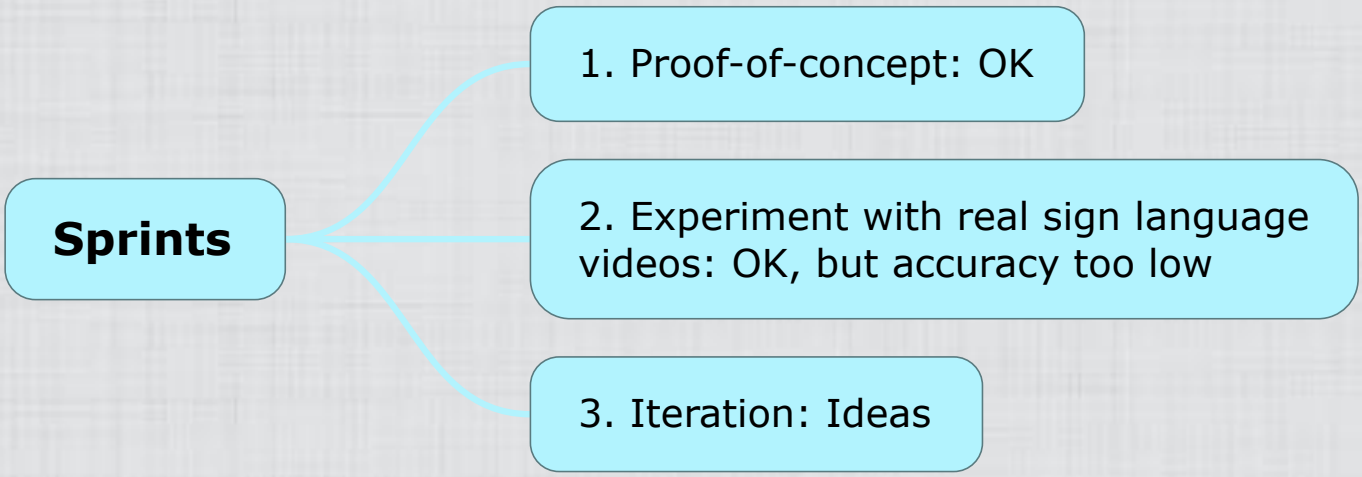

Sprints
1. Proof-of-concept: OK
2. Experiment with real sign language videos: OK, but accuracy too low
3. Iteration: Ideas

1. Proof-of-concept: OK

- Successfully cloned GitHub repository: video classification with Keras+Tensorflow github.com/harvitronix/five-video-classification-methods
- 13.000 human gesture videos, 100 categories, eg Apply Lipstick, Drumming
- 6h training on aws GPU machine,
  65%-74% accuracy on test set,
  25 sec prediction of single video on MacBook

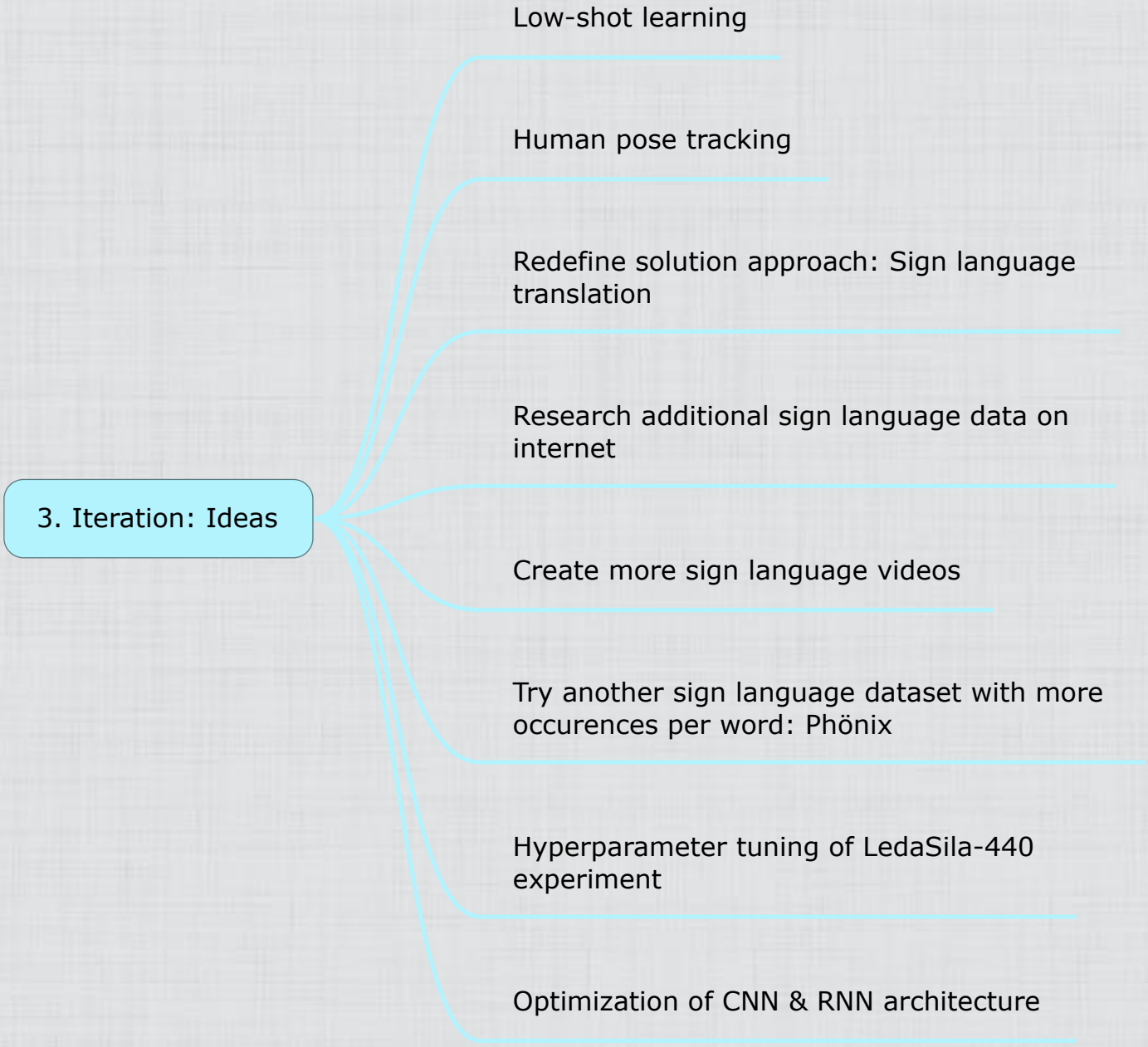
2. Experiment with real sign language videos: OK, but accuracy too low

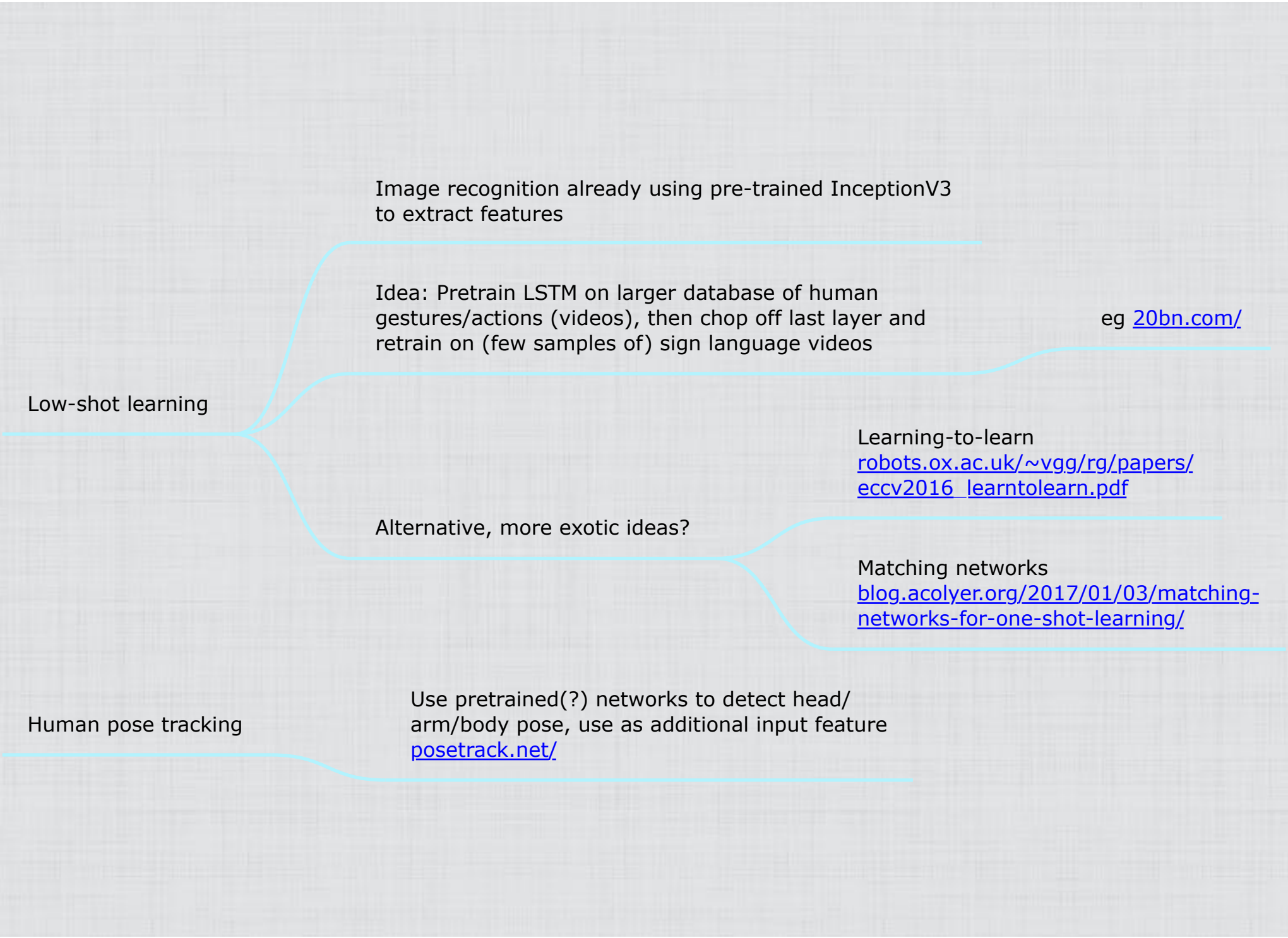
- LedaSila database (Uni Klagenfurt + TU Wien) containing Austrian sign language ÖGS
  33,300 training videos, 15,700 different words, very long tail
  github.com/FrederikSchorr/sign-language/blob/master/01-explore/02-explore-ledasila.ipynb
- Selected 440 videos with top 21 words, min 18 occurences each
- Videos sliced into 20 frames each. Features extracted with InceptionV3. Training with (new) LSTM.
- After 1.000 epochs (with almost no tuning): >30% accuracy on test set (> 80% on training)
- Hypothesis: Algorithm & pipeline work, but not enough training data => overfitting

30% accuracy

After 1.000 epochs (with almost no tuning): >30% accuracy on test set (> 80% on training)

but test loss steadily increasing


3. Iteration: Ideas
Low-shot learning
Human pose tracking
Redefine solution approach: Sign language translation
Research additional sign language data on internet
Create more sign language videos
Try another sign language dataset with more occurences per word: Phönix
Hyperparameter tuning of LedaSila-440 experiment
Optimization of CNN & RNN architecture


Low-shot learning
Image recognition already using pre-trained InceptionV3 to extract features
Idea: Pretrain LSTM on larger database of human gestures/actions (videos), then chop off last layer and retrain on (few samples of) sign language videos
eg 20bn.com/
Alternative, more exotic ideas?
Learning-to-learn robots.ox.ac.uk/~vgg/rg/papers/eccv2016_learntolearn.pdf
Matching networks blog.acolyer.org/2017/01/03/matching-networks-for-one-shot-learning/
Human pose tracking
Use pretrained(?) networks to detect head/arm/body pose, use as additional input feature posetrack.net/

Redefine solution approach: Sign language translation

- instead of recognising single isolated words, directly translate entire sentences www-i6.informatik.rwth-aachen.de/publications/download/1064/CamgozCihanHadfieldSimonKollerOscarNeyHermannBowdenRichard--NeuralSignLanguageTranslation--2018.pdf
- seem to be more datasets, but complexity significantly much higher

Research additional sign language data on internet

- eg SIGNUM by RWTH Aachen (1.000€)
- already asked specialized PhD Oscar Koller: there does not seem to be much more data out there

Create more sign language videos

- Contact local deaf-mute community and/or Humboldt-Uni to record additional signs
- Mecanical turk
- => time (and money?) consuming

Try another sign language dataset with more occurences per word: Phönix

- Total 3174 videos
  Top 16 words with 50-150 utterances, total 2,300 videos
  github.com/FrederikSchorr/sign-language/blob/master/01-explore/01-explore-phoenix.ipynb
- only 1 speaker, very homogenous

Hyperparameter tuning of LedaSila-440 experiment

- Eg more more videos/words: top 84 words with min 15 occurences, total of 1443 videos
- Tuning: Number of frames per video (currently 20), LSTM learningrate, size, init, dropout, ...

Optimization of CNN & RNN architecture

- later