1. (强推)李宏毅2021春机器学习课程 <https://www.bilibili.com/video/BV1Wv411h7kN/?spm_id_from=333.788.recommend_more_video.1>
2. <https://speech.ee.ntu.edu.tw/~hylee/ml/2021-spring.php>
3. Chapter 1, introduction;
4. P1, 2021机器学习相关规定; deep learning; (2021-12-9)
5. P2, 机器学习基本概念简介上; machine learning == looking for function; regression(连续); classification(离散); (1) function with unknown parameters; (2) define loss from training data; (3) optimization; (2021-12-10)
6. P3, 机器学习基本概念简介下; linear model is too simple; sigmoid function; new model, more features; sigmoid->RELU; (2021-12-10)
7. P4, google colab教学; (2021-12-10)
8. P5, pytorch教学 part 1; tensor -- data type; constructor; operators; device; calculate gradient; dataset & DataLoader; torch.nn; MyModel, initialize your model & define layers; compute output of your NN; torch.optim; (2021-12-10)
9. P6, pytorch教学 part 2; cuda out of memory, batch size is too large; (2021-12-10)
10. P7, 作业说明HW1; objective; 疑似病例和确诊病例; task description; data, get from facebook; one-hot vector; training; evaluation metric; Kaggle; submission; grading; (2022-1-20)
11. P8, 深度学习简介; ups and downs of deep learning; (1) define a set of function; (2) goodness of function; (3) pick the best function; fully connected network; trail & error + intuition; backprop; (2021-12-10)
12. P9, 反向传播; gradient descent; chain rule; (2021-12-10)
13. Chapter 2, deep learning;
14. P10, 机器学习任务攻略; framework of ML; model bias; optimization issue; 先跑一些比较小的network或者非深度学习方法; overfitting; CNN是FC针对image任务的特化; public testing set; private testing set; N-fold cross validation; mismatch; (2021-12-13)
15. P11, 类神经网络训练不起来怎么办(1); when gradient is small; optimization fails because... ; local minima; saddle point; Tayler series approximation; saddle point(高维空间) vs local minima(低维空间); (2021-12-13)
16. P12, 类神经网络训练不起来怎么办(2); batch; 大batch不一定时间长; (1) small batch is better; (2) noisy update is better; small batch vs large batch; momentum; gradient descent+momentum; (2021-12-13)
17. P13, 类神经网络训练不起来怎么办(3); customized learning rate; root mean square; learning rate adapts dynamically; RMSProp; learning rate decay; (2021-12-13)
18. P14, 类神经网络训练不起来怎么办(4); classification; classification as regression? Class as one-hot vector; softmax; loss of classification; MSE; CE; (2021-12-13)
19. P15, 类神经网络训练不起来怎么办(5); change landscape; feature normalization; consider deep learning; batch normalization, 适用于big batch; internal covariate shift; (2021-12-14)
20. P16, optimization for deep learning 1; background knowledge; SGD with momentum; why momentum? Adagrad; optimizer, real application; adam vs SGDM; (2021-12-14)
21. P17, optimization for deep learning 2; does Adam need warm-up? Of course! K step forward, one step back; momentum recap; NAG; L2 regularization; something helps optimization--shuffling, dropout, gradient noise, warm-up, curriculum learning, fine-tuning; SGDM和ADAM各自适用情况; no universal optimizer; (2021-12-14)
22. P18, classification; Gaussian distribution; maximum likelihood; probability distribution; 后验概率; (2021-12-14)
23. P19, logistic regression; Step 1, function set; Step 2, goodness of a function; 逻辑回归vs线性回归; Step 3, find the best function; logistic regression + square error; cross entropy better than square error; discriminative vs generative; multi-class classification; limitation of logistic regression, boundary是一条直线; (2021-12-14)
24. P20, 作业说明HW2; classification; (2021-12-30)
25. P21, 作业说明HW2(English version); (2021-12-14)
26. Chapter 3, CNN & self-attention;
27. P22, 卷积神经网络; image classification; 识别关键pattern即可; simplification 1, typical setting; 每个neuron输入一个receptive field; padding是为了补全; simplification 2, 共享参数; benefit of CNN; CNN类似filter; multiple convolutional layer; observation 3; neuron vs filter; pooling -- max pooling, 把图片变小(减小运算量, 可有可无); convolution和pooling交替使用; alphaGo; 围棋和image有共同的特性; alpha Go没用pooling; CNN不可以放缩和旋转图片, 所以需要data augmentation; (2021-12-15)
28. P23, self-attention上; input is a set of vectors, 而且长度会改变; word embedding; what is output? Label; seq2seq; sequence labeling; consider the context? FC can consider the neighbor; how to consider the whole sequence? A window consider the whole sequence? Self-attention; 考虑了整个句子; can be either input or hidden layer; dot-product or addictive; (2021-12-15)
29. P24, self-attention下; 只有WQ, WK, WV三个矩阵是需要学习的; multi-head self-attention; 不同的q负责不同种类的相关性, 关注不同的方面; position encoding; self-attention for speech; self-attention for image; CNN is simplified self-attention; CNN is better for less data, self-attention is better for more data; self-attention vs RNN; RNN遗忘久远的内容, self-attention不会; RNN无法并行化处理, self-attention可以, 所以更有效率; RNN可以被self-attention取代了; self-attention for graph; 减小self-attention的计算量是未来的重点; (2021-12-15)
30. P25, unsupervised learning -- word embedding; 1-of-N encoding; word class; generating word vector is unsupervised; a word can be understood by its context; exploit the context? (1) count based; (2) prediction based - sharing parameters; CBOW; skip-gram; 跨语言embedding; document embedding; beyond bag of word; (2021-12-15)
31. P26, spatial transformer; CNN无法放缩和旋转; 平移图片; image transformation(平移旋转缩放); (2021-12-15)
32. P27, RNN; slot filling; hidden layer is memory; Elman network & Jordan network; Bidirectional RNN; LSTM; RNN就是LSTM, GRU是简化版的LSTM; (2021-12-15)
33. P28, GNN1; classification; generation; spatial-based GNN; NN4G; DGCN; DGC; MoNET; graphSAGE; GAT; GIN; (2021-12-15)
34. P29, GNN2; spectral-based convolution; Fourier Series representation; Fourier transform; spectral graph theory; filtering; ChebNet; GCN; (2021-12-15)
35. P30, HW3; CNN; (2021-12-30)
36. P31, HW3(English version); (2021-12-15)
37. P32, HW4; self-attention; task introduction; Phoneme classification; Speaker classification; dataset; Data Preprocessing; Data formats; Data segmentation during training; sample code; requirements; grading; submission format; deadlines; grading—bonus; hints; regulation; (2021-12-30)
38. P33, HW4(English version); (2021-12-16)
39. Chapter 4, Theory of ML (Prof. Pei-Yuan Wu)
40. Chapter 5, transformer;
41. P34, transformer上; seq2seq; output length is determined by model; speech recognition; machine translation; 硬train一发; text-to-speech; seq2seq for chatbot; NLP == QA, QA == seq2seq; seq2seq for multi-label classification; seq2seq for object detection; input -> encoder -> decoder -> output; encoder is vec2vec; block = self-attention + FC; 残差网络, 输入+输出; layer norm; Add&Norm == residual + layer normalization; 原有block可以换一下顺序; (2021-12-16)
42. P35, transformer下; decoder; auto-regressive(由自己的过去推出自己的未来); begin token; 除了中间部分, encoder和decoder是一样的; masked self-attention, 只能考虑已有的(过去的/左边的)资讯; decoder怎样知道输出的长度? 推文接龙; 需要加上end token; non-autoregressive decoder(NAT); (1) another predictor for output length; (2) ignore after END token; Advantage: parallel; usually worse than AT; encoder-decoder; cross\_attention, q from decoder, k and v from encoder; training; teacher forcing: using ground truth as input; copy mechanism; summarization; 可能会犯低级错误; guided attention; beam search; decoder needs randomness when generating sequence; 不知道怎样优化, 就用强化学习; scheduled sampling, training set不仅有正确的, 还有错误的; (2021-12-16)
43. P36, non-autoregressive sequence generation; RNN is auto-regressive; Vanilla NAT; fertility; sequence-level knowledge distillation; noisy parallel decoding; evolution of NAT; NAT with iterative refinement; mask-predict; insertion transformer; multiple target words to predict?
44. P37, HW5; transformer; (2021-12-16)
45. P38, HW5; (2021-12-16)
46. P39, HW5; (2021-12-16)
47. Chapter 6, generative model;
48. P40, GAN基本概念介绍; Network as generator; y = Net(x+random); video distribution; 让输出不再是单一的输出, 而是一个概率分布; 当任务需要一点创造力的时候, 就要random distribution; 对话中, 每个人可能有不同的答案; GAN; anime face generation; unconditional generation; y = Net(random); discriminator, 越像的输入, 则输出数值越大; generator和discriminator相互进化; (1) fix generator, update discriminator; (2) fix discriminator, update generator; 生成动画人物; 可以生成从未看过的人脸; 两张脸图片的内插; (2021-12-16)
49. P41, 理论介绍与WGAN; 使PG和Pdata距离越近越好; sampling is good enough; GAN is difficult to train; tips for GAN; JS divergence is not suitable; PG and Pdata are not overlapped; Wasserstein distance; 取最小的推土值; WGAN; 怎样求解Wasserstein distance; D has to be smooth enough; (2021-12-17)
50. P42, 生成效能评估与条件式生成; GAN is still challenging; GAN for sequence generation; 不能gradient descent的, 就用reinforcement learning做; scratchGAN; variational autoencoder; flow-based model; 用监督学习做生成模型; evaluation of generation; 靠image classification产生的分布来辨识; diversity -- mode collapse; mode dropping; diversity越多越好; inception score = good quality + large diversity; FID; we don’t want memory GAN; GAN的评估是非常困难的(SSFC的评估也是非常困难的, 谁知道生成的程序有没有用?); conditional generation; text-to-image; 需要文字和图片成对的资料; image-to-image; sound-to-image; GAN生成会动的图片; (2021-12-17)
51. P43, cycle GAN; learning from unpaired data; image style transfer; 真人头像转成二次元人物头像; cycle GAN, 经过两次转换, 输入输出越接近越好; Network很懒, 基本就生成最像的; StarGAN; text style transfer; (2021-12-17)
52. P44, deep generative model上; PixelRNN; 创造很难evaluate; 难以确定最佳的生成顺序, 生成速度很慢; drawing from scratch needs some randomness; variational autoencoder(VAE); 可以固定向量中的某几维; VAE写诗; (2021-12-17)
53. P45, deep generative model下; why VAE? VAE加上了noise, 使一个点变成了一个范围区间; loss加上限制不能太小; estimate the probability distribution; P(x) = sum(P(m)\*P(x|m)); EM algorithm; maximizing likelihood; connection with network; conditional VAE; 根据一个数字画出其他风格相近的数字; problems of VAE; VAE does not try to simulate real image; GAN; 拟态的演化; evolution of generation; generator可以产生database从未有过的image; GAN -- toy example; GAN is difficult to optimize; (2021-12-18)
54. P46, Flow-based generative model; 最大似然估计==最小KL散度; flow-based model可以直接maximize likelihood; Jacobian matrix; determinant; 行列式是高维空间中的体积; change of variable theorem; p(x’)\*|det(Jf)| = pai(z’); dim(z) = dim(x); 多个G叠加; 训练的实际是G-1; coupling layer; 交替使用; 1\*1 convolution; W can shuffle channels; (2021-12-18)
55. P47, HW6; GAN; (2021-12-18)
56. P48, HW6, English version; (2021-12-19)
57. Chapter 7, self-supervised learning;
58. P49, 芝麻街与进击的巨人; 自监督模型都是以芝麻街命名; BERT是非常巨大的模型, 340 million参数; (2021-12-19)
59. P50, BERT简介; self-supervised learning; 一部分作为模型的输入, 另一部分作为模型的标注; 没有label就是unsupervised learning; masking input; transformer encoder就是Bert架构, Bert训练的是encoder; randomly mask/replace some tokens; 输出接近被遮挡的word; next sentence prediction; 判断两个句子是否相接; 这一招不太有用, 可能因为此任务太简单了; sentence order prediction; downstream task可能和天空没有关系; pre-train->fine-tune; GLUE scores; how to use Bert? Case 1, 判断句子正面还是负面; pre-train better than random; Case 2, 词性标注; Case 3, 自然语言推断; case 4, 问答(答案一定在文章里), 找出正确答案在文章中的起始位置和结束位置; training Bert is expensive; BERT胚胎学; pre-training a seq2seq model; MASS/BART; T5 - comparison; (2021-12-19)
60. P51, Bert的奇闻异事; why does Bert work? Embedding vector代表了word的意思; similar embedding has similar meaning, context is being considered; You shall know a word by the company it keeps; Bert是复杂版本的CBOW, 而且Bert还可以考虑上下文, 所以Bert抽取的word是contextualized word embedding; apply Bert to protein&DNA, 效果也很好; cross-lingual alignment; 只有资料量足够大才有效; 语言差异可以用向量表示; (2021-12-19)
61. P52, GPT的野望; predict next token; they can do generation; how to use GPT? Few-shot learning, no gradient descent; one-shot learning; zero-shot learning; beyond text; Image-simCLR; BYOL; speech; speech GLUE; (2021-12-19)
62. P53, auto-encoder上, 基本概念; auto-encoder类似self-supervised, 不需要人为标注书签; 使encoder输入和decoder输出越接近越好, 类似于cycle GAN; embedding == representation == code (low dim, bottleneck); dimension reduction; why auto-encoder? 化繁为简, 抽象出最本质的东西; de-noising auto-encoder; 还原加入noise之前的结果; bert的mask可以视为noise; Bert == de-noising auto-encoder; (2021-12-19)
63. P54, auto-encoder下, 领结变声器与更多应用; feature disentangle; 对向量维数代表的内容进行区分; application, voice conversion; 把我的声音转成新垣结衣的声音; discrete latent representation; 强迫embedding vector为one-hot; VQVAE; text as representation; embedding的不是向量, 而是文字; 或许文字就是文章的摘要; 其实不是, 再加上GAN使之合乎文法; 也是cycle GAN的另一种形式; tree as embedding; decoder可以视为generator; 用来compress; anomaly detection; input is similar or not; 用来判别欺诈; (2021-12-19)
64. P55, Bert and its family; pre-train & fine-tune; 通过大量语料训练语感, 再做少量题就可以了; pre-train model; represent each token by an embedding vector; 这种token不考虑上下文; English word as token; contextualized word embedding; LSTM, self-attention layers, tree-based model(擅长处理文法结构严谨的, 比如数学表达式); bigger model; smaller model; Albert; network compression; how to fine-tune; w1w2[sep]w3w4w5; 看到CLS要产生和整个句子有关的token; output: (1) class for each token; (2) copy from input(extraction-based QA); (3) general sequence; only fine-tune task specific or fine-tune all (better); only fine-tune adaptor part; weighted features; (2021-12-19)
65. P56, other models; how to pre-train; context vector; treat model as translation encoder; self-supervised learning; predict next token; 使用LSTM的ELMo; 使用self-attention的GPT; they can do generation; bidirectional LSTM (disadvantage, 各自只考虑了一半的资讯); BERT masking/replacing input; transformer, no limitation on self-attention; BERT类似CBOW, 只不过不受window size限制; masking input; whole word masking; phrase level & entity level (ERNIE); SpanBert, mask掉一个很长的范围; span boundary objective; useful in coreference; XLNet; predict next token只能看到之前的token; XLNet打乱sentence token顺序; BERT cannot talk(lack generation ability, 只能当作encoder)? Given partial sentence, predict next token; LM model (auto-regressive model) born for; BERT never seen partial sentence; MASS/BART; how to pre-train seq2seq? 需要对输入进行一定程度的破坏, 在输出还原; input corruption; deletion/permutation(bad)/rotation(bad)/text infilling(good); UniLM; Bidirectional LM(BERT) + left-to-right LM(GPT) + seq2seq LM(MASS/BART); replace or not? 把token替换成其他词汇, 输出判定置换的token; yes/no is easier, every position is used; small BERT; similar to GAN; 由于任务简单, 同等运算量下效果更好; representation for whole sentence OR sentence embedding, by sentence company; skip thought, predict next sentence; quick thought, 让相邻的句子embedding尽可能接近, 避开生成的任务; next sentence prediction (NSP)效果不太好; sentence order prediction(SOP), 使相接的句子转向, used in ALBERT; structBERT; pre-train运算量很大, 一般人做不了; T5 - comparison; ERNIE, add external knowledge; audio BERT; (2021-12-20)
66. P57, multi-lingual BERT; 把所有语言统统倒进BERT; zero-shot reading comprehension; train on English QA, test on Chinese QA, 结果也挺好, 比翻译后训练的结果好; XTREME; MRR; word2vec and Glove cannot do well even with more data; how alignment happens? Different languages share some common tokens; code switching; intermediate language; 不同语言的资讯藏在哪里? English average - Chinese average; attribute representation; (2021-12-20)
67. P58, GPT3; very big; want zero-shot learning; few-shot learning; one-shot learning; zero-shot learning; closed book QA; LM can do generation; 写新闻; 造句; 算术; Turing advice challenge; apply GPT3 on graph; (2021-12-20)
68. P59, unsupervised learning -- linear method; (1) clustering & dimension reduction; (2) generation; clustering; K-means; HAC; distributed representation; dimension reduction; 流形降维; 很多维数是不需要的; feature selection; PCA; maximize variance; PCA-decorrelation; 将特征值由大到小提取; another point of view; linear combination of basic component; weakness of PCA; unsupervised & linear; Eigen digit; face; non-negative matrix factorization (NMF), 强迫相加不能相减; NMF on MNIST; NMF on face; matrix factorization; 填补缺失的value, 用于推荐系统; latent semantic analysis; (2021-12-21)
69. P60, unsupervised learning -- neighbor embedding; t-SNE; 非线性降维; manifold learning; locally linear embedding; keep relation unchanged; Laplacian Eigenmaps; high density connection; t-SNE; 把数据聚成一堆一堆; (2021-12-21)
70. P61, HW7; BERT; (2021-12-21)
71. P62, HW8; anomaly detection; (2021-12-21)
72. Chapter 8, explainable AI / adversarial attack;
73. P63, adversarial attack上; robust; example of attack; 加入小noise, 把cat识别成star fish; noise越小越好, 识别差越大越好; L2-norm; L-infinity; attack approach; update input, not parameters; FGSM; (2021-12-21)
74. P64, adversarial attack下; white box v.s. black box; 知道参数的攻击是white box; black box attack; train a proxy network; 挺容易成功的; the attack is so easy! Adversarial examples are not bugs, they are features; 原因不在Model, 而在data; one pixel attack; universal adversarial attack; beyond image; attack in the physical world; backdoor in model; Defense; passive defense; smoothing; image compression; generator; 随机防御方法; adversarial training; 给攻击成功的图片标上正确的label; data augmentation; (2021-12-21)
75. P65, explainable ML上; correct is not clever; interpretable v.s. powerful; decision tree; make people comfortable; which component is crucial? Saliency map; smooth grad, 给图片加上不同的noise; gradient saturation; how NN process input data? Visualization; probing; (2021-12-22)
76. P66, explainable ML下; global explanation; what does a filter detect? What does a digit look like for CNN? Constraint from generator; 让人开心而不是机器开心; 用linear model模仿未知model (哪怕模拟一小部分); LIME; (2021-12-22)
77. P67, more about adversarial attack上; attack on image; one pixel attack; how do we find the exact pixel and value? Differential evolution; one pixel attack recap; DE used in one pixel attack; (2021-12-22)
78. P68, more about adversarial attack下; attacks on ASR; attacks on ASV; wake up words; hidden voice attack; 听声辨位; signal preprocessing; perturbation; time domain inversion; random phase generation; high frequency addition; time scaling; (2021-12-22)
79. P69, HW9; Explainable AI; (2021-12-30)
80. P70, HW10; Adversarial Attack; (2021-12-30)
81. Chapter 9, domain adaption;
82. P71, domain adaption; 当测试资料和训练资料有差异时, 怎样保证accuracy; transfer learning; domain shift; knowledge of target domain; (1) little but labeled, train by source data, fine-tune by target data; (2) large amount of unlabeled data; feature extractor, learn to ignore colors; domain adversarial training; feature extractor (generator) + label predictor; 让feature extractor的结果分不出差异, 骗过domain classifier; limitation; outlook; universal domain adaption; (3) little & unlabeled; domain generalization; (2021-12-22)
83. P72, HW11; Adaptation; (2021-12-30)
84. Chapter 10, Privacy v.s. ML (Prof. Pei-Yuan Wu);
85. Chapter 11, Reinforcement learning;
86. P73, Reinforcement learning 1; challenging to label data in some tasks; what is RL? Machine learning == looking for function; playing video game; (1) function with unknown; input, observation; output, actions; (2) define loss from training data; -reward; (3) optimization; reward之和越大越好; actor具有随机性, environment也有随机性; how to control your actor; optimize cross entropy; (2021-12-22)
87. P74, Reinforcement learning 2; reward delay; Version 1, cumulative reward; Version 2, discount factor; Version 3, reward有正有负; policy gradient; data collection is in training process; theta can only update once; time-consuming; on-policy; off-policy; difference between actor to train and actor to interact; need randomness; (2021-12-22)
88. P75, Reinforcement learning 3; actor critic; evaluating how good it is when observing s; value function, 需要在游戏结果之前预测; Monte-Carlo approach, 不适合永不结束的游戏; temporal-difference approach, 利用递推式训练; compared with average; Version 4, advantage actor-critic; parameters of actors and critics can be shared; deep Q network; (2021-12-23)
89. P76 Reinforcement learning 4; reward shaping; sparse reward, reward=0 in most cases; define extra reward; VizDoom; 扣血就惩罚; 加上了人类对知识的理解; curiosity-based, but meaningful; (2021-12-23)
90. P77, Reinforcement learning 5; sometimes reward is hard to define; imitation learning; 言传身教; behavior cloning; inverse RL; learn reward function from expert demonstration; the teacher is always the best; framework of IRL; actor == generator, reward == discriminator; (2021-12-23)
91. P78, HW12; RL; (2021-12-30)
92. P79, deep RL; agent observe environment, and act; supervised learning learn from teacher, RL learn from experience; learning a chat-bot; playing video game; difficulty of RL; reward delay; policy-based approach; learning an actor; RL is also looking for a function; neural network as actor; probability of taking an action; goodness of actor; gradient ascent; add a bias; critics; A3C; (2021-12-23)
93. P80, proximal policy optimization; on policy; off policy; issue of importance sampling; PPO is easier than TRPO; PPO2; experimental result; (2021-12-23)
94. P81, Q learning 1; critic related with actor; value, estimate critic; Monte-Carlo approach; Temporal-difference approach ; 不需要把游戏玩到底, 用递推式; TD比MC更常见; TD和MC估测结果不同; state-action value function, cumulated reward; loop: find a better actor; target network; exploration; epsilon (decay during training) greedy; Boltzmann Exploration; replay buffer; 利用过去的经验; (2021-12-28)
95. P82, Q learning 2; tips of Q-learning; double DQN; Q经常被高估; dueling DQN; 改network架构; prioritized replay; multi-step, balance between TD and MC; noisy net; noise on action OR parameters; distributional Q-function; rainbow; (2021-12-28)
96. P83, Q learning 3; Q learning for continuous actions; gradient ascent (因为是找最大值); design a network to make optimization easy; (2021-12-28)
97. Chapter 12, life long learning;
98. P84, life long learning 1; 其实新旧任务差别不大; forget; 同时学能记住, 依次学就忘了; multi-learning是life long learning的upper bound; transfer learning只在意new model; evaluation; (2021-12-23)
99. P85, life long learning 2; selective synaptic plasticity; why forgetting? Error surface不同; only change unimportant parameters; each parameter has a guard; bi需要人为制定; 看看改变theta对结果的影响; not forget, but intransigence; bi可以算, 方法各不同; 任务学习顺序不同, 结果相差很大; gradient episodic memory; 改变gradient descent下降方向; GEM偷存了先前的资料; additional neural resource allocation; progressive NN; 不断新增额外的参数; PackNet; 每次只用一部分网络; CPG; memory reply; generating Task 1 and send to Task 2; curriculum learning; (2021-12-24)
100. P86, 神经网络压缩1; network pruning; 剪掉network的一些参数; 树大必有枯枝; importance of a weight; importance of a neuron; 一次剪掉一点参数; weight pruning; network becomes irregular, hard to implement and speed up; 补0并没有实际变小; 95%的参数可以归零, 但运算并没有加快; neuron pruning; large network is easier to optimize; 把大网络变小有可能, 直接训练小网络不可能; 大乐透解说; 正负号是初始化参数的关键; (2021-12-24)
101. P87, 神经网络压缩2; knowledge distillation; teacher net, student net; 让学生的输出逼近老师的输出, 哪怕老师是错的; 使用多个模型求平均; temperature for softmax; parameter quantization; using less bits to represent a value; weight clustering; binary weights; architecture design; depthwise convolution; filter number == input channel number; pointwise convolution, 1\*1, 考虑channel之间的关系; 减少了参数; dynamic computation; NN自由调整运算量; dynamic depth; dynamic width; 可以同时用多个策略; (2021-12-24)
102. P88, HW13; network compression; (2021-12-30)
103. P89, HW14; lifelong learning; (2021-12-24)
104. P90, conjecture of deep learning; local minimum == global minimum; analyzing Hessian; 当参数多, local minimum概率极小; saddle在loss大的地方, local minimum在loss小的地方; (2021-12-24)
105. Chapter 13, meta learning;
106. P91, meta learning 1; learn to learn; learn hyperparameters; machine learning == looking for a function; (1) function with unknown parameters; (2) define loss from training data; training tasks containing training data & testing data; compute loss based on testing data of training task; (3) optimization; what is meta learning? Phi, learnable parameters; 用测试资料来计算Loss; 无法微分就用RL硬做; 测试集中只需要少量的标记数据; testing data中的资料是不能碰的; few-shot是目标, meta是手段; find a function F that finds a function f; across task (outer loop) training & testing; (2021-12-25)
107. P92, meta learning 2; review gradient descent; Case 1, learning to initialize; MAML, 训练任务和测试任务差距不大, 类似domain adaption & transfer learning; multi-task learning is baseline of MAML; rapid learning OR feature reuse? Almost No Inner Loop; Case 2, learn to optimize; Case 3, network architecture search (NAS); RL硬做; DARTS, differential architecture search; Case 4, data augmentation; sampling reweighting; beyond gradient descent; learning to compare; application; few-shot image classification; N-way K-shot, N classes, K examples; Omniglot; (2021-12-25)
108. P93, MAML 1; meta learning == learn to learn; life-long, one model for all tasks; meta, how to learn a new model; (2021-12-25)
109. P94, MAML 2; define a set of learning algorithm; network structure & init & update让机器自己来学, 之前是人为制定的超参数; meta learning就是学习超参数; 需要很多任务来评估; (2021-12-25)
110. P95, MAML 3; training tasks (training examples & testing examples); testing tasks (training examples & testing examples); few-shot是目的, meta是手段; L(F) = sum(l); (2021-12-25)
111. P96, MAML 4; Omniglot; N-way K-shot, N classes, K examples; (2021-12-25)
112. P97, MAML 5; MAML; 学习初始化参数init, 限制network structure必须一样; MAML(loss计算的是训练以后的表现) difference between pre-train (loss计算的是现在的表现); train update once, test update multiple times; 为了节省计算资源, 为了防止over-fitting; (2021-12-25)
113. P98, MAML 6; toy example; (2021-12-25)
114. P99, MAML 7; warning of math; 假装没有看到二次微分; (2021-12-26)
115. P100, MAML 8; implementation of MAML; translation; (2021-12-27)
116. P101, MAML 9; Reptile; 朝四步以后的方向走一步; 介于pre-traini和MAML之间; learn network structure OR learning rate; output parameter directly? (2021-12-27)
117. P102, metric-based 1; face verification; meta learning; siamese network; (2021-12-27)
118. P103, metric-based 2; binary classification; 不同角度的同一张人脸投影到相似位置; learn to ignore the background; (2021-12-27)
119. P104, metric-based 3; 5-way 1-shot; prototypical network; learn by cross-entropy as typical classification; matching network; relation network; few-shot learning for imaginary data; add generator; (2021-12-27)
120. P105, train + test as RNN; MANN; SNAIL; (2021-12-27)
121. P106, END; looking for function; input, matrix & sequence; METHOD: GAN, BERT, domain adaption, RL, attack & defense, explainable, network compression, life-long learning, meta learning; APPLICATION: CV, NLP(translation, QA), speech recognition, RL; find a problem you care about and try to solve it; NIPS, ICLR, AAAI, ICML; 范例程序稍微改改就能用在其他地方; (2021-12-27)
122. HW15: Meta Learning
123. 李宏毅2020机器学习深度学习(完整版)国语 <https://www.bilibili.com/video/BV1JE411g7XF?from=search&seid=4463720420343295823&spm_id_from=333.337.0.0>
124. P1, 机器学习; python is good; ML == looking for a function minimizing loss; what kind of function do you want? (1) Regression wants scalar; (2) binary classification; multi-class classification; (3) generation; supervised learning; supervised v.s. reinforcement; unsupervised learning; 寻找方法: gradient descent; (2022-1-8)
125. P2, Rule; (2022-1-8)
126. P3, regression; step 1, model; step 2, goodness of function; step 3, best function; gradient descent; unilinear function; model selection; overfit; what are the hidden factors? Back to step 1, redesign the model; back to step 2, regularization; (2022-1-8)
127. P4, basic concept; estimator; bias and variance of estimator; bias v.s. variance; deal with large bias; more feature, more complex; deal with large variance; more data; regularization; model selection; public testing set, private testing set; cross validation; N-fold validation; (2022-1-8)
128. P5, gradient descent 1; learning rate; adaptive learning rate; adagrad; stochastic gradient descent; feature scaling; gradient descent theory; (2022-1-9)
129. P6, gradient descent 2; (2022-1-9)
130. P7, gradient descent 3; (2022-1-9)
131. P8, optimization for DL 1; background; SGD; SGDM; adagrad; RMSProp; Adam; (2022-1-9)
132. P9, optimization for DL 2; RAdam v.s. SWATS; normalization; CV用SGDM, NLP用Adam; (2022-1-9)
133. P10, classification; how to do classification; two boxes; prior; probability from class - feature; max likelihood; modifying model; 用mean和covariance衡量模型的好坏; probability distribution; 假设所有feature相互独立, 就是naïve Bayes classifier; posterior probability; (2022-1-20)
134. P11, logistic regression; step 1, function set; step 2, goodness of a function; cross entropy; step 3, find the best function; cross entropy v.s. square error (梯度下降更快); discriminative v.s. generative; benefit of generative model; 善于脑补, 适用于数据较少的场合; multi-class classification; limitation of logistic regression; feature transformation; cascading logistic regression models; (2022-1-31)
135. P12, deep learning; ups and downs of deep learning; multi-layer perceptron; step 1, function set; step 2, goodness of a function; step 3, find the best function; fully connected feedforward network; deep = many hidden layers; 需要使用ResNet才能搞定100多层的network, Fully connected根本train不起来; matrix operation; hidden layer可以看作feature extractor; 以图像识别举例, 浅层提取边缘信息, 深层提取高维抽象信息; trial & error and intuition; loss for an example; total loss; gradient descent; backprop; deeper is better? (2022-2-3)
136. P13, backprop; gradient descent; chain rule; forward pass; backward pass; (2022-2-3)
137. P14, tips for deep learning; do not always blame overfitting; vanishing gradient; ReLU; maxout; hard to find optimal parameters; momentum; adam; early stopping; regularization; L1; weight decay; dropout; dropout is a kind of ensemble; testing of dropout; (2022-2-4)
138. P15, why deep; 高瘦强于矮胖; 因为分层有利于模组化, 不同层提取不同层次的信息; modularization - speech; phoneme; Gaussian Mixture Model; 所有函数都可以用一层神经网络表示; deeper means less parameters; analogy; hidden layers相当于剪窗花的对折; end-to-end learning; less engineering labor, machine learns more; (2022-2-1)
139. P16, pytorch tutorial; tensors and numpy; tensor([1, 2]); tensor.view == numpy.reshape; broadcasting; computing graph; cuda; 自动微分; linear regression; sequential; convolution; (2022-2-6)
140. P17, CNN; why CNN for image? A neuron connects to a small region; subsampling will not change object; the whole CNN; filter某个框中的数字 == FC中的一个weight; two neurons share weight; max pooling; what does CNN learn? Deep dream; deep style; playing go; speech; (2022-2-8)
141. P18, GNN1; NN; CNN; RNN; transformer; graph; classification; generation(开发新药); 从人物关系推断凶手; a node can learn from its neighbors; spatial-based; spectral-based; NN4G(neural network for graph); DCNN(diffusion-convolution neural network); MoNET(mixture model network); GraphSAGE; GAT(graph attention network); GIN(graph isomorphism network); (2022-3-3)
142. P19, GNN2; spectral-based convolution; N-dim vector space; Fourier Series representation; time domain basis; frequency domain basis; Fourier transform; spectral graph theory; eigenvectors; interpreting vertex frequency; filtering; ChebNet; GCN; benchmark tasks; results; application to GCN; drop edge; graph generation; VAE-based model; GAN-based model; AR-based model; GNN roadmap; (2022-4-29)
143. P20, RNN (Part I); slot filling; 1-of-N encoding; beyond 1-of-N encoding, 归于other; word hashing; example application; neural network needs memory; the output of hidden layer is stored in the memory; example; multiple hidden layer; Elman Network & Jordan Network; bidirectioinal RNN; LSTM; input gate, output gate, forget gate; (2022-4-30)
144. P21, RNN (Part II); learning target; RNN is not easy to train; the error surface is either very deep or very deep, 原因在于RNN把同一套参数在不同的位置反复使用; clipping; LSTM can handle gradient vanishing (not gradient explode); the influence never disappeared unless forget gate is closed; gated recurrent unit (GRU), simpler than LSTM; clockwise RNN; SCRN; more application; many-to-one; sentiment analysis; key term extraction; many-to-many; speech recognition; add an extra symbol to represent “null”; CTC; machine translation; don’t know when to stop; syntactic parsing; sequence to sequence auto-encoder; chat-bot; attention-based model; machine’s memory; neural Turing machine; reading comprehension; ask machine a question; visual machine answering; speech question answering; attention在文中划重点; simple baselines; RNN v.s. structured learning; 单向RNN没有考虑整个句子; cost and error are not related; DNN+HMM; is structured learning practical? GAN is a kind of structured learning; deep and structured will be the future; (2022-5-1)
145. P22, semi-supervised; introduction; 用data的feature不算偷, 用label才算偷; transductive learning, unlabelled data is the testing data; inductive learning, unlabelled data is not the testing data; collecting data is easy, but collecting labelled data is expensive; the distribution of the unlabelled data tell us something; 聚类假设: 处于相同聚类的样本更可能具有相同标记; 流型假设: 处于很小的局部区域的样本更相似，更可能具有相同标记；; supervised generative model; semi-supervised generative model; low-density separation; self-training; slides中应该是加到训练好的data; self-training用硬标签, 生成模型用软标签; it looks like class 1, then it is class 1; entropy-based regularization; 看分布集中不集中; semi-supervised SVM; smoothness assumption; 近朱者赤近墨者黑; 文章分类; cluster and then label; graph-based approach; 启发式, 根据经验根据直觉怎么爽怎么来; graph construction: define similarity, add edge, edge weight equals to similarity; define the smoothness of label; graph Laplacian; L = D – W; looking for better representation;
146. -