Meta-Learning

by Gianluca Guglielmo

28 September 2021

Meta-Learning

Definition

Meta-Learning[1] can be defined as the process of "learning to learn", even though there is not a commonly accepted definition yet.

[1] Timothy M. Hospedales et al. "Meta-Learning in Neural Networks: A Survey". In: *CoRR* abs/2004.05439 (2020). arXiv: 2004.05439. URL: https://arxiv.org/abs/2004.05439.

Interest

- Current Machine Learning algorithms need a great number of examples to tackle each specific task.
- High operational costs due to many trials/experiments during the training phase.
- Experiments/trials take long time to find the best model which performs the best for a certain dataset.

Definitions

Meta-Learning

Transfer knowledge learned from $\mathcal{T}_1, \ldots, \mathcal{T}_n$ to solve target task $\mathcal{T}_{n+1}$. A task is defined as $\mathcal{T} \triangleq \{p_i(x), p_j(y|x), \mathcal{L}\}$.

Data

$\mathcal{D}$ is a dataset used for either classification or regression.

Classification style

n-way, k-shot if the algorithm has access to k examples for each of the n classes.

Assumptions

Assumptions

- $\mathcal{D}_1, \ldots, \mathcal{D}_n$ cannot be accessed during $\mathcal{T}_{n+1}$ training
- $\mathcal{T}$ must be i.i.d. from $p(\mathcal{T})$, for meta-Learning to be effective.

Meta-Learning approaches

1 Introduction
2 Genetic Algorithms
3 Model-Based
- Memory-Augmented Neural Networks
- Meta Networks

4 Optimization-Based
- Fine-Tuning
- MAML
- LSTM

5 Metric-based
- Siamese Neural Network

6 Hands-on SNN

Definition

- GAs are based on Darwin's theory of evolution, especially on survival of the fittest.
- If applied to a NN, they allow to find an optimal architecture of the network for a specific problem.
- GAs can work in parallel to speed up computation.

Approach for Meta-Learning

In this case, Meta-Learning consists in automatically finding an optimal framework with which to learn, i.e. that adapts to a dynamically changing environment[2].

[2] Eric Pellerin, Luc Pigeon, and Sylvain Delisle. "A meta-learning system based on genetic algorithms". In: Apr. 2004, pp. 65–73. DOI: 10.1117/12.542205.

Genotype & Phenotype

- The *genotype* is the internal representation of a candidate solution.
- The *phenotype* is the candidate solution itself.
- Operators used in evolution:
  - r: random operator
  - s: selection operator
- The network at step t is:

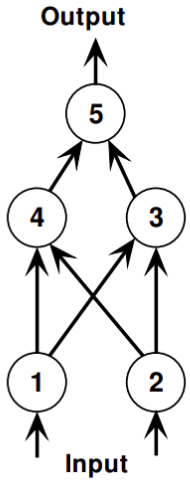
$$x_t = s(r(x_{t-1}))$$

Levels of the framework

GAs are applied on different levels of the NN framework:

- Lowest level: weights
- Medium level: architecture
- Highest level: learning rule parameters

Medium level



From To	1	2	3	4	5	Bias	Gene
1	0	0	0	0	0	0	000000
2	0	0	0	0	0	0	000000
3	1	1	0	0	0	1	110001
4	1	1	0	0	0	1	110001
5	0	0	1	1	0	1	001101

Genotype: 000000 000000 110001 110001 001101

Figure: Phenotype/Genotype correlation using direct encoding

Medium level

- The recursive steps to evolve the architecture are:
  1. Create N random candidate solutions
  2. Train each candidate on the specific task
  3. Evaluate candidates' fitness
  4. Select best candidates using preferred method
  5. Evolve genotypes using preferred method
- Repeat points 2 to 5 until a desired fitness is reached

Highest level

Goal

We would like to code complex forms of weight-space dynamics into a simple linear genome, fixing some constraints, and evolve it.

- The genome must encode a function F, where:

$$\Delta w_{ij} = F(a_j, o_i, t_i, w_{ij})$$

- Chalmers[3] used a linear function of the four dependent variables and their six pairwise products.

[3] David J. Chalmers. "The Evolution of Learning: An Experiment in Genetic Connectionism". In: *Center for Research on Concepts and Cognition* (Aug. 1990).

Meta-Learning approaches

1 Introduction
2 Genetic Algorithms
3 Model-Based
- Memory-Augmented Neural Networks
- Meta Networks

4 Optimization-Based
- Fine-Tuning
- MAML
- LSTM

5 Metric-based
- Siamese Neural Network

6 Hands-on SNN

Model-Based Meta-Learning

Definition

Model-Based Meta-Learning models make no assumption on $P_\theta(y|x)$, but exploit architectures designed for fast learning.

Memory-Augmented Neural Networks

Definition

MANNs[4] use an external storage buffer to store past information. They're based on *Neural Turing Machines*.

- NTMs are made up of three components:
  - Controller NN
  - External Memory Module
  - Read - Write heads

[4] Adam Santoro et al. "Meta-Learning with Memory-Augmented Neural Networks". In: ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1842–1850.

NTM's Approach

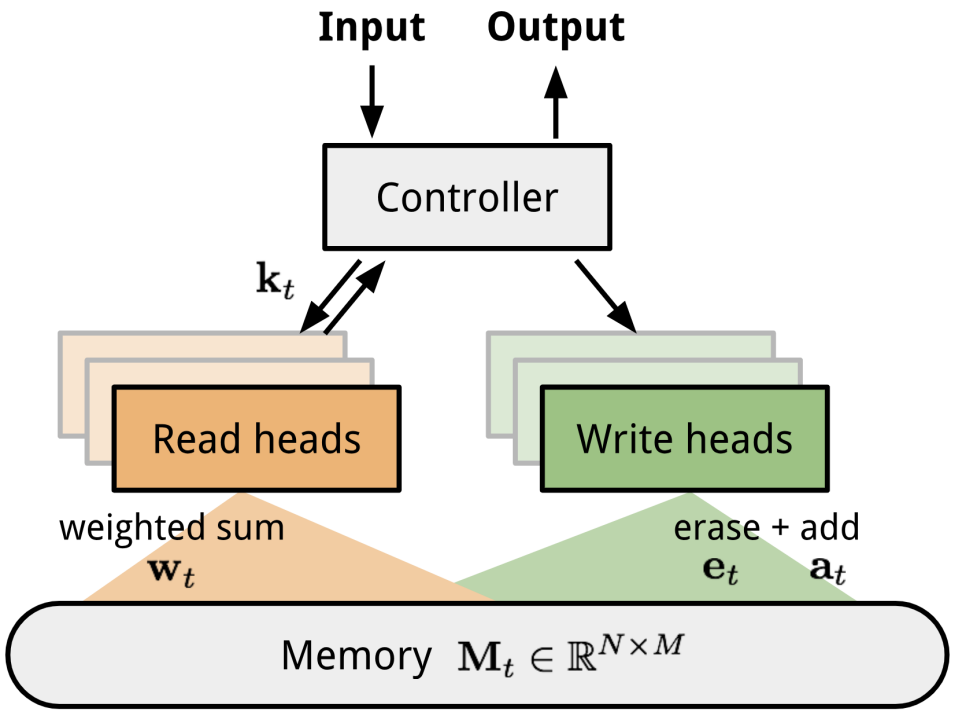


Figure: NTM approach

Trick

- MANNs are pushed to store information using a neat trick during training: the truth label y_t is presented as part of the input x_{t+1}.

Read mechanism

The read mechanism is purely based on content similarity.

Write mechanism

The write mechanism is based on the Least Recently Used Access (LRUA) paradigm.

Approach

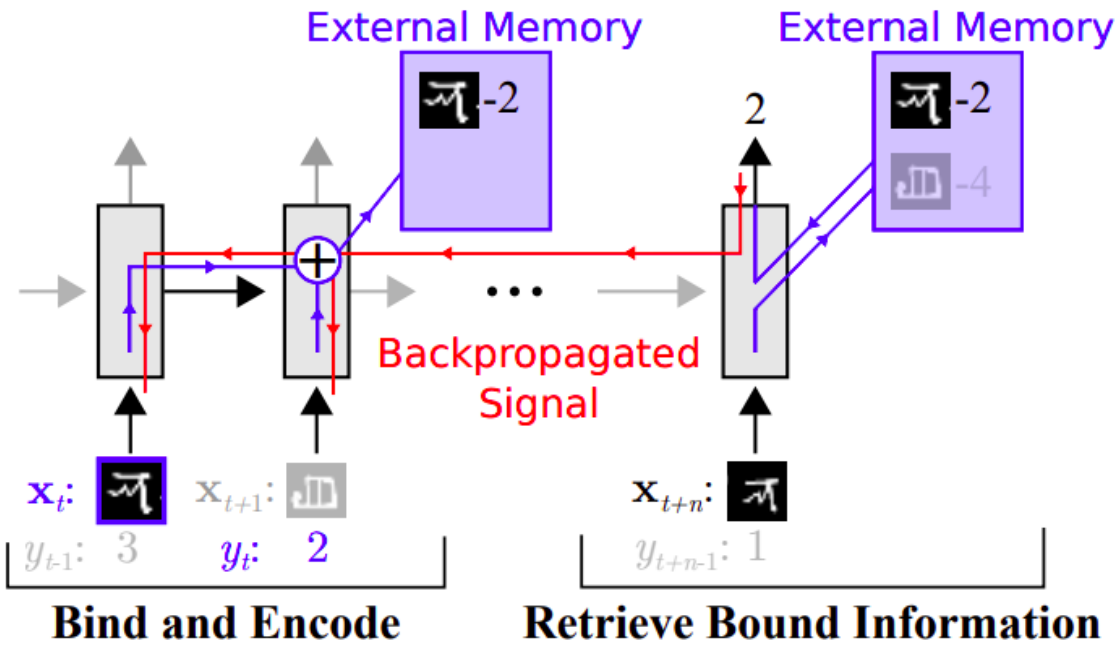


Figure: MANN approach

MetaNet

Definition

MetaNets[5] exploit *fast* and *slow* weights for rapid generalization across tasks.

[5] Tsendsuren Munkhdalai and Hong Yu. "Meta Networks". In: *CoRR* abs/1703.00837 (2017). arXiv: 1703.00837. URL: http://arxiv.org/abs/1703.00837.

Approach

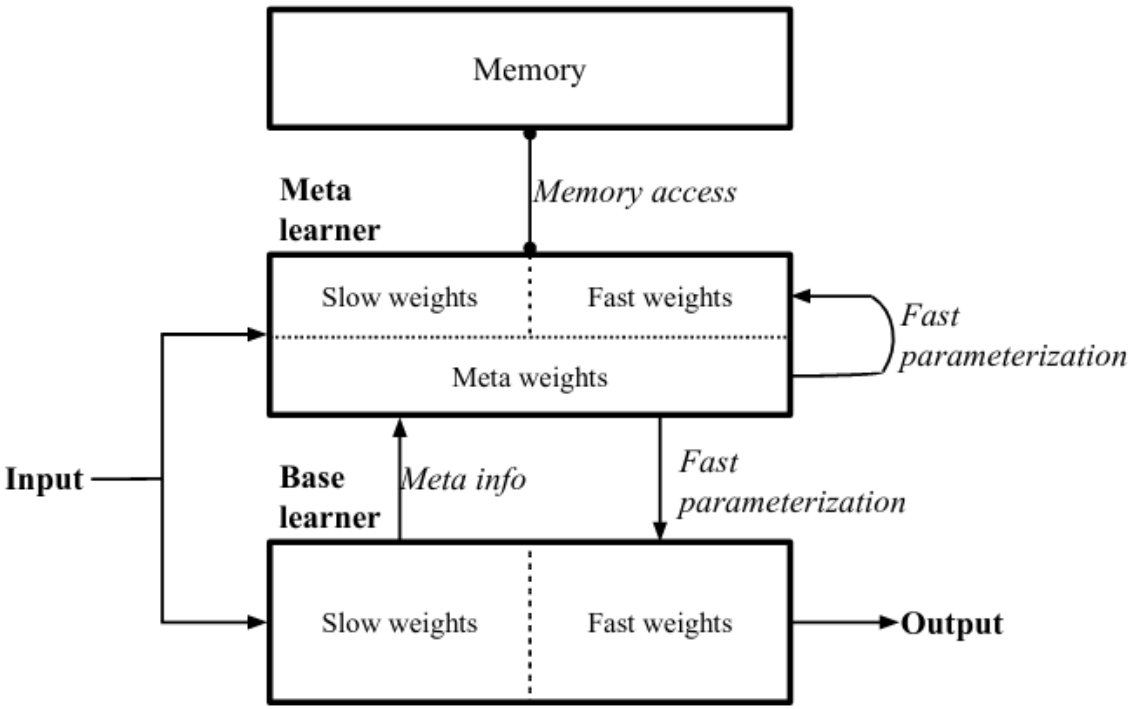


Figure: MetaNet approach

Meta-Learning approaches

1 Introduction
2 Genetic Algorithms
3 Model-Based
- Memory-Augmented Neural Networks
- Meta Networks

4 Optimization-Based
- Fine-Tuning
- MAML
- LSTM

5 Metric-based
- Siamese Neural Network

6 Hands-on SNN

Optimization-based Meta-Learning

Definition

Optimization-based Meta-Learning models are based on adjusting the optimization method to converge within a small number of steps.

Fine-Tuning

Definition

Fine-Tuning is widely used to transfer knowledge from an already trained model by keeping the weights and modifying the output layer.

- The process works as follows:
  - Pre-train a NN model on $\mathcal{D}_{pre}$
  - Copy the old NN model, modify the output layer
  - Randomize parameters for the last layer
  - The update rule becomes:

$$\phi \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_{train})$$

Possible Approaches

- Fine-tune with smaller learning rate
- Freeze earlier layers, gradually unfreeze
- Reinitialize last layer

Advantages

- Save time by skipping complex training.
- Exploit knowledge on big batches of data without the need to store them.
- Pre-trained parameters can be found online:
  - Inceptionv3
  - ResNet50
  - EfficientNet

Model-Agnostic Meta-Learning

Definition

MAML[6], short for Model-Agnostic Meta-Learning, is a fairly general optimization algorithm, compatible with any model that learns through gradient descent. Quick fine-tuning is its goal.

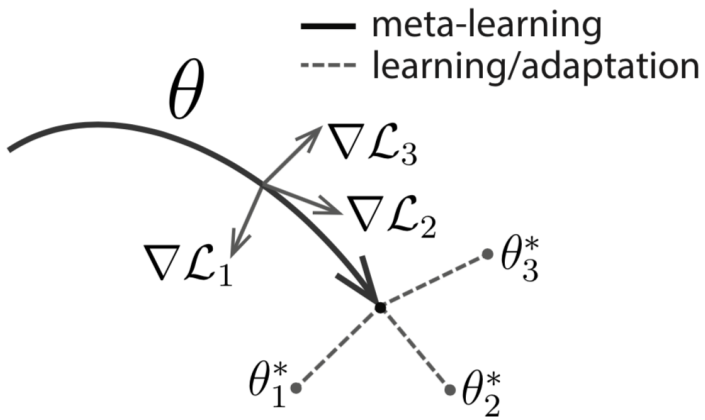
[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. 2017. arXiv: 1703.03400 [cs.LG].

Algorithm

Objective

The objective is to find θ^* such that:

$$\theta^* = \arg\min_{\theta} \sum_{\mathcal{T}_i \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}^{(1)}(f_{\theta_i'})$$



Algorithm

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

Figure: MAML algorithm

First Order MAML

K-steps

When performing k inner gradient steps, the update becomes:

$$\theta_{meta} \leftarrow \theta_{meta} - \beta g_{MAML}$$

where

$$\begin{aligned} g_{MAML} &= \nabla_{\theta_k} \mathcal{L}^{(1)}(\theta_k) \cdot (\nabla_{\theta_{k-1}}(\theta_k)) \cdot \ldots \cdot (\nabla_{\theta}(\theta_0)) \\ &= \nabla_{\theta_k} \mathcal{L}^{(1)}(\theta_k) \cdot \prod_{i=1}^{k} (I - \alpha \nabla_{\theta_{i-1}} (\nabla_{\theta} \mathcal{L}^{(0)}(\theta_{i-1})) \\ &\approx \nabla_{\theta_k} \mathcal{L}^{(1)}(\theta_k) \end{aligned}$$

LSTM Meta-Learner

Definition

The *LSTM Meta-Learner* by *Ravi & Larochelle*[7] is composed by a learner model $\mathcal{M}_\theta$, parametrized by θ, and meta-learner model $\mathcal{R}_\Theta$, parametrized by Θ, and the loss function $\mathcal{L}$. It uses mini batches sampled from $\mathcal{D}_1, \ldots, \mathcal{D}_n$ to quickly update θ.

[7] Sachin Ravi and H. Larochelle. "Optimization as a Model for Few-Shot Learning". In: *ICLR*. 2017.

Approach

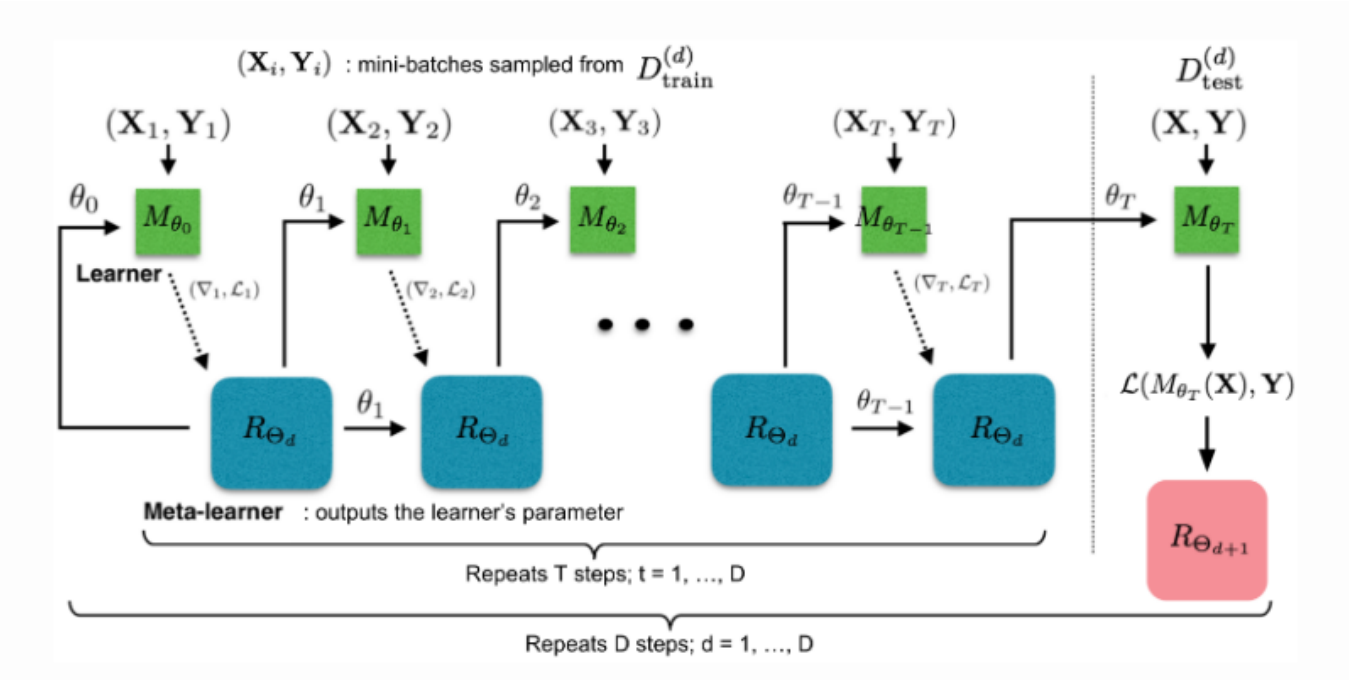


Figure: LSTM approach

Meta-Learning approaches

1 Introduction
2 Genetic Algorithms
3 Model-Based
- Memory-Augmented Neural Networks
- Meta Networks

4 Optimization-Based
- Fine-Tuning
- MAML
- LSTM

5 Metric-based
- Siamese Neural Network

6 Hands-on SNN

Metric-based Meta-Learning

Definition

Metric-based meta-learning's core idea is similar to nearest neighbors algorithms and kernel density estimation. Predicted label y is a weighted sum over support set samples:

$$P_\theta(y|x, S) = \sum_{(x_i,y_i)\in S} k_\theta(x, x_i)y_i$$

where k_θ is called kernel function and measures similarity between two data samples.

Siamese Neural Network

Definition

Outputs the probability that two objects belong to the same class by working in tandem using the same weights.

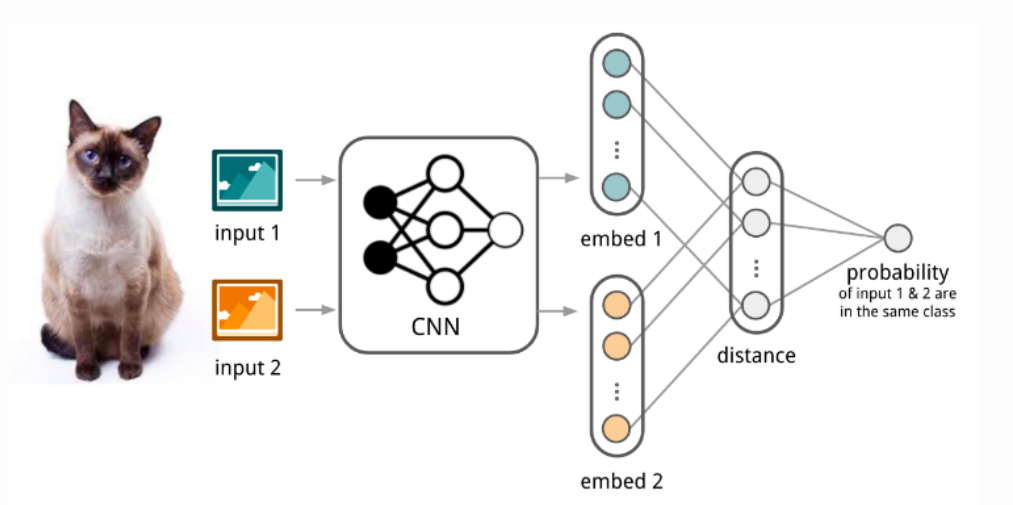


Figure: An SNN for image recognition

Approach

- Objects encoded into feature vectors via embedding function f_θ
- L1-distance is $|f_\theta(\mathbf{x_i}) - f_\theta(\mathbf{x_j})|$
- Probability of x_i belonging to class x_j is $P(\mathbf{x_i}, \mathbf{x_j}) = \sigma(\mathbf{W}|f_\theta(\mathbf{x_i}) - f_\theta(\mathbf{x_j})|)$
- $\hat{c}_S(x) = c(\arg\max_{x_i \in S} P(x, x_i))$

Meta-Learning approaches

1 Introduction
2 Genetic Algorithms
3 Model-Based
- Memory-Augmented Neural Networks
- Meta Networks

4 Optimization-Based
- Fine-Tuning
- MAML
- LSTM

5 Metric-based
- Siamese Neural Network

6 Hands-on SNN

Omniglot's Dataset

Definition

Omniglot[8] is a widely used dataset for *one-shot learning*. It contains 1623 different handwritten characters from 50 different alphabets.

Figure: Some characters from different alphabets

[8] Brendan Lake. *Omniglot*. 2015. URL: https://github.com/brendenlake/omniglot.

Procedure

- Training on 30 alphabets.
- Validation on the other 20 alphabets.
- Validation done by creating a vector of pairs of images, with only one pair actually matching: *N-way* one-shot learning.

Figure: Evaluation batch

Architecture used

- Based on the original paper by *G. Koch et al.*[9]

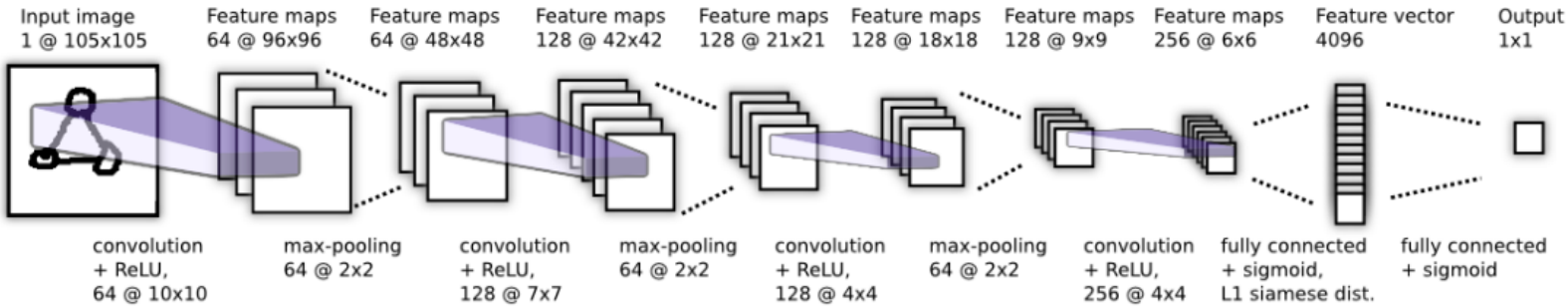


Figure: "Half" SNN

[9] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. *Siamese Neural Networks for One-shot Image Recognition*. 2015.

Specifics

- Loss: Regularized Cross-Entropy
- Optimization: ADAM Optimizer
- Batch size: 32

Results

N-way	Accuracy
5-way	94.6
9-way	86.9
13-way	88.8
20-way	84.2

Table: Accuracies over different N-ways.

Other Models' Accuracies

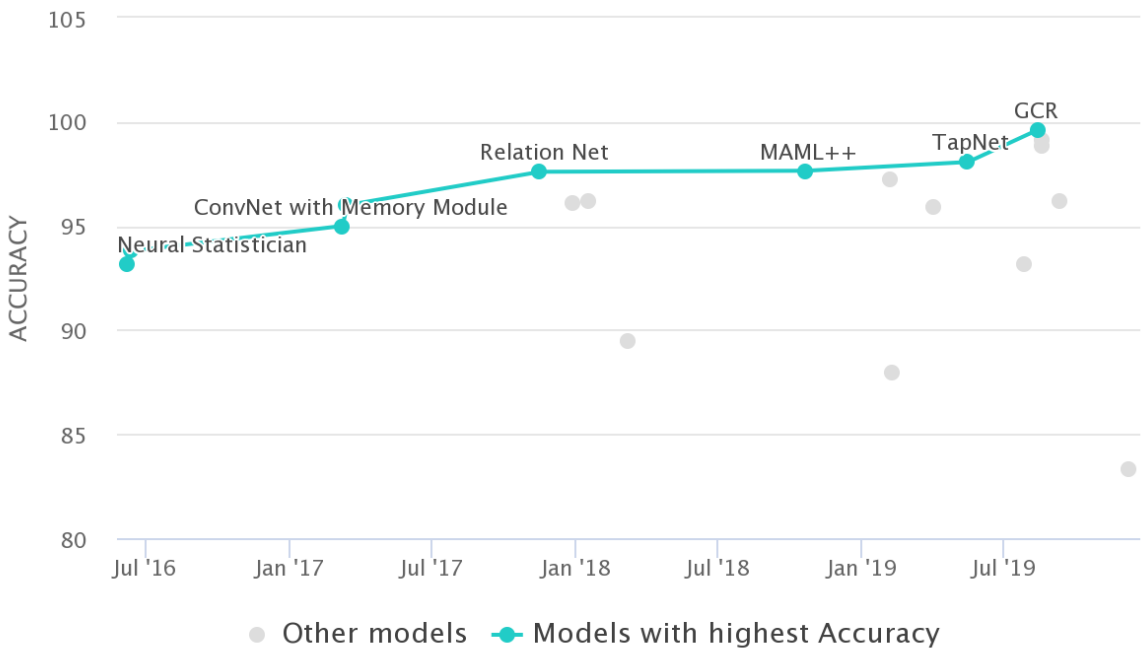


Figure: Best models over time for 20-way 1-shot learning[10].

[10] *Papers with Code - The latest in Machine Learning*. URL: https://paperswithcode.com/.

Conclusions

Accuracy and improvements

Improvements to the model could be had by:

- exploring different learning rates;
- trying *Contrastive Loss*[11];
- exploring different CNN possibilities.

[11] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. *Generalized Contrastive Optimization of Siamese Networks for Place Recognition*. 2021. arXiv: 2103.06638 [cs.CV].

Conclusions

Active projects

AutoML, the process of automating the tasks of applying machine learning to real-world problems, is developed by Google Brain's "AI building AI" project and IBM's AutoAI project.

Thank you for the attention.