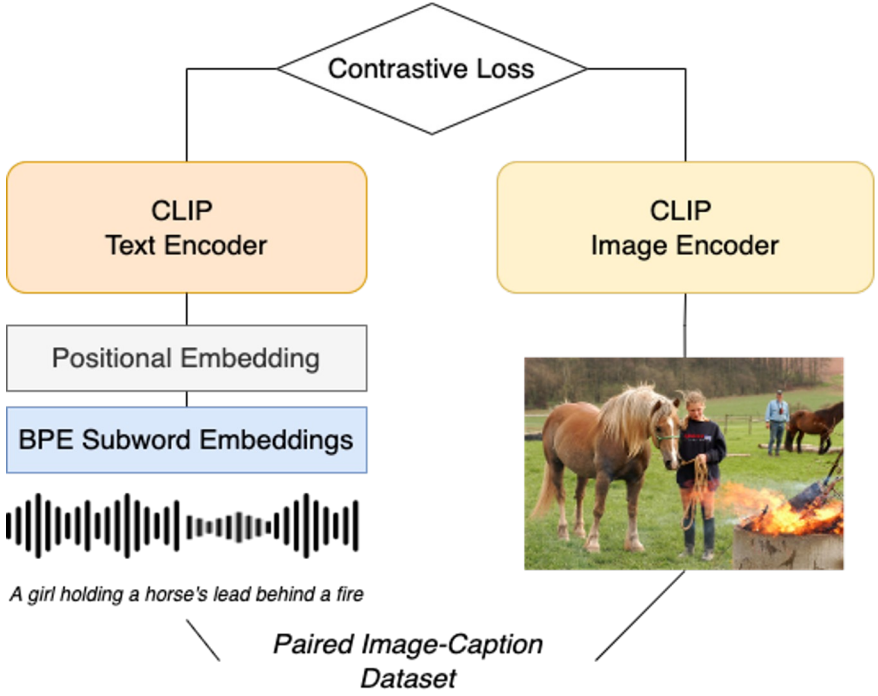
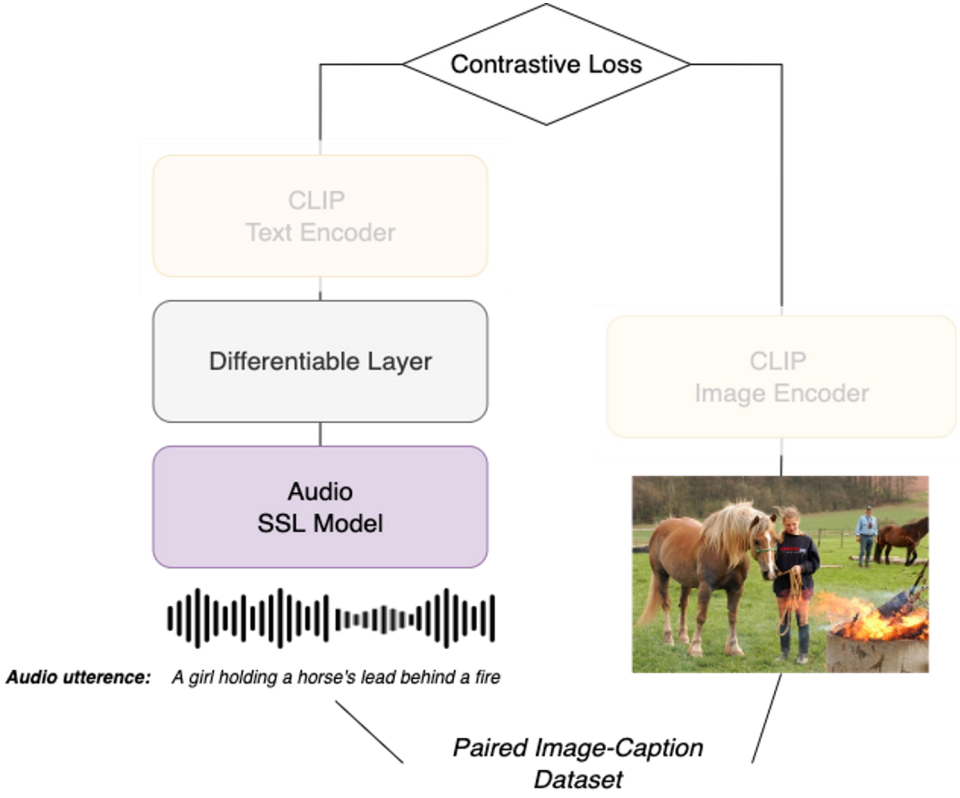
Cascaded CLIP



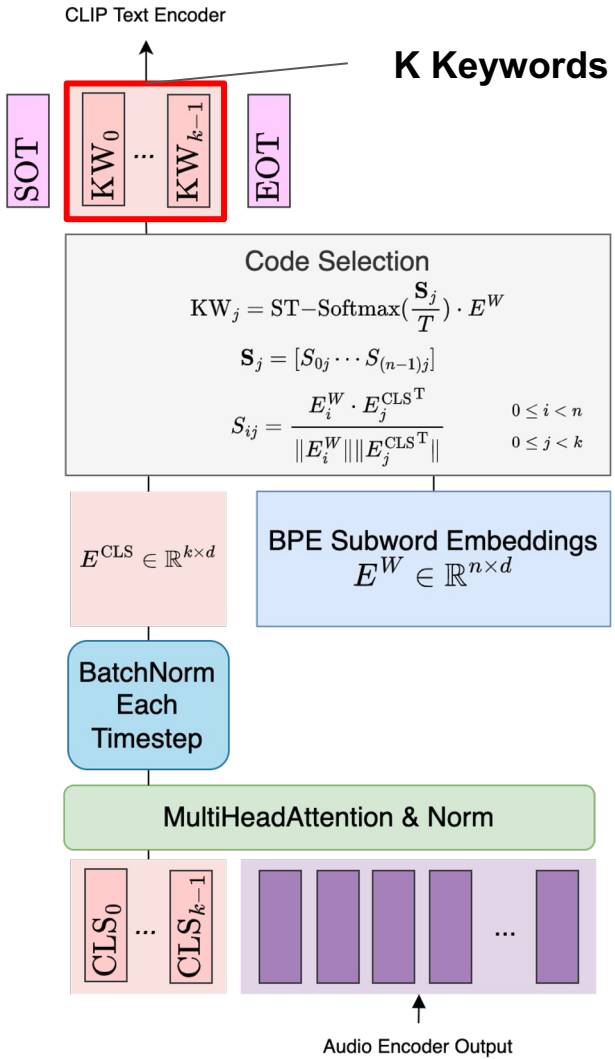
Original CLIP Setting



Cascaded CLIP Setting

Cascaded CLIP

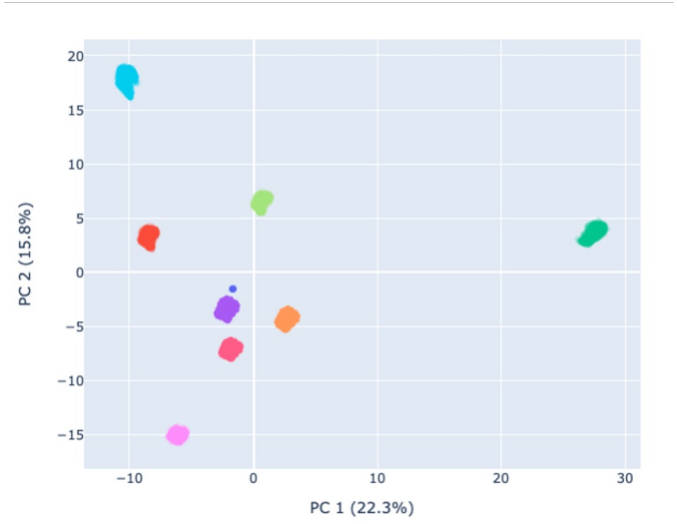
- Dataset : Flick8K
- Loss Function: InfoNCE
- bsz = 240, Adam (lr = 1e-4)
- Cosine logits temperature T = 0.1



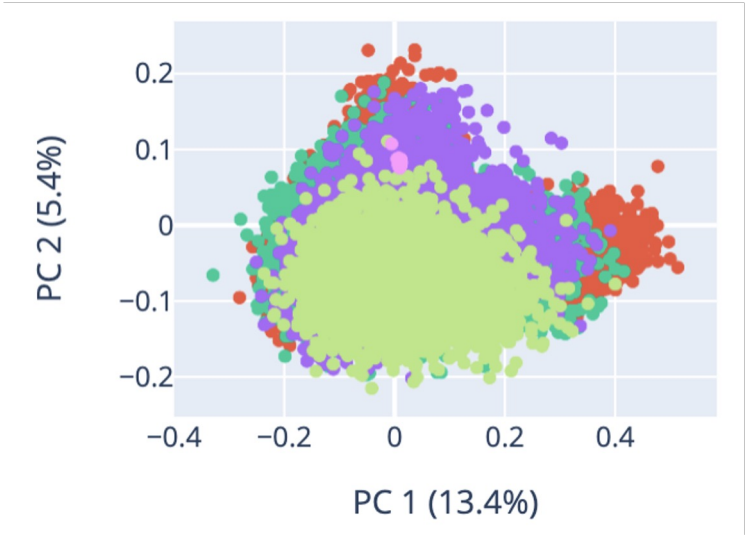
BatchNorm

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

w/o BatchNorm



w/ BatchNorm



Detokenize KW (w/o BN)

"gold": "Two girls play on a skateboard in a courtyard"

"neighbors": "nationalpark</w>", 0.1894872784614563

"gold": "An outdoor ice skating rink full of people",

"neighbors": "foxsports</w>", 0.1776472628116607 7

"gold": "A white dog watching a black dog in the air",

"neighbors": "frequency</w>", 0.1845706701278686 5

Detokenize KW (w/ BN)

Input: **A woman is throwing water on a child in a plastic swimming pool in a rural area**

kw0 "participating</w>", "participate</w>", "participates</w>", "participant</w>", "attending</w>",

kw1 "child</w>", "children</w>", "baby</w>", "boy</w>", "kid</w>",

kw2 "relaxing</w>", "relax</w>", "peaceful</w>", "calming</w>", "relax",

kw3 "enjoying</w>", "enjoy</w>", "enjoys</w>", "celebrating</w>", "having</w>",

kw4 "swim</w>", "swimming</w>", "swim", "swims</w>", "swimmers</w>",

kw5 "swimming</w>", "swim</w>", "swim", "swims</w>", "swimmer</w>",

kw6 "bikin", "bikini</w>", "kini</w>", "manne", "naked</w>", "thong</w>",

kw7 "kid", "kid</w>", "kids</w>", "child</w>", "children</w>",

Recall

- The ***recall degrades*** a lot when confining our speech encoder's output to the subword embeddings, but it provides our model with ***the ability to map audio to some keywords***.

	recall_mean@1	recall_mean@5	recall_mean@10
Cascaded	6.29	19.77	29.77
Parallel	23.00	56.20	65.90

Zerospeech 2021 Semantic

	LibriSpeech	Synthetic
CascadedCLIP (Flickr8k)	16.00	7.53
Hubert Base	15.89	2.19
Fast Vgs	23.55	15.8
Baseline	15.09	9.6

SUPERB

	semantics			content		
	IC_Acc (⬆️)	SF_F1 (⬆️)	SF_CER (⬇️)	PR (⬇️)	KS (⬆️)	Qbe (⬆️)
Cascaded CLIP	98.05	87.8	26.6	**5.26**	96.75	7.36
Hubert Base	98.34	88.53	25.3	5.41	96.3	7.36
Fast Vgs+	98.97	88.15	27.12	7.76	97.27	5.26

Conclusion & Next Steps

- Conclusion:
  - Cascaded models have better performances on SUPERB ***Content*** downstream task.
  - Cascaded models have the ability to map some ***spoken words*** in the utterence directly to ***semantic related subwords***.
- Next steps:
  - Focusing on improving ***Content***-related downstream tasks on SUPERB
    - Compare performances of downstream tasks for *Parallel CLIP*, *Cascaded CLIP* and *Parallel+Cascaded CLIP*
  - Changing SSL models
  - Moving to larger dataset: SpokenCOCO or Places Audio Dataset

Leveraging Cascaded-SpeechCLIP

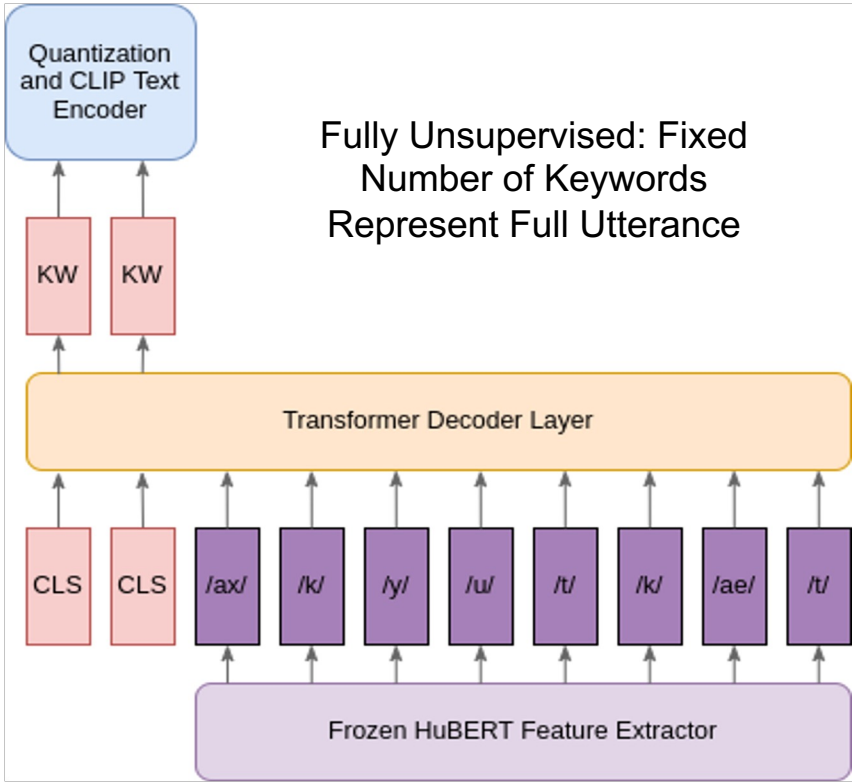
Goal: Given a dataset of images and corresponding spoken captions, use the mapping learned from the input speech to a sequence of CLIP tokens to generate English text.

- If the input language is English, this corresponds to ASR with no parallel text supervision (similar to wav2vec-u 2.0) *and* the additional restriction of no pronunciation lexicon.
- If the input language is non-English, this corresponds to S2T translation with no source language text *and* no parallel source and target language supervision.

Challenge: CLIP is largely invariant to word order, and nearly as accurate when given a few keywords as when given a full transcription.

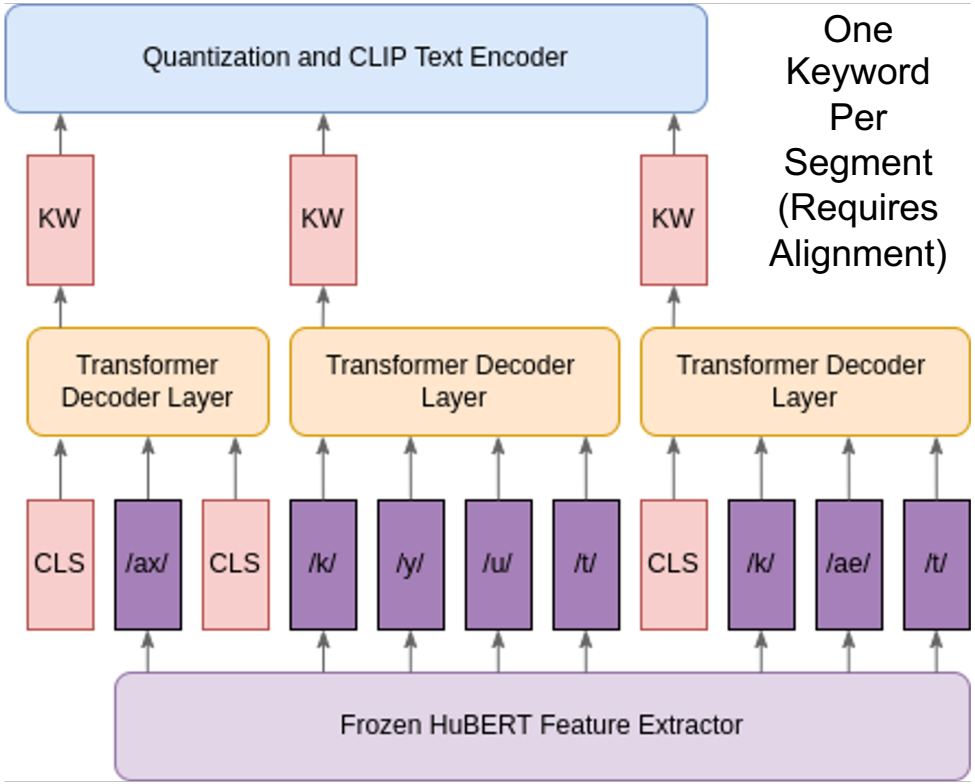
Response: Predict CLIP tokens on a word-by-word basis, rather than all at once for a full utterance.

- Our initial approach uses ground truth alignments to segment the utterance. If the second word in an utterance spans 0.5-1.2s, and HuBERT frames are extracted at 50Hz, the second CLIP token generated is predicted based only on HuBERT frames 25-60.
- Ultimately, we aim to automatically segment input speech to eliminate the additional supervision of ground truth alignments.



Fully Unsupervised: Fixed Number of Keywords Represent Full Utterance

“a cute cat”



One Keyword Per Segment (Requires Alignment)

“a cute cat”

Example Predicted Sequences

a|shirt man|hand feeding|giving a|have giraffe|animals food|pizza with|taking his|giving mouth|have

photograph|seen of|has an|have outdoor|same arena|which that|into looks|into neat|taking

man|hand giving|leading the|taking peace|delicious sign|streets to|have a|have lady|giving taking|giving a|having picture|right with|into a|into smart|making phone|phone

Word Discovery Analysis on SpokenCOCO Val Set

Methodology:

1. Use our trained model to label each segment in the validation set
2. For each ground truth word that appears, collect a list of every label it is assigned
3. If a word appears less than 10 times total, remove it from the set of words being analyzed
4. If a word *w* appears *N* times and receives its most common label *k* times, assign the model a recognition score of *k/N* for that word

Average Recognition Score: **61.46%**
31.15% of words get a score of 75% or higher

Pattern Type	Examples
Perfect Match	bathroom→bathroom (93.75%) kitchen→kitchen (100%) skateboard→skateboard (95.38%) vegetables→vegetables (82.76%) truck→truck (97.62%)
Semantically Related	elephants→cattle (89.29%) dark→seen (100%) parked→stopped (92.68%) rock→forest (100%)
Bucketing	street, train, traffic, intersection → cars meat, food, plate, sandwich → food soccer, tennis, court → frisbee woman, men, women, girls, guy → players skis, snow, skier, skiing, ski → skiing
Default Token	a, on, the, has, white, in, his, and, at, with, is, their, its, up, to, from, that, green, next, front, various, one, some, …. → into (total: 66!)
Unexpected	benches→appears (76.92%) snowy→following (85.71%) glass→together (78.57%) player, baseball, bat, batter → toothbrush

Next Steps and Future Directions

Step 1: Add formal supervision alongside semantic supervision

- Minimize perplexity of predicted sequences according to a pretrained LM over CLIP tokens
- Adversarial training with a discriminator that sees predicted sequences and non-parallel ground truth captions tokenized by CLIP

Step 2: Automatic segmentation

- Use pretrained VG-HuBERT (Peng and Harwath 2022)
- Allow segmenter to be fine-tuned as part of Cascaded-SpeechCLIP

Step 3: Speech to text translation to English with no parallel data

- Input Places400k Hindi and Japanese spoken captions
- Adapt frame representation extraction and segmentation modules by language