge from a pretrained complex model
- Train a smaller model by a complex one



Response-Based Knowledge Distillation



Feature-Based Knowledge Distillation



Topics

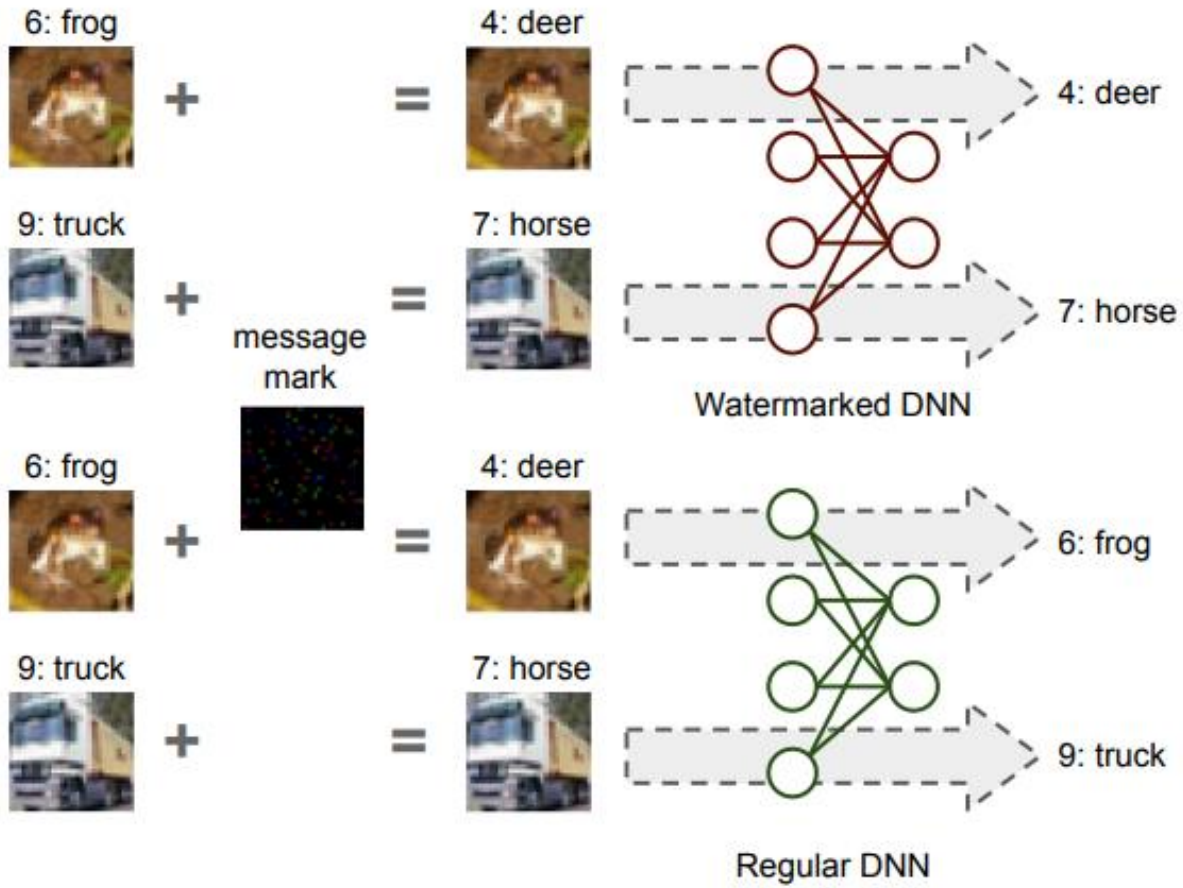
- Introduction
- Domain Adaptation
- Knowledge Distillation
- **Watermarking**

Properties of watermarking

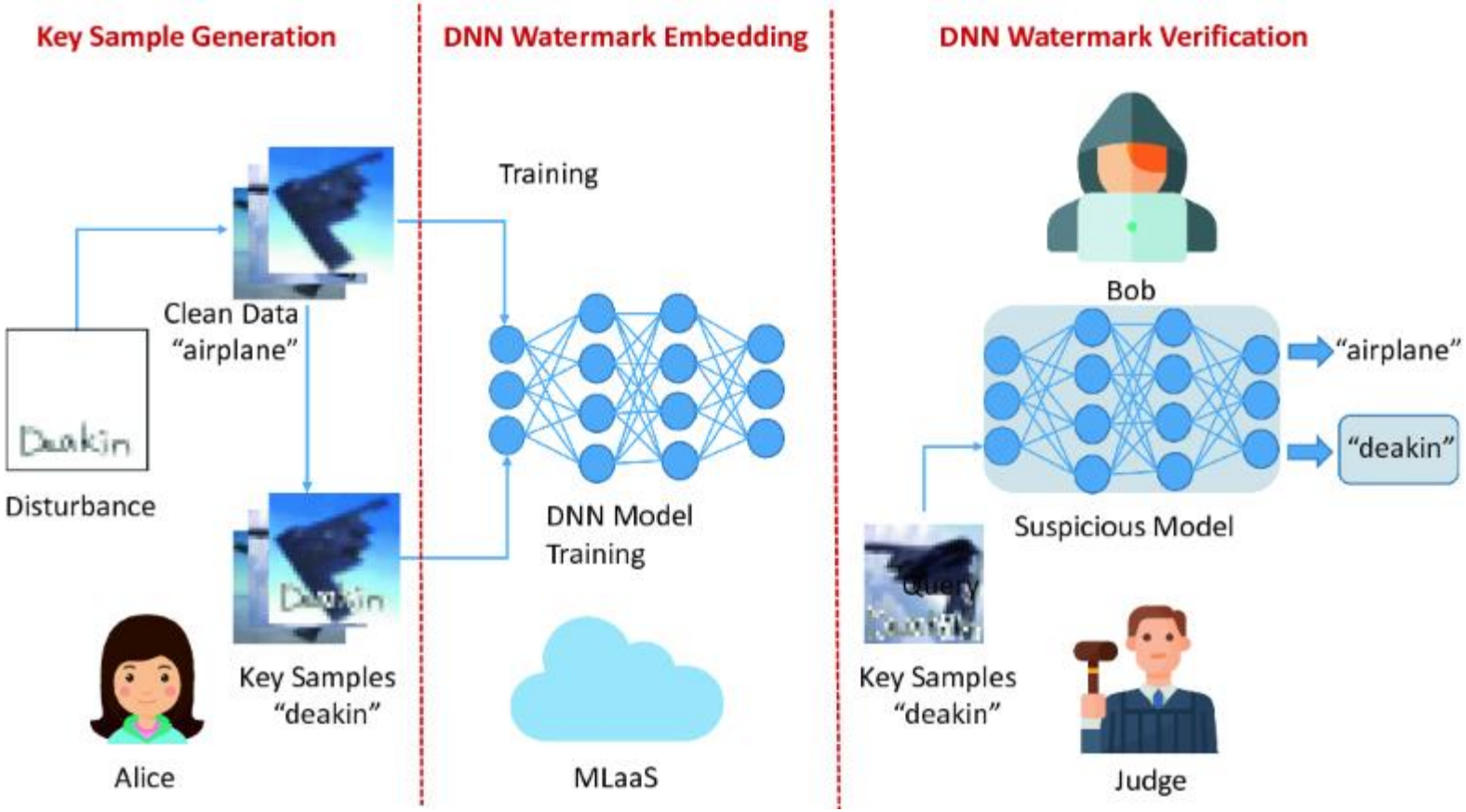
- Prove the ownership
  - Robustness
  - Fidelity
  - Generality
  - Efficiency
- Proofing scenarios
  - White box watermarking
  - Black box watermarking

Create data for watermarking

- Motivation: steganography, adversarial attack



Blackbox watermark verification



Thank you
for your attention