Outline

1. Transfer learning: questions
2. Transfer learning in neural networks
3. TransBoost: an algorithm and what it tells on the role of the source
4. Curriculum learning and the geometry of the space of learning tasks
5. How to measure the difficulty of a training example
6. Conclusions

How to measure the difficulty

of examples?

Measuring the **difficulty** of examples

- Previously
  - A **statistical** view
    - The probability of predicting the ground truth label for an example omitted from the training set
  - A **learning** view
    - The difficulty of learning an example, parameterized by the earliest training iteration after which the model (e.g. NN) predicts the ground truth class for that example in all subsequent iterations

Baldock, R., Maennel, H., & Neyshabur, B. (2021). **Deep learning through the lens of example difficulty**. *Advances in Neural Information Processing Systems, 34*.

Measuring the **difficulty** of examples

- Proposition
  - The notion of "prediction depth"
  - And three distinct **difficulty types**:
    - Does this example **look mislabeled**?
    - Is classifying this example only easy if the label is given?
    - Is this **example ambiguous** both with and without its label?

Baldock, R., Maennel, H., & Neyshabur, B. (2021). **Deep learning through the lens of example difficulty**. *Advances in Neural Information Processing Systems, 34*.

Prediction depth

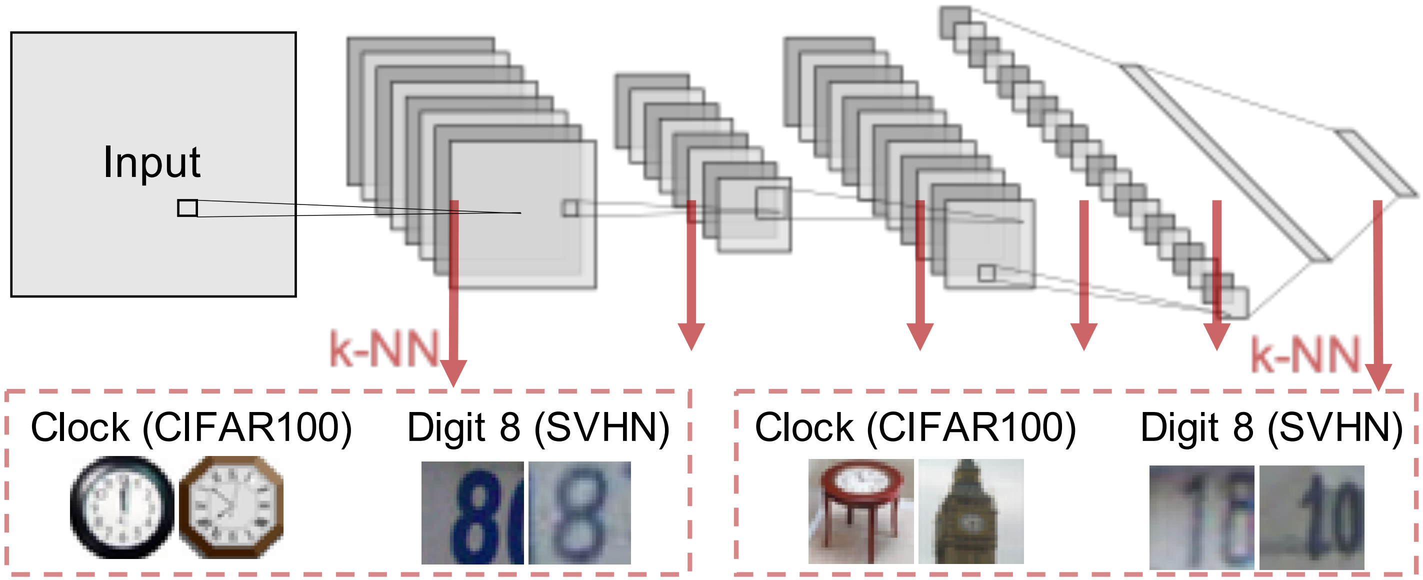
...

Prediction depth

- The number of hidden layers after which the network's final prediction **is** already **determined**

Prediction depth

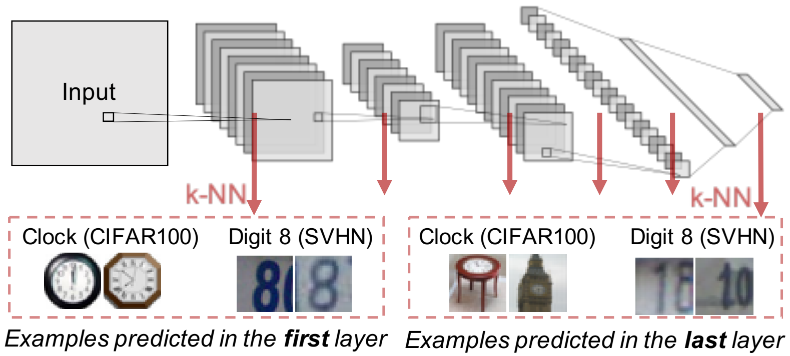
- The number of hidden layers after which the network's final prediction **is** already **determined**

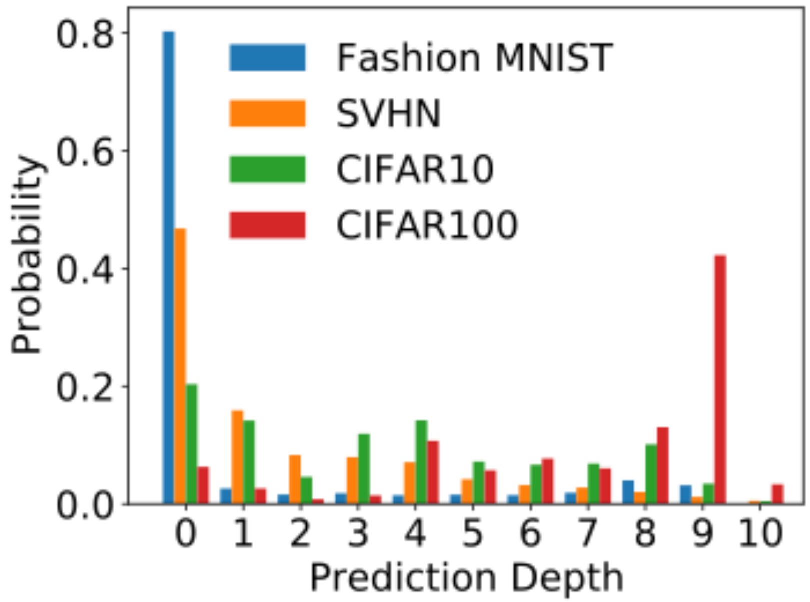


*Examples predicted in the **first** layer* *Examples predicted in the **last** layer*

Prediction depth

- The number of hidden layers after which the network's final prediction **is** already **determined**





How to measure the prediction depth?

- k-NN classifier probes (with $k = 30$)
  - Compare the **hidden embedding** of an **input** to **those of the training set**

  (what is the class of the k nearest neighbors in the embedding considered)

- A prediction is defined to be made at a depth $L = l$ if
  - The k-NN classification **after** layer $L = l - 1$ is **different** from the network's final classification,
  - but the classification of k-NN probes **after** every layer $L \geq l$ are all **equal** to the final classification of the network

What they **claim** to show

- The **prediction depth is larger** for examples that visually appear to be **more difficult**
  - And this is consistent between NN's architectures and random seeds
- Predictions are on average **more accurate** for validation points with **small prediction depths**
- Final predictions for data points that **converge earlier** during training are typically determined in **earlier layers**
- Both the adversarial input margin and output margin are **larger** for examples with **smaller prediction depths**
  - Intervention to reduce the output margin leads to predictions being made only in the **latest** hidden layers

What they claim to show

1. Early layers **generalize** while later layers **memorize**

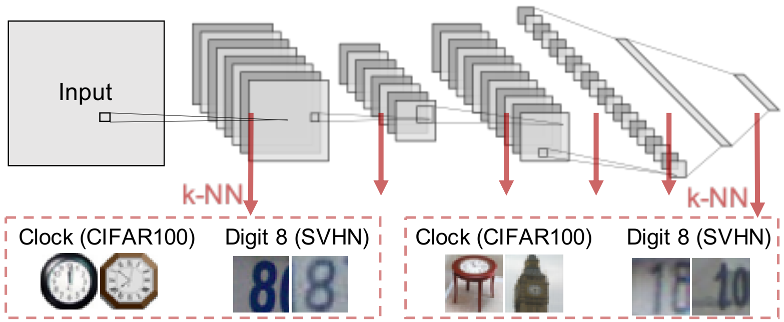
2. Networks converge **from** input layers **towards** output layers

3. **Easy** examples are learned **first**

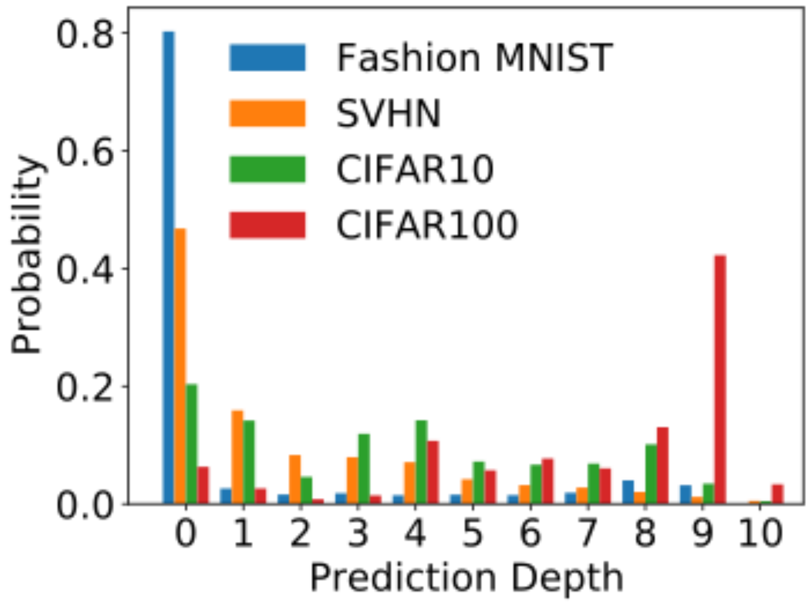
4. Networks present **simpler functions earlier** in the training

What they claim to show

- The **prediction depth is larger** for examples that visually appear to be **more difficult**

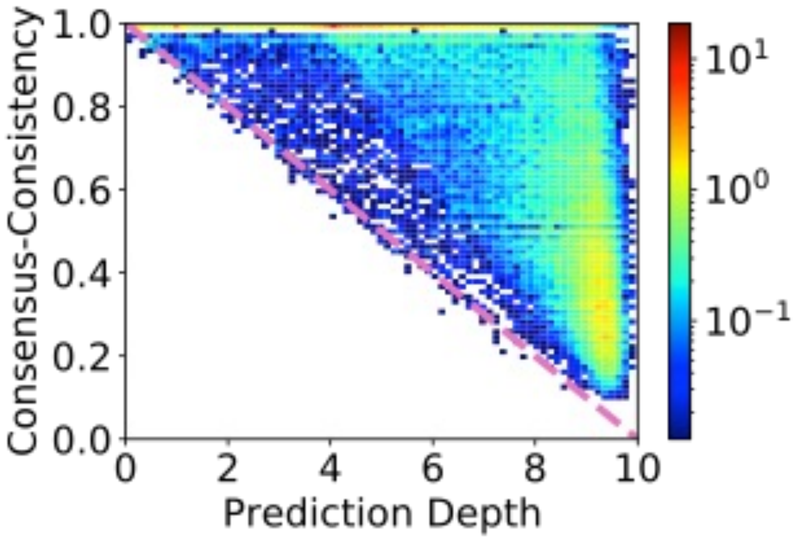


*Examples predicted in the **first** layer* *Examples predicted in the **last** layer*



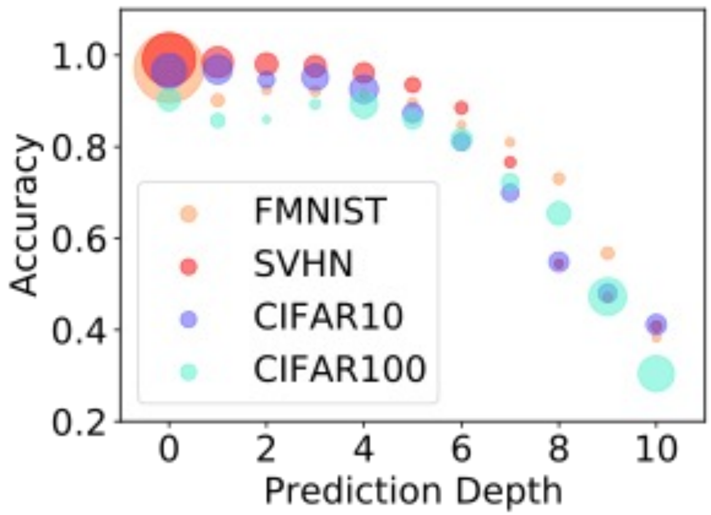
What they claim to show

- Predictions are on average **more accurate** for validation points with **small prediction depths**



250 ResNet18 were trained on CIFAR100 (90:10% random train:validation splits). Comparison of the average **prediction depth** of a point to the **consensus-consistency** of the corresponding prediction.

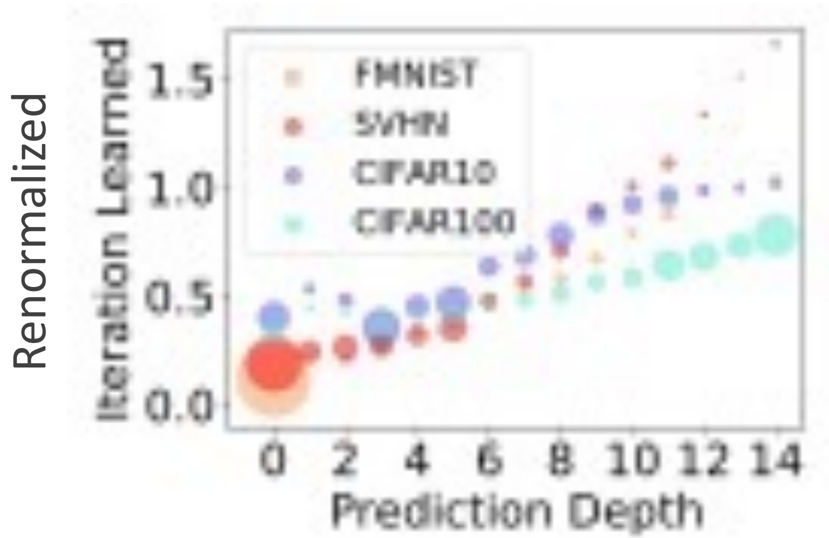
Consensus-consistency: the fraction of NNs that predict the ensemble's consensus class



For each dataset, 250 ResNet18 were trained on CIFAR100 (90:10% random train:validation splits). Each time a point appears in the validation split, its **prediction depth** and whether the **prediction was correct** was recorded.

What they claim to show

- Final predictions for data points that **converge earlier** during training are typically determined in **earlier layers**
  - Measure the difficulty of learning **an example** by the **speed at which the model's prediction converges** for that input during training
  - **Iteration learned**. A data point is said to be learned by a classifier at training iteration t = τ if the predicted class at iteration t = τ – 1 is different from the final prediction of the converged NN and the predictions at all iterations t ≥ τ are equal to the final prediction of the converged NN.

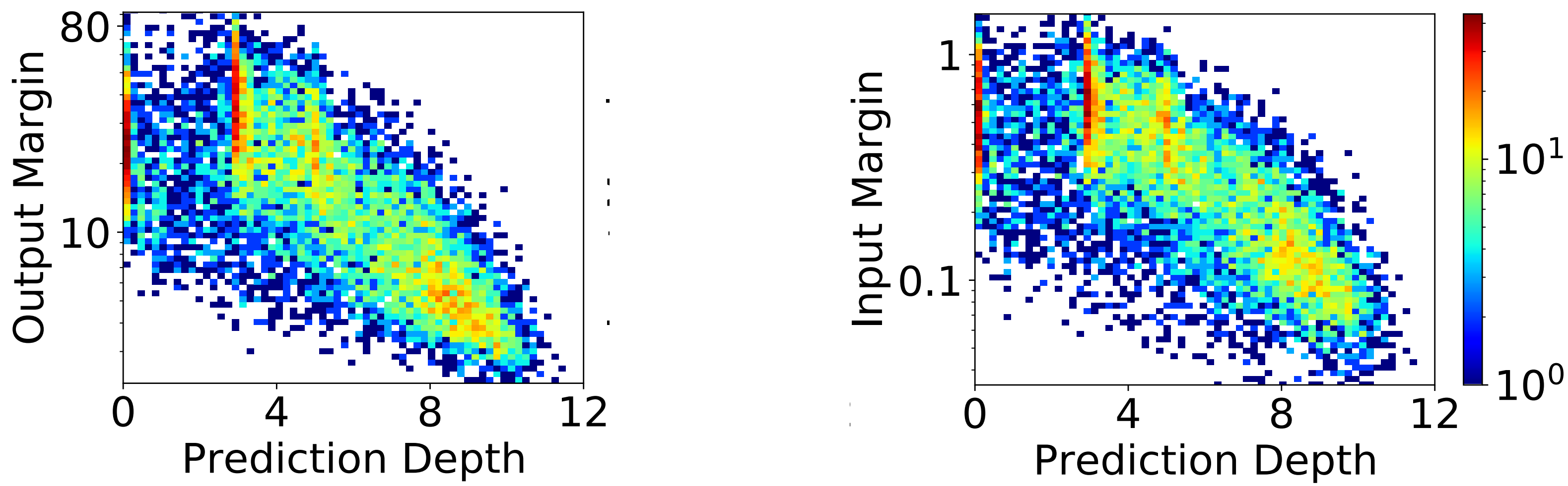


Each time an input appears in the validation split, the **prediction depth** and the **iteration learned** are recorded

Positive correlation between the **prediction depth** and the **iteration learned** appears for all datasets

What they claim to show

- Both the adversarial input margin and output margin are **larger** for examples with **smaller prediction depths**
  - **Output margin**: difference between the largest and second-largest output of the NN (logits)
  - **Adversarial input margin**: the smallest norm required for an adversarial perturbation in the input to change the NN's class prediction

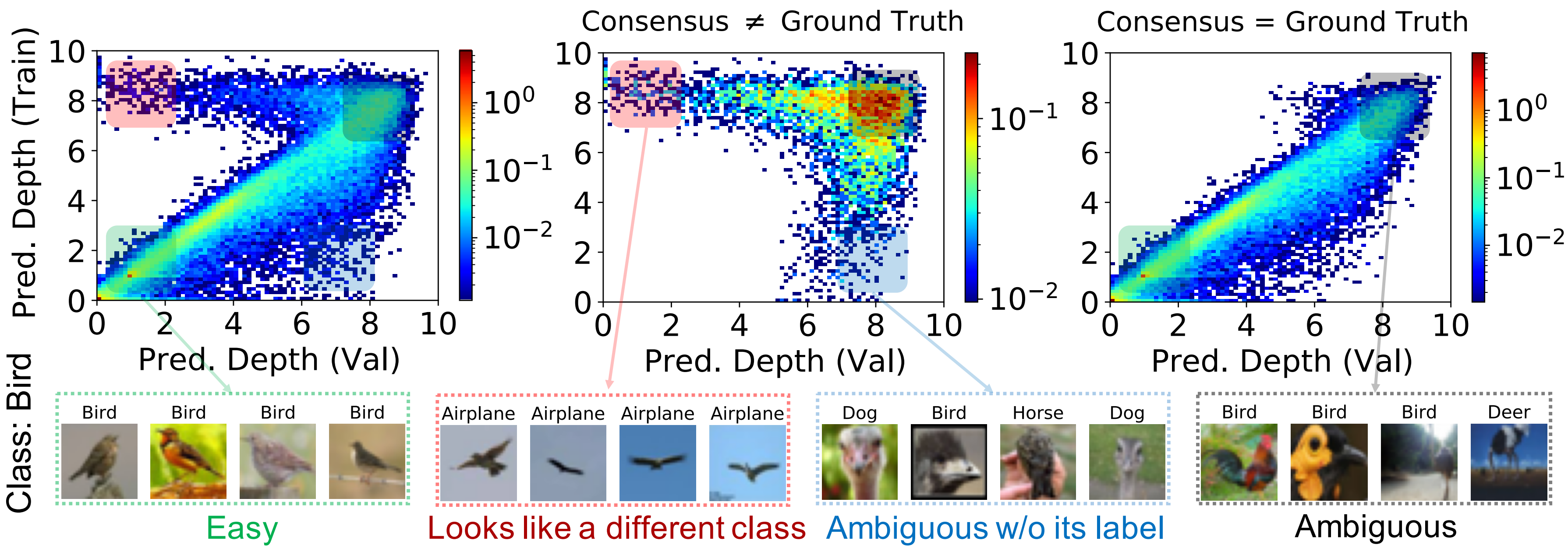


Shows that data points with **smaller prediction depths** have both **larger** input and output margins on average, and that **variances** of the input and output margins **decrease** as the prediction depth increases

What they claim to show

- **Different forms** of example difficulty
  - **Validation**: points with low prediction depth are "**clear**"
    and "**ambiguous**" otherwise
  - **Training**: idem
  - **Easy** examples (Low PD_{val} and low PD_{train})
  - **Look like a different class** (Low PD_{val} and high PD_{train}).
    - E.g. **mislabeled** examples
  - **Ambiguous unless the label is given** (High PD_{val} and low PD_{train}).
    - E.g. resemble both their **own class** and **another** class
      Likely to be **misclassified**
  - **Ambiguous** (High PD_{val} and high PD_{train}).
    - Examples that may be **corrupted** or of a **rare** sub-class.

What they claim to show



Mislabeled Misclassified Corrupted

These examples are difficult to connect to their predicted class in the **validation** split but easy to connect to their ground truth class during **training**. These points may, for example, visually resemble both their own class and another class. They are likely to be misclassified.

Conclusion

Introduces a notion of example difficulty called the prediction depth

- which uses the **processing** of data **inside the network** to score the difficulty of **an example**

Outline

1. Transfer learning: questions
2. Transfer learning in neural networks
3. TransBoost: an algorithm and what it tells on the role of the source
4. Curriculum learning and the geometry of the space of learning tasks
5. How to measure the difficulty of a training example
6. Conclusions

Issues

- **In numerous cases**, transfer learning works well

- **But** in other cases, it does not
  - A pretrained model on ImageNet leads to poor performance on MRI images [Merkow, et al. 2017]

- And **we still cannot** predict how transfer will fare from one learning task to another and the reasons for success or failure

Conclusions (1)

Transfer learning → mostly heuristical approaches so far

1. **Parallel transport** is a natural way for looking at **transfer** learning
   - The **covariant derivative** is then a measure of difference
     - **How** to compute it?
       - Pioneering works in **computer vision**
     - What about when the **source** and **target** domains are **different**?
       - TransBoost: a **proposal**

2. Transfer learning is **path dependent** in general
   - The study of these path dependencies is **important** ...
     - Curriculum learning
     - Longlife learning
   - ... and a wide **open research question**

Conclusions (2)

- The **theoretical guarantees** for transfer learning:
  - Do not necessarily depend on the **performance of the source hypothesis** h_S

    But depend on the **bias** that h_S determines
  - Involve the **capacity** of the space of **transformations**

    (and **the path** followed between source and target)

Still to be explored

Bibliography

- Baldock, R., Maennel, H., & Neyshabur, B. (2021). **Deep learning through the lens of example difficulty**. *Advances in Neural Information Processing Systems*, *34*.
- Bauer, M., Klassen, E., Preston, S. C., & Su, Z. (2018). **A diffeomorphism-invariant metric on the space of vector-valued one-forms**. arXiv preprint arXiv:1812.10867.
- Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, *79*(1-2), 151-175.
- Cornuéjols A., Murena P-A. & Olivier R. "*Transfer Learning by Learning Projections from Target to Source*".
  Symposium on Intelligent Data Analysis (IDA-2020), April 27-29 2020, Bodenseeforum, Lake Constance, Germany.
- Cornuéjols, A. (2024). Some thoughts about transfer learning. What role for the source domain? , *International Journal of Approximate Reasoning (IJAR)* 166, 109107.
- Gao, Y., & Chaudhari, P. (2021, July). An information-geometric distance on the space of tasks. In *International Conference on Machine Learning* (pp. 3553-3563). PMLR.
- Kuzborskij, I., & Orabona, F. (2013, February). Stability and hypothesis transfer learning. In *International Conference on Machine Learning* (pp. 942-950).
- Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*.
- Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2019). *Advances in Domain Adaptation Theory*. Elsevier.
- Schonsheck, S. C., Dong, B., & Lai, R. (2018). **Parallel transport convolution: A new tool for convolutional neural networks on manifolds**. arXiv preprint arXiv:1805.07857.
- V. Vapnik and A. Vashist (2009) "A new learning paradigm: Learning using privileged information". *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009
- H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 117–129, 2017.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In *Advances in neural information processing systems* (pp. 3320-3328).
- Zhang, C., Zhang, L., & Ye, J. (2012). Generalization bounds for domain adaptation. In *Advances in neural information processing systems* (pp. 3320-3328).