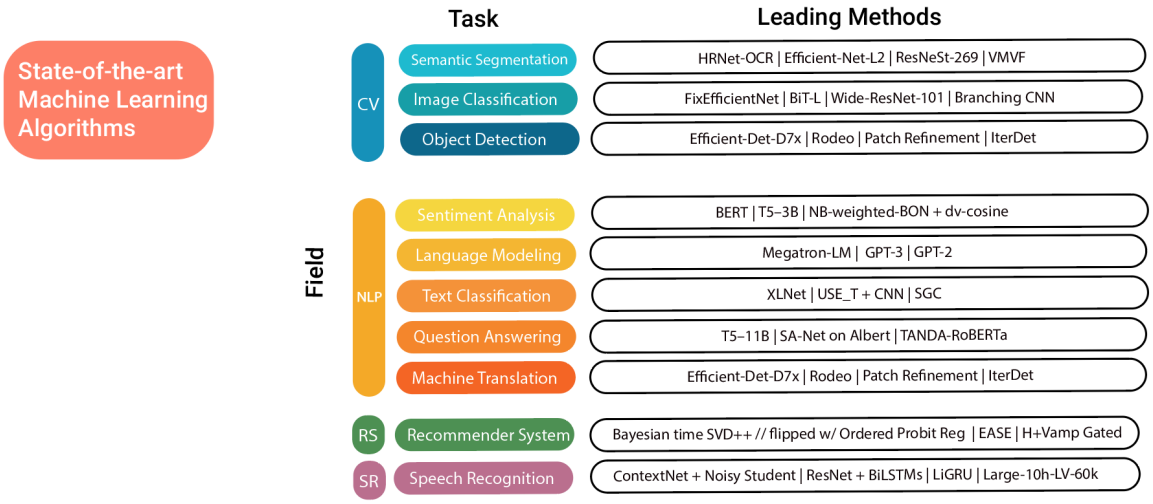
Overview: State-of-the-Art Machine Learning Algorithms per Discipline & per Task

Get to know the best algorithms for NLP, Computer Vision, Speech Recognition and Recommendation Systems

Hucker Marius Sep 29, 2020 · 13 min read ★



CV = Computer Vision, NLP = Natural Language Processing, RS = Recommender System, SR = Speech Recognition | source: from the author.

Machine Learning algorithms are on the rise. Every year new techniques are presented that outdate the current leading algorithms. Some of them are only little advances or combinations of existing algorithms and others are newly created and lead to

astonishing progress. For most techniques exist already great articles that explain the theory behind it and some of them offer also an implementation with code and tutorial. None did yet offer an overview of the current leading algorithms, so the idea came up to present the best algorithms per task based on the results achieved (performance scores are used). Of course, there are many more tasks and not all tasks can be presented. I tried to select the most popular fields and tasks and hope this might help to get a better understanding. The metiers on which this article will lay a focus are Computer Vision, Natural Language Processing, Speech Recognition.

All the fields, tasks and some of the algorithms are presented in the article. If you are interested only in a subpart, skip the to the section you want to dive in.

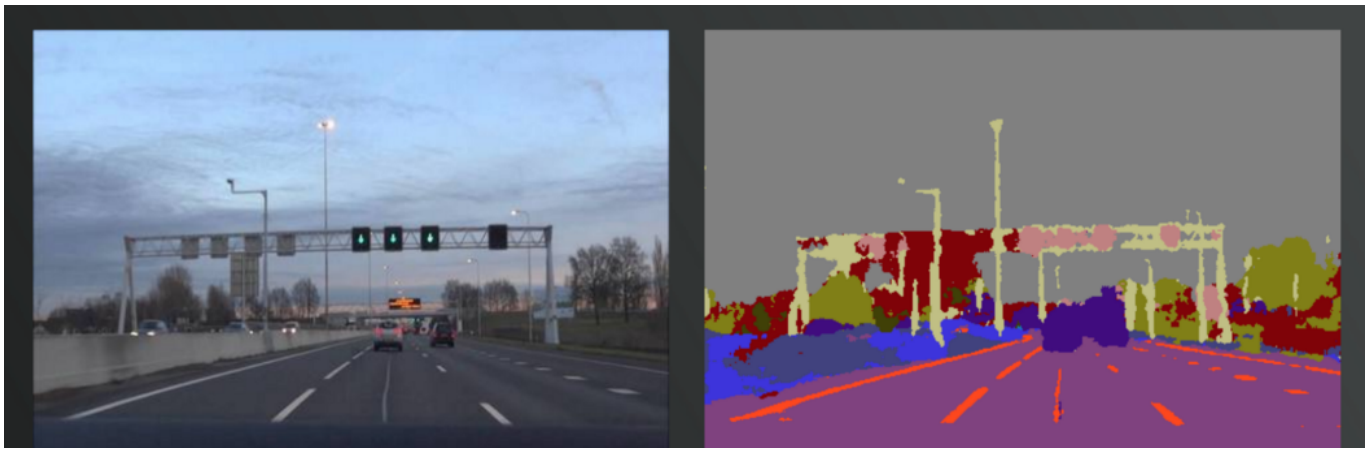
. . .

Computer Vision

Computer Vision is one of the most researched and most popular fields in machine learning. It is utilized to solve many everyday problems and consecutively involved in multiple applications, from which the most popular is currently the vision of self-driving cars. The tasks on which we'll take a look are *semantic segmentation, image classification and object detection.*

Semantic Segmentation

Semantic Segmentation can be seen as understanding the structures and components of an image on a pixel level. Methods for semantic segmentation try to make predictions about the structures and objects in an image. For a better understanding a semantic segmentation of a street scene can be seen below:



Semantic Segmentation with SegNet https://mi.eng.cam.ac.uk/projects/segnet/

The current leading algorithm **HRNet-OCR** was presented in 2020 by Tao et al. from Nvidia. It achieved a Mean Intersection Over Union (Mean IOU) of 85,1%. HRNet-OCR scales the image and uses a dense mask for each scale. The predictions of all scales are then "combined by performing pixel-wise multiplication between masks with the predictions followed by pixel-wise summation among the different scales to obtain the final results" [1].

Check out the Github to the technique:
https://github.com/HRNet/HRNet-Semantic-Segmentation

Other top-tier techniques (Method — Dataset):

- Efficient-Net-L2+NAS-FPN — PASCAL VOC
- ResNeSt-269 — PASCAL Context
- VMVF —ScanNet

. . .

. . .

Image Classification

Other than Semantic Segmentation, Image Classification, does not focus on the areas on the image, but on the image as a whole. This discipline tries to classify each image by

assigning a label.

source: image by author.

The **FixEfficientNet** has been presented first with the corresponding paper on the 20th April 2020 from the Facebook AI Research Team [2][3]. It is currently the state-of-the-art and has the best results on the ImageNet Dataset with 480M params, a top-1 accuracy of 88.5%, and top-5 accuracy of 98,7%. FixRes is the short form for Fix Resolution and tries to keep a fixed size for either the RoC (Region of Classification) used for train time or the crop used for test time. The EfficientNet is a compound scaling of the dimensions of a CNN which improves both accuracy and efficiency.

For more information to the FixEfficientNet, read this.

Other top-tier techniques (Method — Dataset):

- BiT-L — CIFAR-10
- Wide-ResNet-101 — STL-10
- Branching/Merging CNN + Homogeneous Filter Capsules — MNIST

Object Detection

Object detection is the task of recognizing instances of objects of a certain class within an image.

The current leading Object Detection technique is the **Efficient-Det D7x** first presented by the Google Brain Team (Tan et al.) in 2020 [4]. It achieved an AP50 (For more on AP50: Average Precision with a fixed IoU threshold at 50) of 74,3 and a box AP of 55,1.

The Efficient-Det is a combination of **EfficientNets** with Bidirectional Feature Pyramid Networks (**BiFPNs**).

As shortly explained above, the **EfficientNet** is a compound scaling of the dimensions of a CNN which improves both accuracy and efficiency. For more on EfficientNet, you can click here.

In Computer Vision a typical approach to increase the accuracy is the creation of multiple copies of the same image with different resolutions. This results in a so-called Pyramid due to the arrangement of the smallest image as the top layer and the biggest image as the bottom layer. The Feature Pyramid Network represents such a pyramid. Bidirectional means that there is not only a top-down approach but simultaneously a bottom-up approach. Every bidirectional path is used as a feature network layer and this leads to the BiFPNs. It helps with increasing accuracy and speed. For more information on BiFPNs, click here.

Other top-tier techniques (Method — Dataset):

- RODEO — PASCAL VOC
- Patch Refinement — KITTI Cars Easy
- IterDet — CrowdHuman

. . .

Natural Language Processing

A common definition for Natural Language Processing is the following:

> *NLP is a subfield of AI that gives the machines the ability to read, understand and derive meaning from human languages.*

NLP tasks vary in a broad range and as the definition reveals, all of them try to deduct some meaning from our language and perform calculations based on our language and its components. Algorithms based on NLP can be found in various applications and industries. Just to name a few applications which you might encounter every day such as

translators, social media monitoring, chatbots, spam filters, grammar check in Microsoft word or messengers and virtual assistants.

Sentiment Analysis

Sentiment Analysis is a field of Text Mining and is used to interpret and classify emotions in text data. One of the current leading algorithms is **BERT** which achieved an accuracy of 55.5 on the SST-5 Fine-grained classification dataset in 2019. The original paper was published by the Google AI Team [5].

BERT stands for **Bidirectional Encoder Representations from Transformers** and applies a bidirectional training of the Transformer technique. Transformer technique is an attention model used for language modeling which was previously only applied in one direction. Either to parse a text from left-to-right or from right-to-left. For more details, read this great article.

Other top-tier techniques (Method — Dataset):

- T5–3B — SST-2 Binary classification
- NB-weighted-BON + dv-cosine — IMDb

Language Modeling

Language Modeling is the task of predicting the next words or letters in a text based on the existing text/previous words.The GPT-2 model was given two sentences about a herd of unicorns living in the Andens and it created an astonishing story. You can read it here.

In Language Modeling one of the best performing algorithms can be found in **Megatron-LM**. This model and the paper were first presented in 2019 by the Nvidia team. A model similar to GPT-2 was trained on 8300 billion params. It was able to reduce the current state-of-the-art score of 15.8 to a test perplexity of only 10.8. The dataset used was the WikiText103 [6].

The model makes use of the Transformer Network. In their work, a transformer layer is made up of a self-attention block followed by a two-layer, multi-layer perceptron (MLP). In each of the blocks, model parallelism is used. This helps to reduce communication and keeps the GPUs compute-bound. The computation of the GPUs is duplicated to increase the speed of the model.

Other top-tier techniques (Method — Dataset):

- GPT-3 — Penn Treebank
- GPT-2 — WikiText2, Text8, enwik8

Machine Translation

Machine Translation is used in applications such as Google Translate or www.deepl.com. It is used to translate a text in another language using an algorithm.

One of the most promising algorithms in this field is the **Transformer Big +BT.** It was presented in this paper in 2018 by the Google Brain Team. In general, Transformers are state-of-the-art for dealing sequences and for machine translation. Transformers do not use recurrent connections but parse instead sequences simultaneously [7].

The input is represented in green is given to the model (blue) and transformed to the output (purple) . GIF source

As you can see in the gif above the input and output differ. This is due to the two different languages, the input is for example in English, while the output language is german. For the sake of increasing speed parallelization is a key aspect of the model. This problem is tackled by using CNN together with attention models. The self-attention helps to increase the speed and the focus on certain words, while CNN is used for parallelization [8]. For more on transformers read this great article. The authors applied **back-translation (BT)** for their training. In this method, the training dataset is translated into the target language and the algorithm translates it back to the original language. The performance can then be observed perfectly [7].

Other top-tier techniques (Method — Dataset):

- MAT+Knee — IWSLT2014 German-English
- MADL — WMT2016 English-German
- Attentional encoder-decoder +BPE — WMT2016 German-English

Text Classification

Text Classification is the task of assigning a certain category to a sentence, a text, or a word. The current leading algorithm on three different datasets (DBpedia, AG News and IMDb) is **XLNet.**

The paper and the technique **XLNet** was first presented in 2019 by the Google AI Team. It improved the leading algorithm BERT in 20 tasks. The method that **XLNet** pioneered in is called **Permutation Language Modeling**. It makes use of a permutation of the words. Imagine you got 3 words in the following order [w1,w2,w3]. All permutations are then retrieved, here 3*2*1 = 6 permutations. Obviously, with long sentences, this leads to numerous permutations. All words positioned before the prediction word (e.g. w2) are used for prediction [9]:

```
w3 w1 w2
w1 w2 w3
w1 w3 w2
 …
```

In row 1, w3 and w1 are used for the prediction of w2. In row 2 only w1 is used for prediction and so on. To get a better understanding of the technique you can read more about it here.

Other top-tier techniques (Method — Dataset):

- USE_T + CNN — TREC-6
- SGC — 20News

Question Answering

Question Answering is the task of training an algorithm to answer questions (often based on reading comprehension). This task is part of Transfer Learning due to the

learning on a given text database and the storing of knowledge to answer the questions to a later point in time.

With the **T5–11B** the Google AI Team achieved state-of-the-art benchmarks on four different datasets: GLUE, SuperGLUE, SQuAD and CNN/Daily Mail. T5 stands for the five T's in Text-to-Text Transfer Transformer, while 11B stands for the 11 Billion Dataset with which the algorithm was trained. In contrast to BERT and other great algorithms the T5–11B does not output a label to the input sentence. Instead, as the name already shows, the output is a text string as well [10].

source: https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

The authors of the paper have rigorously evaluated and refined dozens of existing NLP tasks to take the best ideas to their model. These included experiments on model architectures, pre-training objectives, unlabeled datasets, training strategies and scale as the authors describe [10]:

> **model architectures**, *where we found that encoder-decoder models generally outperformed "decoder-only" language models;*
>
> **pre-training objectives**, *where we confirmed that fill-in-the-blank-style denoising objectives (where the model is trained to recover missing words in the input) worked best and that the most important factor was the computational cost;*
>
> **unlabeled datasets**, *where we showed that training on in-domain data can be beneficial but that pre-training on smaller datasets can lead to detrimental overfitting;*

> **training strategies**, *where we found that multitask learning could be close to competitive with a pre-train-then-fine-tune approach but requires carefully choosing how often the model is trained on each task;*
>
> *and* **scale**, *where we compare scaling up the model size, the training time, and the number of ensembled models to determine how to make the best use of fixed compute power [11]*

The full T5–11B model is more than thirty times the size of existing NLP models such as BERT.

Other top-tier techniques (Method — Dataset):

- T5–11B — SQuAD1.1 dev
- SA-Net on Albert — SQuAD2.0
- TANDA-RoBERTa — WikiQA

. . .

Recommendation Systems

YYou have most probably already seen and used diverse kinds of recommendation systems. Your favorite online shop or platform uses it to suggest similar products in which you might be interested.

One of the current leading algorithms in this field is the **Bayesian time SVD++.** It was presented in 2019 by the Google Team and achieved SOTA benchmarks on the MovieLens100K Dataset. The Google Team tried multiple diverse methods and combinations of methods until they found the leading combination of a Bayesian Matrix Factorization and a timeSVD++. The Bayesian Matrix Factorization model was trained using a Gibbs sampling. More on the model and all methods tried out you can find here [12].

Other top-tier techniques (Method — Dataset):

- H+Vamp Gated — MovieLens 20M

- EASE — Million Song Dataset
- Bayesian timeSVD++ flipped w/ Ordered Probit Regression — MovieLens 1M

Speech Recognition

As well as Recommender Systems Speech Recognition takes part in our everyday life. There are more and more applications utilizing speech recognition in form of virtual assistants such as Siri, Cortana, Bixby, or Alexa.

One of the leading algorithms in this field is the **ContextNet + SpecAugment-based Noisy Student Training with Libri-Light** first introduced 2019 by the Google Team, the paper [13].

As the name reveals this method combines a ContextNet with Noisy Student Training. The ContextNet is a CNN-RNN-Transducer. The model consists of an audio encoder for the input audio, a label encoder for producing the input label, and a joint network of both to decode. For the label encoder, a LSTM is used and the audio encoder is based on a CNN. The Noisy Student Training is a sort of semi-supervised learning that uses unlabeled data to improve accuracy [13].

> *"In noisy student training, a series of models are trained in succession, such that for each model, the preceding model in the series serves as a teacher model on the unlabeled portion of the dataset. The distinguishing feature of noisy student training is the exploitation of augmentation, where the teacher produces quality labels by reading in clean input, while the student is forced to reproduce those labels with heavily augmented input features." [13]*

The Libri Light refers to the unlabeled audio dataset on which the model is trained and which is derived from audio books.

Other top-tier techniques (Method — Dataset):

- ResNet + BiLSTMs acoustic model — Switchboard + Hub500
- LiGRU + Dropout + BatchNorm + Monophone Reg — TIMIT
- Large-10h-LV-60k — Libri-Light test-clean

• • •

Conclusion

The last decade has brought a breakthrough in multiple disciplines and tasks. New technologies, algorithms, and applications have been discovered and developed and we are still at the beginning. This was made mainly possible through two developments: 1) growing databases that made it possible to feed the algorithm with enough data and 2) the technological development of processors, RAM, and graphic cards made it possible to train more complex algorithms that need more computing power. Furthermore, the half-life period of an algorithm being state-of-the-art shrinks also with increasing investments in Data Science and with more and more people being interested in the field of Data Science and Machine Learning. Consecutively, this article might be already out of date in a year. But for now, these are leading techniques which help in the progress of creating better and better algorithms.

For the case you know other methods or disciplines that should be added, you can comment or contact me. I appreciate your feedback and hope you enjoyed reading this article!

. . .

References:

[1] Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical Multi-Scale Attention for Semantic Segmentation. *ArXiv:2005.10821 [Cs]*. http://arxiv.org/abs/2005.10821

[2] Touvron, H., Vedaldi, A., Douze, M., & Jégou, H. (2020b). Fixing the train-test resolution discrepancy: FixEfficientNet. *ArXiv:2003.08237 [Cs]*. http://arxiv.org/abs/2003.08237

[3] Touvron, H., Vedaldi, A., Douze, M., & Jégou, H. (2020a). Fixing the train-test resolution discrepancy. *ArXiv:1906.06423 [Cs]*. http://arxiv.org/abs/1906.06423

[4] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and Efficient Object Detection. *ArXiv:1911.09070 [Cs, Eess]*. http://arxiv.org/abs/1911.09070

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

[6] Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2020). Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *ArXiv:1909.08053 [Cs]*. http://arxiv.org/abs/1909.08053

[7] Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. *ArXiv:1808.09381 [Cs]*. http://arxiv.org/abs/1808.09381

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

[9] Touvron, H., Vedaldi, A., Douze, M., & Jégou, H. (2020b). Fixing the train-test resolution discrepancy: FixEfficientNet. *ArXiv:2003.08237 [Cs]*. http://arxiv.org/abs/2003.08237

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

[11] https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

[12] Rendle, S., Zhang, L., & Koren, Y. (2019). On the Difficulty of Evaluating Baselines: A Study on Recommender Systems. *ArXiv:1905.01395 [Cs]*. http://arxiv.org/abs/1905.01395

[13] Park, D. S., Zhang, Y., Jia, Y., Han, W., Chiu, C.-C., Li, B., Wu, Y., & Le, Q. V. (2020). Improved Noisy Student Training for Automatic Speech Recognition. *ArXiv:2005.09629 [Cs, Eess]*. http://arxiv.org/abs/2005.09629

Machine Learning | Algorithms | Data Science | NLP | Cv