INTRODUCTION

Computer vision for SLR (visual-spatial) encoding, front end

Machine translation for SLT (grammatical), back end, seq2seq

Goal: find methods without glosses needed

Best: Transformer + input embeddings from ResNet50 or pose based landmark features

GSL (controlled and constrained), ASL+CSL (less controlled)

Five parameters: handshape, location, palm orientation, body movement, facial

no representation in written language -> encode visually and spatially

models: with and without attention, reinforcement learning, transformer

**glosses**

once spelled words can be stored at a location in space and be referred to

classifiers

glosses include location, palm orientation, repetitions

**text-to-text translation**

LSTM (long term dep problems), attention mechanism, Transformers, GPT, BERT

**Video captioning**

Seq2seq video to text: two stacked LSTMs

Hierarchical RNNs

Reinforcement video captioning

Longer sequences: transformers, BERT, Vision and Language BERT, Transformer-XL

**Gesture Recognition**

Sequence modelling, classification problem

CNN (early, late, mid fusion temporal), recurrent networks

Openpose, openface library

Separate signs -> continuous SLR (need to consider prior context)

**Reinforcement learning**

Problems with LSTM seq2seq learning: teacher forcing -> exposure bias, non differentiable issues (no back propagation of evaluation metric

* RL solves those problems

**Video Sign Language Translation**

Word level sign language translation is easier (classification task)

Continuous sign language translation:

* long term dependencies, output series of words dependent on each other
* video captioning, temporal information included
* DeepASL: deep hierarchical bidirectional RNN, two way communication

Fang: DeepASL: Enabling ubiquitous and non-intrusive word and sentence-level

sign language translation, 2017

* Long/short term dependencies/fusion
* Connectionist temporal classification loss (CTC)

Graves: Connectionist temporal classification:

Labelling unsegmented sequence data with recurrent neural networks, 2006

* others
* Camgöz NMT, 2018: RNN+encoder/decoder attention

SPECIFIC MODELS FOR SLT

RNNs:

! CNN fix input

* RNN Arbitrary no of input frames

! RNN vanishing gradients (only useful for few frames)

* LSTM and GRU (gates/memory cell)

! LSTM and GRU vanishing gradients for long sequences

* Transformers and RL

Sequence to Sequence Modeling:

* Usually stacked LSTMs
* From frames extracted features fed one by one into encoder
* Encoder hidden state carries semantically rich features
* Use hidden states (latent embedding) in decoder
* Produce one word at a time with previous word

Sequence to Sequence Modeling with Attention:

* combination of seq2seq and attention mechanism
* linear combi of all timesteps encoder outputs (context) is additionally passed to decoder
* decoder output combined with context vector
* two different attention mechanisms: luong: multiplication, bahdanau: concatenation

Transformers:

* Camgoz: train features on CNN-LSTM-HMM architecture then transformer for transl
* LSTM: 100 output tokens, Transformer: 1000 output tokens
* Difference to RNN: all inputs at the same time, positional encoding
* Encoder decoder layers repeated N times
* Each layer: encoder: 2 sublayers, decoder: 3 sublayers
* K, q, v, multi-head -> concatenated in the end
* Then two layer FNN with ReLU
* Decoder second layer: query from previous layer, k,v from encoder
* Residual connections all across

Sequence to Sequence Modeling wth Reinforcement Learning:

* Environment: video, states: frames, agent: transformer with policy, action: prediction of words, reward = 0 until end of sentence
* Metric: CIDEr or WER
* Word sampling e.g. greedy
* Self critic loss function

DATASETS

German Sign Language:

* Grey homogeneous background, contrasts to clothes of signers
* PHOENIX Weather
* 25 fps
* 210x260px
* 9 signers
* 7096 training videos, 6811 unique utterances
* 1078 unique words, 1042 repeated 2-10 times
* Uneven distribution of videos per signer
* Only weather related sentences
* Better than other two datasets because controlled environment, most training samples and good distributions of val and test

American Sign Language:

* Different color scales
* 38 different topics
* 25fps
* 216x218px to 312x324px
* 7 signers
* 1.5-2mins (sub-videos) -> not matching anymore
* 1457 unique utterances

Chinese Sign Language:

* More complex background, more background (small signers), different backgrounds
* 50000 videos
* 10000 different utterances total
* 50 signers (50% female) each 1000 utterances
* Even distribution of videos per signer
* 30fps
* 1920x1080px
* Not native signers since standardized language is new

INPUT FEATURES

OpenPose Features:

* Body 25, hand2x 21, face 70 joint/landmark locations (x, y) -> 137\*2=274er vector
* Rescaling for CLSDataset
* Framt to frame smoothing (median filter, savitzky golay filter)

CNN Features:

* Mutlidimentional feature vector from alexnet and resnet pretrained on imagenet
* Pretrained features on imagenet:
  + Alexnet 4096 dim
  + Resnet50 24048 dim
  + Efficientnetb7 2560 dim
  + Inceptionv3 2048
* Ent to end trained:
  + Inceptionv1 1024 dim
  + resnet

Kmeans from OpenPose Features:

* 9521 clusters for similar poses (based on keypoints and cluster center)

ANALYSIS AND RESULTS

Performance Metrics

* BLEU Scores:
  + Compare n grams
  + fails if length of predicted sentence is less than gt -> brevity penalty
* OTHERS??

Results

* GSL training 7096 videos, valid 519 test 642
* CSL (sentence repetition) train 1790, 450d
* ASL 1031 training videos 340 test videos
* Training Details

Analysis of Results

* Variants of OpenPose for SLT

Easy to interpret for humans, hands most important keypoints, but not enough alone (combine with face and body), best if all three

* Choosing among Vanilla RNN, GRU and LSTM

Openpose keypoints input: rnn, gru, lstm

Rnn the worst -> gru, lstm better with long term dependencies

* Ablations with different CNN Features

Resnet50, alexnet, efficientnet-b7, inceptionv1, v3 as feature extractors, pretrained on imagenet

Resnet50 provides best results with transformer (deeper networks better than shallow, wider networks not better)

* Experiments and ablations on GSL Dataset

Transformer performs best due to more weights, openpose and resnet50 feature similar but openpose features less expensie

* Experiments and ablations on CSL Dataset

Dataset too small, many abnormalities

Attention mechanisms better with openpose than with resnet (openpose is centered, scaled and cut) -> due to structure of dataset

* Experiments and ablations on ASL Dataset

With attention->longer ngrams better (also GSL, CSL)

Seq2seq with attention best (not transformer)

Openpose, cnn feature approx. same

Structure of dataset is important, asl lies between gsl and csl

* Additional Experiments and Analysis using RL

Possible next step: include attention in rl model

* Human as an Oracle

Human performance compared to ground truth is not good, especially with openpose videos

CONCLUSION

Transformer with Openpose or ResNet50 input features the best (on GSL dataset)