# End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira,
Ivan Laptev, Josef Sivic, Andrew Zisserman


Julia Guerrero Viu
Seminar on Current Works in Computer Vision
Summer Semester 2020

UNI FREIBURG

# Outline

1. Motivation
2. Related Work
3. The MIL-NCE objective
4. Experiments
5. Conclusion and discussion impulses

# Motivation

- **Objective:** Learning video representations
- **Challenge:** Most current video representation models require extensive annotations. Annotating videos is expensive and not scalable
- **Possible solution:** Leveraging narrated videos that are available at scale on the web

# Motivation

- **Narrated / instructional videos:** Videos that include an oral description of what it is happening
- **HowTo100M:** dataset of 100 million clips-narrations from YouTube


- **Challenges:** Narration supervision is weak and noisy.
  In particular, weak alignment between text and image (~50% misalignment)

# Motivation

In this work...

- Learning embeddings of video and text in a self-supervised manner directly from uncurated instructional videos

- **MIL-NCE objective:** New specific loss to address misalignments in narrated videos

MIL
Multiple Instance Learning

**MIL-NCE**

NCE
Noise Contrastive Estimation

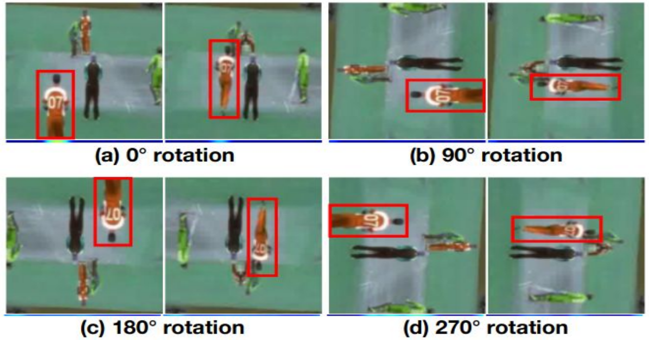
# Related Work

1. Self-supervised learning on videos
2. Joint video-language
3. Multiple Instance Learning
4. Noise Contrastive Estimation

# Self-supervised learning on videos

- Use metadata from social media videos as labels [Ghadiyaram et al. 2019]
- **Self-supervised**: learn a proxy task with labels taken directly from videos

Geometric transformations [Jing et al. 2018]




...

Predicting the future [Han et al. 2019]

- Domain gap between curated and uncurated videos [Caron et al. 2019]

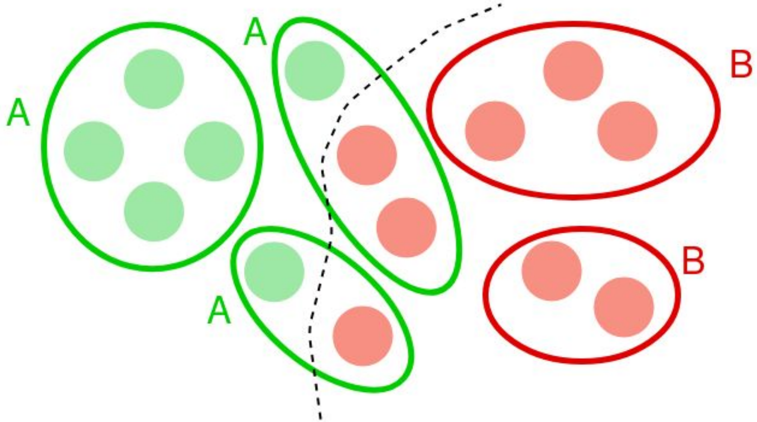
# Joint video-language

- Learn a **joint embedding space** for visual and textual data
    - Supervised: Manual annotated datasets
    - Self-Supervised: Exploit **semantic** information from natural language (audio speech or Automatic Speech Recognition)

- [Miech et al. HowTo100M 2019] and [Sun et al. CBT 2019]:
  Self-supervised using ASR but leverage **pre-trained** visual representations on ImageNet and Kinetics

# Multiple Instance Learning (MIL)

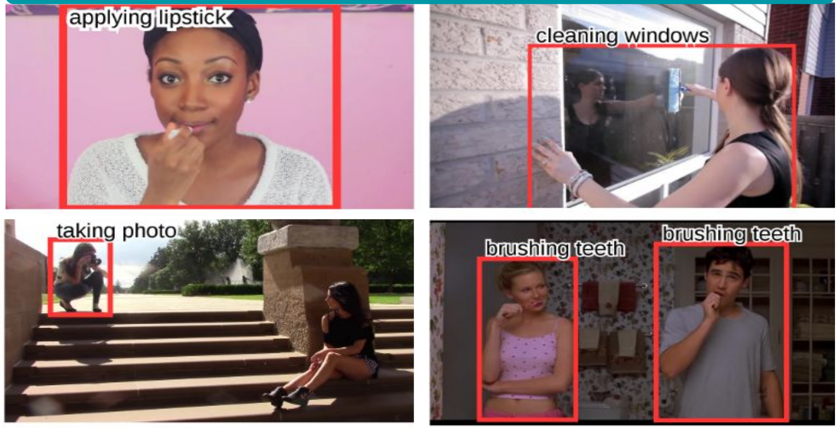
- Weakly supervised learning with **labeled sets** of many samples instead of individual labels per sample
- Multiple instances of the same object, where the label refers to the object
- MIL deals with problems with incomplete knowledge



# Multiple Instance Learning (MIL)

- MIL applied to video understanding
  - Max-pooling: MIL-SVM
  - Discriminative clustering: DIFFRAC

Action localization [Weinzaepfe et al. 2016]



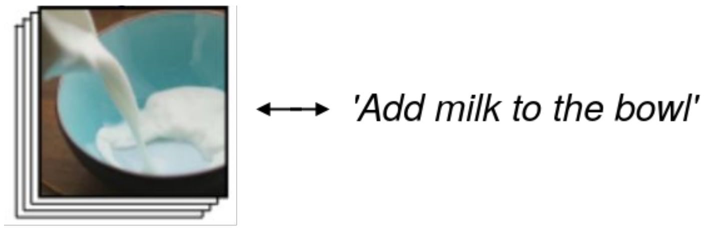

Person recognition in movies [Miech et al. 2017]

# Noise Contrastive Estimation (NCE)

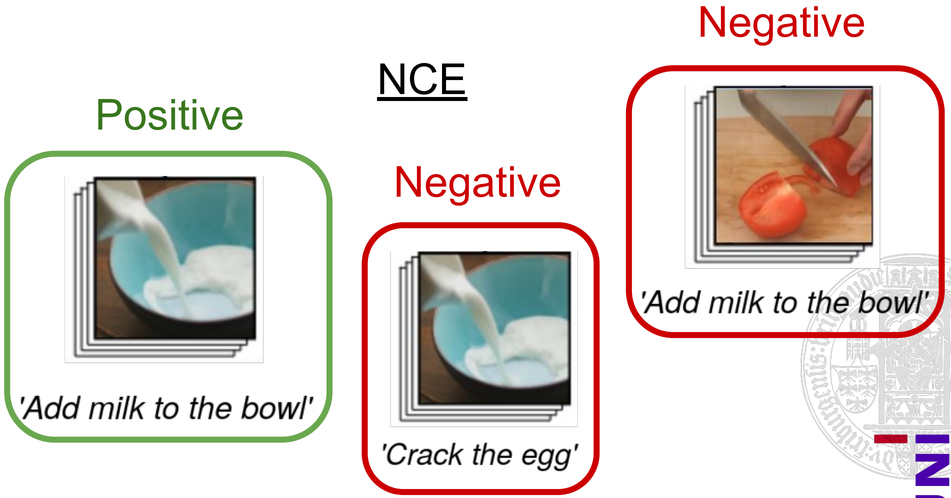
- Discriminate between samples from a ‘real’ distribution and an artificially generated noise distribution

- Used to train classifiers with a very large number of classes

Conventional Classifier




NCE



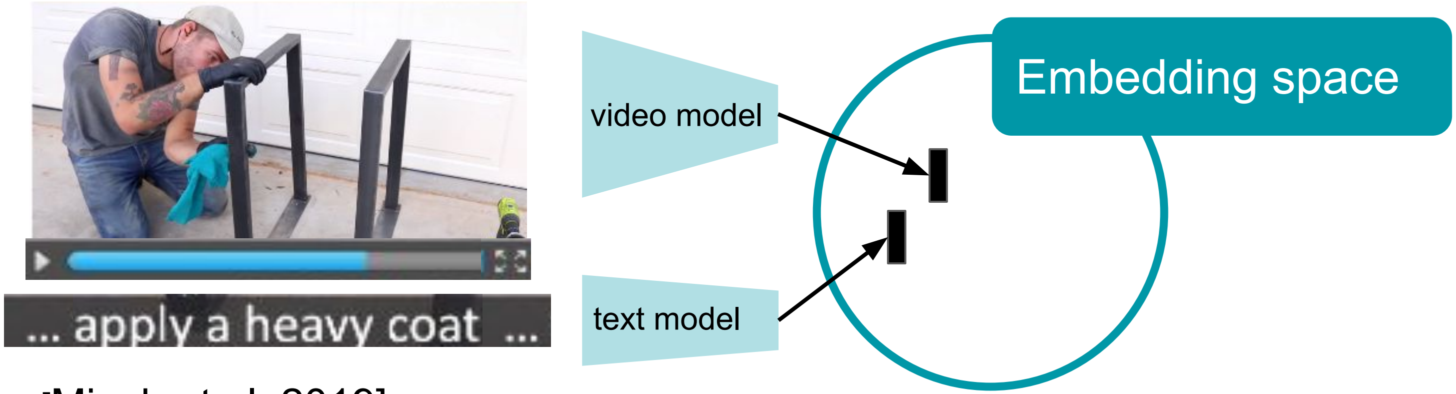
# Noise Contrastive Estimation (NCE)

- [Hénaff et al. 2019] and [Van den Oord et al. 2018] apply NCE to self-supervised learning using *InfoNCE* loss

- [Sun et al. CBT 2019] apply NCE loss to video-text representation learning:
  - Different way of constructing the negative samples

# MIL-NCE objective

- Learning joint embedding space from **video** and **text**
- Embedding similarity when text and video content **semantically similar**




[Miech et al. 2019]

# MIL-NCE objective

$$\max_{f,g} \sum_{i=1}^{n} \log \left( \frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

UNI FREIBURG

# MIL-NCE objective

**f** and **g**: embedding functions

**x**: video-clip **y**: narration

$$\max_{f,g} \sum_{i=1}^{n} \log \left( \frac{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum\limits_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

**P**: positive candidates **N**: negative candidates

- “Maximizing ratio of the sum of positive candidate scores to the sum of negative samples scores, where score is exponentiated dot product of the embeddings”

# MIL-NCE objective

Building the MIL-NCE objective step by step…

1. Simple joint probabilistic model
2. MIL contribution: Multiple options for matching video with narration
3. NCE contribution

# Joint probabilistic model

- **Input:** Set of *n* video-text pairs from the joint data distribution

$$\{(x_i, y_i)\}_{i=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n$$

- **Output:** Two parametrized functions *f* and *g* that map video and text to a *d*-dimensional vector space

$$f : \mathcal{X} \to \mathbb{R}^d \qquad g : \mathcal{Y} \to \mathbb{R}^d$$

- **Probability** of a matching pair (*x* , *y*) can be estimated up to a constant as:

$$\boxed{p(x, y; f, g) \propto e^{f(x)^\top g(y)}}$$

UNI FREIBURG

# MIL contribution

- Key idea: Consider **multiple options** to match a video with a narration

- Given a video-clip *x*, *K* positive narrations that are close in time
  Joint probability of *x* happening with any of the $y_k$ (mutually exclusive):

$$p(\cup_k\{(x, y_k)\}) = \sum_k p(x, y_k) \propto \sum_k e^{f(x)^\top g(y_k)}$$

- **Symmetric** joint probability $(x, y) \longrightarrow \mathcal{P} = (x^k, y^k)_{k=1}^{K}$

$$\boxed{p(\mathcal{P}) \propto \sum_{(x,y)\in\mathcal{P}} e^{f(x)^\top g(y)}}$$

# NCE contribution

- A lot of all possible pairs of video-text: intractable for *generative* loss
- *Discriminative* loss: softmax version of NCE [Jozefowicz et al. 2016]

$$\max_{f,g} \sum_{i=1}^{n} \log\left(\frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x_i)^\top g(y_i)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x_i)^\top g(y_i)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}}\right)$$

# NCE contribution

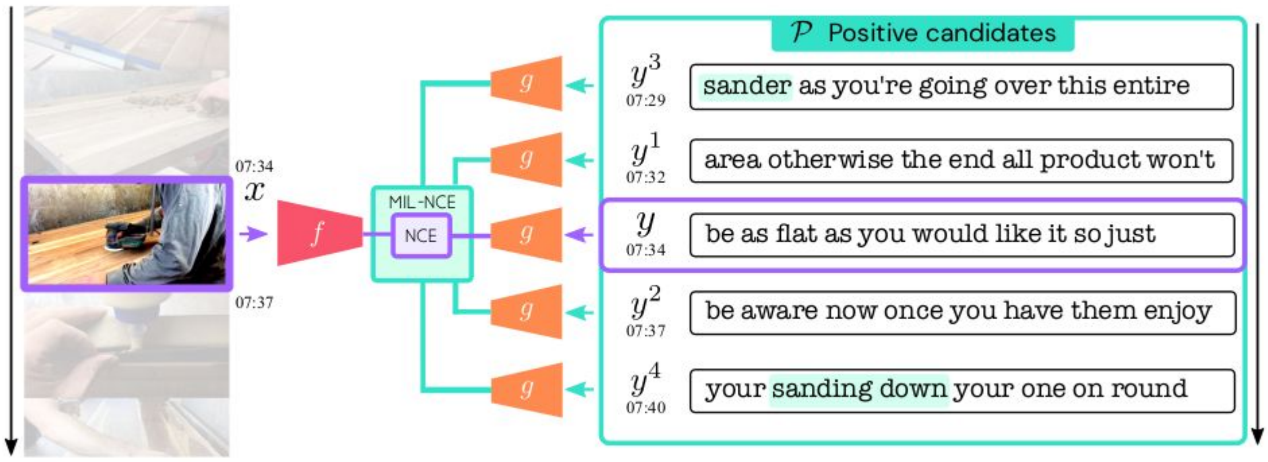
- A lot of all possible pairs of video-text: intractable for *generative* loss
- *Discriminative* loss: softmax version of NCE [Jozefowicz et al. 2016]

**MIL-NCE**

$$\max_{f,g} \sum_{i=1}^{n} \log \left( \frac{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum\limits_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

# MIL-NCE objective

- NCE: Discriminate between positive and negative candidates
- Model a symmetric joint probability between text and video
- MIL: Solve the **temporal misalignments** between text and video



# Experiments

1. Implementation details

2. Downstream tasks

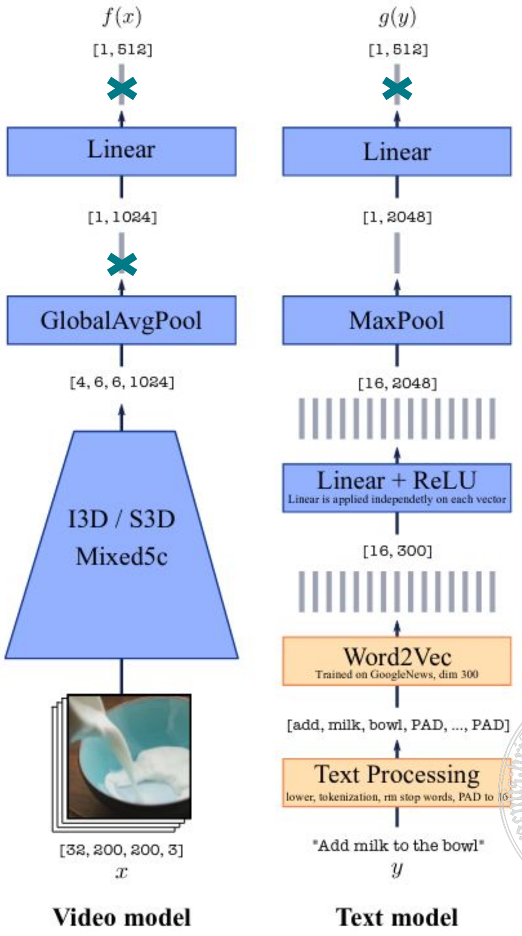
3. Ablation studies

4. Comparison to state-of-the-art

# Implementation details

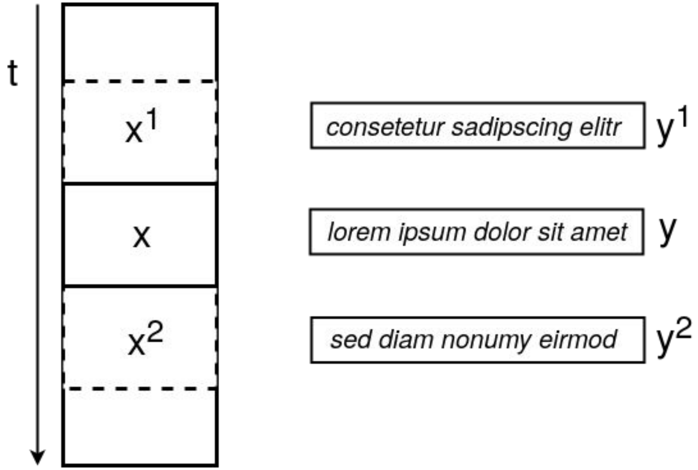
- **Video-model:**
  I3D [Carreira et al. 2017] / S3D [Xie et al. 2018]
- **Text-model:**
  word2Vec pre-trained on Google News [Mikolov et al. 2013]

- Train on **HowTo100M dataset**
  - 120M pairs → 15 years
  - 3.2 seconds video (32 frames)
  - Automatic Speech Recognition
  - max. 16 words narration

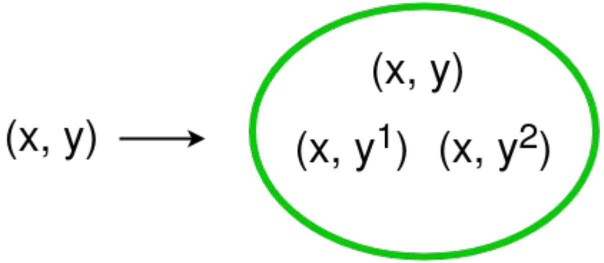


# Implementation details

- **Positive samples**



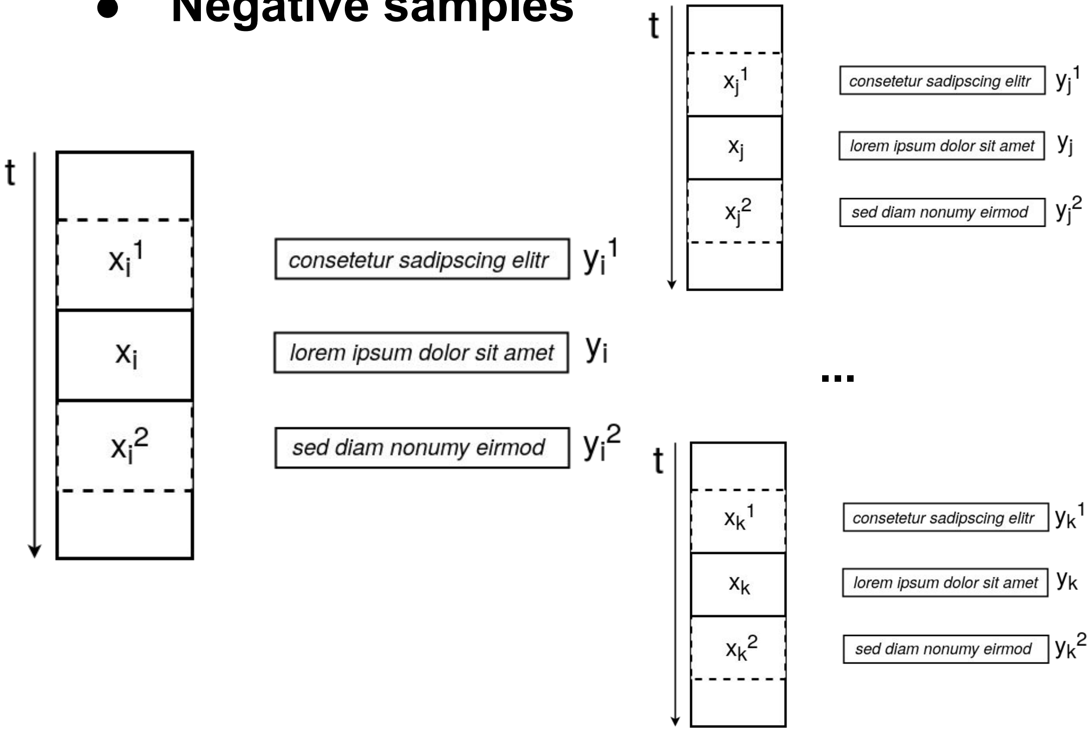

Size of the positive set: $|P| = 3$



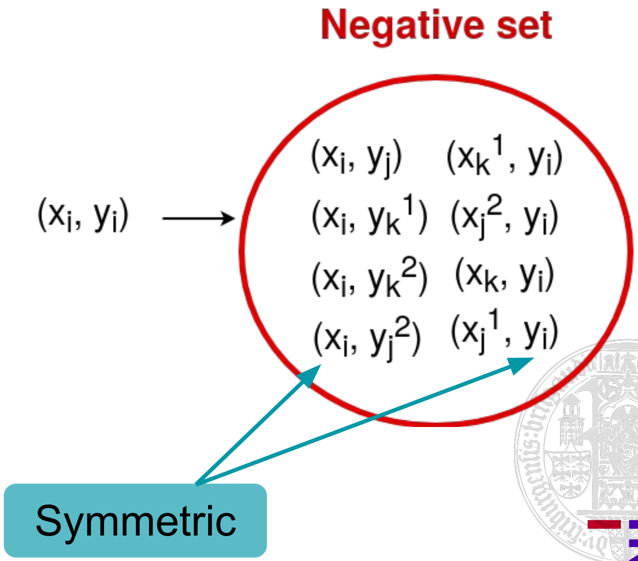

**Positive set**

# Implementation details

- **Negative samples**




Size of the negative set: $|N| = 8$



# Downstream tasks

## Action Recognition


**datasets**
HMDB-51
UCF-101
Kinetics700

**metric**
accuracy

## Temporal Action Localization


**dataset**
Youtube8M Segments

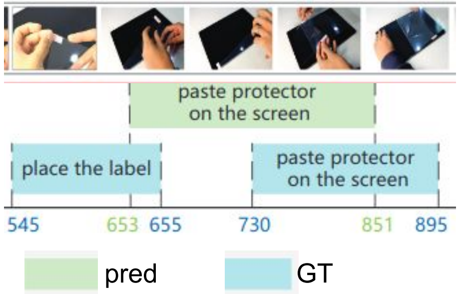
**metric**
mAP

## Action Step Localization


**dataset**
CrossTask (CTR)

**metric**
average recall

## Temporal Action Segmentation




**dataset**
COIN

**metric**
frame accuracy (FA)

## Text-to-Video Retrieval

**Input:** cut tomato

**Output:**

**datasets**
YouCook2 (YR10)
MSR-VTT (MR10)

**metric**
recall@K

UNI FREIBURG

# Ablation Studies

(a) **Training loss**

| Loss | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Binary-Classif | 18.5 | 23.1 | 32.6 | 44.2 | 68.5 |
| Max margin | 16.3 | 24.1 | 29.3 | **56.2** | 76.6 |
| NCE | **29.1** | **27.0** | **35.6** | 55.4 | **77.5** |

(b) **Negatives per positive**

| $\|\mathcal{N}\|$ | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| 64 | 26.0 | 25.5 | 33.1 | 56.1 | 76.0 |
| 128 | 27.1 | 26.4 | 33.3 | **57.2** | 76.2 |
| 256 | 28.7 | 28.7 | **36.5** | 56.5 | **77.5** |
| 512 | **28.8** | **29.0** | 35.6 | 55.4 | 77.4 |

(c) **Number of positive candidate pair**

| | NCE | MIL-NCE | | | | |
|---|---|---|---|---|---|---|
| $\|\mathcal{P}\| \rightarrow$ | 1 | 3 | 5 | 9 | 17 | 33 |
| YR10 | 29.1 | 33.6 | **35.0** | 33.1 | 32.4 | 28.3 |
| MR10 | 27.0 | 30.2 | **31.8** | 30.5 | 29.2 | 30.4 |
| CTR | 35.6 | **37.3** | 34.2 | 31.8 | 25.0 | 25.0 |
| HMDB | 55.4 | **57.8** | 56.7 | 55.7 | 54.8 | 51.4 |
| UCF | 77.5 | 79.7 | **80.4** | 79.5 | 78.5 | 77.9 |

(d) **MIL strategy**

| Method | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Cat+NCE | 31.9 | 30.8 | **35.2** | 56.3 | 78.9 |
| Max+NCE | 32.3 | 31.3 | 32.2 | 55.3 | 79.2 |
| Attn+NCE | 32.4 | 30.2 | 33.4 | 55.2 | 78.4 |
| MIL-NCE | **35.0** | **31.8** | 34.2 | **56.7** | **80.4** |

(e) **Symmetric vs asymmetric negatives**

| Negatives | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| $(x\|y)$ | 34.4 | 29.0 | 33.9 | 55.1 | 78.1 |
| $(y\|x)$ | 19.3 | 19.4 | 28.2 | **57.1** | 79.2 |
| $(x, y)$ | **35.0** | **31.8** | **34.2** | 56.7 | **80.4** |

(f) **Language models**

| Text model | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| LSTM | 16.6 | 15.6 | 23.8 | 53.1 | 80.1 |
| GRU | 16.8 | 16.9 | 22.2 | 54.7 | **82.8** |
| Transformer | 26.7 | 26.5 | 32.7 | 53.4 | 78.4 |
| NetVLAD | 33.4 | 29.2 | **35.5** | 51.8 | 79.3 |
| Ours | **35.0** | **31.8** | 34.2 | **56.7** | 80.4 |

# Ablation Studies

(a) **Training loss**

| Loss | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Binary-Classif | 18.5 | 23.1 | 32.6 | 44.2 | 68.5 |
| Max margin | 16.3 | 24.1 | 29.3 | **56.2** | 76.6 |
| NCE | **29.1** | **27.0** | **35.6** | 55.4 | **77.5** |

(b) **Negatives per positive**

| $\|\mathcal{N}\|$ | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| 64 | 26.0 | 25.5 | 33.1 | 56.1 | 76.0 |
| 128 | 27.1 | 26.4 | 33.3 | **57.2** | 76.2 |
| 256 | 28.7 | 28.7 | **36.5** | 56.5 | **77.5** |
| 512 | **28.8** | **29.0** | 35.6 | 55.4 | 77.4 |

(c) **Number of positive candidate pair**

| | NCE | MIL-NCE | | | | |
|---|---|---|---|---|---|---|
| $\|\mathcal{P}\| \rightarrow$ | 1 | 3 | 5 | 9 | 17 | 33 |
| YR10 | 29.1 | 33.6 | **35.0** | 33.1 | 32.4 | 28.3 |
| MR10 | 27.0 | 30.2 | **31.8** | 30.5 | 29.2 | 30.4 |
| CTR | 35.6 | **37.3** | 34.2 | 31.8 | 25.0 | 25.0 |
| HMDB | 55.4 | **57.8** | 56.7 | 55.7 | 54.8 | 51.4 |
| UCF | 77.5 | 79.7 | **80.4** | 79.5 | 78.5 | 77.9 |

(d) **MIL strategy**

| Method | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Cat+NCE | 31.9 | 30.8 | **35.2** | 56.3 | 78.9 |
| Max+NCE | 32.3 | 31.3 | 32.2 | 55.3 | 79.2 |
| Attn+NCE | 32.4 | 30.2 | 33.4 | 55.2 | 78.4 |
| MIL-NCE | **35.0** | **31.8** | 34.2 | **56.7** | **80.4** |

(e) **Symmetric vs asymmetric negatives**

| Negatives | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| $(x\|y)$ | 34.4 | 29.0 | 33.9 | 55.1 | 78.1 |
| $(y\|x)$ | 19.3 | 19.4 | 28.2 | **57.1** | 79.2 |
| $(x, y)$ | **35.0** | **31.8** | **34.2** | 56.7 | **80.4** |

(f) **Language models**

| Text model | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| LSTM | 16.6 | 15.6 | 23.8 | 53.1 | 80.1 |
| GRU | 16.8 | 16.9 | 22.2 | 54.7 | **82.8** |
| Transformer | 26.7 | 26.5 | 32.7 | 53.4 | 78.4 |
| NetVLAD | 33.4 | 29.2 | **35.5** | [illegible] | [illegible] |
| Ours | **35.0** | **31.8** | 34.2 | [illegible] | [illegible] |

UNI FREIBURG

# Ablation Studies

(a) **Training loss**

| Loss | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Binary-Classif | 18.5 | 23.1 | 32.6 | 44.2 | 68.5 |
| Max margin | 16.3 | 24.1 | 29.3 | **56.2** | 76.6 |
| NCE | **29.1** | **27.0** | **35.6** | 55.4 | **77.5** |

(b) **Negatives per positive**

| $\|\mathcal{N}\|$ | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| 64 | 26.0 | 25.5 | 33.1 | 56.1 | 76.0 |
| 128 | 27.1 | 26.4 | 33.3 | **57.2** | 76.2 |
| 256 | 28.7 | 28.7 | **36.5** | 56.5 | **77.5** |
| 512 | **28.8** | **29.0** | 35.6 | 55.4 | 77.4 |

(c) **Number of positive candidate pair**

| | NCE | MIL-NCE | | | | |
|---|---|---|---|---|---|---|
| $\|\mathcal{P}\| \rightarrow$ | 1 | 3 | 5 | 9 | 17 | 33 |
| YR10 | 29.1 | 33.6 | **35.0** | 33.1 | 32.4 | 28.3 |
| MR10 | 27.0 | 30.2 | **31.8** | 30.5 | 29.2 | 30.4 |
| CTR | 35.6 | **37.3** | 34.2 | 31.8 | 25.0 | 25.0 |
| HMDB | 55.4 | **57.8** | 56.7 | 55.7 | 54.8 | 51.4 |
| UCF | 77.5 | 79.7 | **80.4** | 79.5 | 78.5 | 77.9 |

(d) **MIL strategy**

| Method | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Cat+NCE | 31.9 | 30.8 | **35.2** | 56.3 | 78.9 |
| Max+NCE | 32.3 | 31.3 | 32.2 | 55.3 | 79.2 |
| Attn+NCE | 32.4 | 30.2 | 33.4 | 55.2 | 78.4 |
| MIL-NCE | **35.0** | **31.8** | 34.2 | **56.7** | **80.4** |

(e) **Symmetric vs asymmetric negatives**

| Negatives | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| $(x\|y)$ | 34.4 | 29.0 | 33.9 | 55.1 | 78.1 |
| $(y\|x)$ | 19.3 | 19.4 | 28.2 | **57.1** | 79.2 |
| $(x, y)$ | **35.0** | **31.8** | **34.2** | 56.7 | **80.4** |

(f) **Language models**

| Text model | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| LSTM | 16.6 | 15.6 | 23.8 | 53.1 | 80.1 |
| GRU | 16.8 | 16.9 | 22.2 | 54.7 | **82.8** |
| Transformer | 26.7 | 26.5 | 32.7 | 53.4 | 78.4 |
| NetVLAD | 33.4 | 29.2 | **35.5** | [illegible] | [illegible] |
| Ours | **35.0** | **31.8** | 34.2 | [illegible] | [illegible] |

# Ablation Studies

**(a) Training loss**

| Loss | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Binary-Classif | 18.5 | 23.1 | 32.6 | 44.2 | 68.5 |
| Max margin | 16.3 | 24.1 | 29.3 | **56.2** | 76.6 |
| NCE | **29.1** | **27.0** | **35.6** | 55.4 | **77.5** |

**(b) Negatives per positive**

| $\|\mathcal{N}\|$ | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| 64 | 26.0 | 25.5 | 33.1 | 56.1 | 76.0 |
| 128 | 27.1 | 26.4 | 33.3 | **57.2** | 76.2 |
| 256 | 28.7 | 28.7 | **36.5** | 56.5 | **77.5** |
| 512 | **28.8** | **29.0** | 35.6 | 55.4 | 77.4 |

**(c) Number of positive candidate pair**

| | NCE | MIL-NCE | | | | |
|---|---|---|---|---|---|---|
| $\|\mathcal{P}\| \rightarrow$ | 1 | 3 | 5 | 9 | 17 | 33 |
| YR10 | 29.1 | 33.6 | **35.0** | 33.1 | 32.4 | 28.3 |
| MR10 | 27.0 | 30.2 | **31.8** | 30.5 | 29.2 | 30.4 |
| CTR | 35.6 | **37.3** | 34.2 | 31.8 | 25.0 | 25.0 |
| HMDB | 55.4 | **57.8** | 56.7 | 55.7 | 54.8 | 51.4 |
| UCF | 77.5 | 79.7 | **80.4** | 79.5 | 78.5 | 77.9 |

**(d) MIL strategy**

| Method | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Cat+NCE | 31.9 | 30.8 | **35.2** | 56.3 | 78.9 |
| Max+NCE | 32.3 | 31.3 | 32.2 | 55.3 | 79.2 |
| Attn+NCE | 32.4 | 30.2 | 33.4 | 55.2 | 78.4 |
| MIL-NCE | **35.0** | **31.8** | 34.2 | **56.7** | **80.4** |

**(e) Symmetric vs asymmetric negatives**

| Negatives | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| $(x\|y)$ | 34.4 | 29.0 | 33.9 | 55.1 | 78.1 |
| $(y\|x)$ | 19.3 | 19.4 | 28.2 | **57.1** | 79.2 |
| $(x, y)$ | **35.0** | **31.8** | **34.2** | 56.7 | **80.4** |

**(f) Language models**

| Text model | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| LSTM | 16.6 | 15.6 | 23.8 | 53.1 | 80.1 |
| GRU | 16.8 | 16.9 | 22.2 | 54.7 | **82.8** |
| Transformer | 26.7 | 26.5 | 32.7 | 53.4 | 78.4 |
| NetVLAD | 33.4 | 29.2 | **35.5** | [illegible] | [illegible] |
| Ours | **35.0** | **31.8** | 34.2 | 56.7 | [illegible] |

# Ablation Studies

(a) **Training loss**

| Loss | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Binary-Classif | 18.5 | 23.1 | 32.6 | 44.2 | 68.5 |
| Max margin | 16.3 | 24.1 | 29.3 | **56.2** | 76.6 |
| NCE | **29.1** | **27.0** | **35.6** | 55.4 | **77.5** |

(b) **Negatives per positive**

| $\|\mathcal{N}\|$ | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| 64 | 26.0 | 25.5 | 33.1 | 56.1 | 76.0 |
| 128 | 27.1 | 26.4 | 33.3 | **57.2** | 76.2 |
| 256 | 28.7 | 28.7 | **36.5** | 56.5 | **77.5** |
| 512 | **28.8** | **29.0** | 35.6 | 55.4 | 77.4 |

(c) **Number of positive candidate pair**

| | NCE | MIL-NCE | | | | |
|---|---|---|---|---|---|---|
| $\|\mathcal{P}\| \rightarrow$ | 1 | 3 | 5 | 9 | 17 | 33 |
| YR10 | 29.1 | 33.6 | **35.0** | 33.1 | 32.4 | 28.3 |
| MR10 | 27.0 | 30.2 | **31.8** | 30.5 | 29.2 | 30.4 |
| CTR | 35.6 | **37.3** | 34.2 | 31.8 | 25.0 | 25.0 |
| HMDB | 55.4 | **57.8** | 56.7 | 55.7 | 54.8 | 51.4 |
| UCF | 77.5 | 79.7 | **80.4** | 79.5 | 78.5 | 77.9 |

(d) **MIL strategy**

| Method | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| Cat+NCE | 31.9 | 30.8 | **35.2** | 56.3 | 78.9 |
| Max+NCE | 32.3 | 31.3 | 32.2 | 55.3 | 79.2 |
| Attn+NCE | 32.4 | 30.2 | 33.4 | 55.2 | 78.4 |
| MIL-NCE | **35.0** | **31.8** | 34.2 | **56.7** | **80.4** |

(e) **Symmetric vs asymmetric negatives**

| Negatives | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| $(x\|y)$ | 34.4 | 29.0 | 33.9 | 55.1 | 78.1 |
| $(y\|x)$ | 19.3 | 19.4 | 28.2 | **57.1** | 79.2 |
| $(x, y)$ | **35.0** | **31.8** | **34.2** | 56.7 | **80.4** |

(f) **Language models**

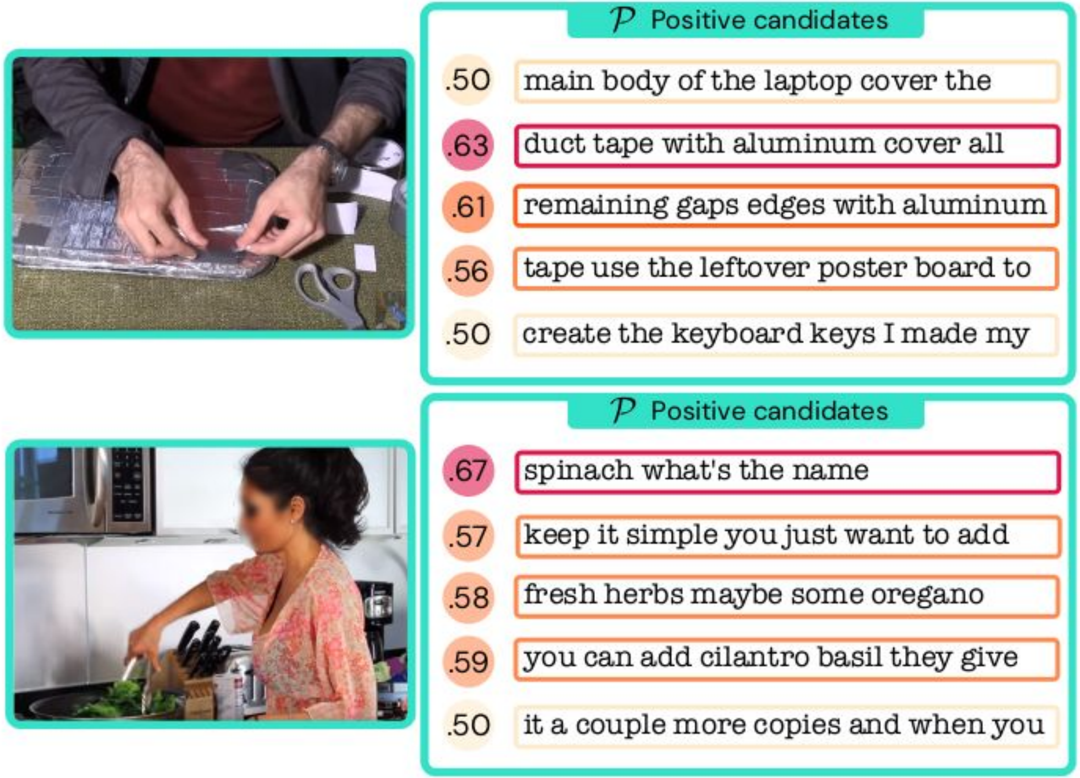
| Text model | YR10 | MR10 | CTR | HMDB | UCF |
|---|---|---|---|---|---|
| LSTM | 16.6 | 15.6 | 23.8 | 53.1 | 80.1 |
| GRU | 16.8 | 16.9 | 22.2 | 54.7 | **82.8** |
| Transformer | 26.7 | 26.5 | 32.7 | 53.4 | 78.4 |
| NetVLAD | 33.4 | 29.2 | **35.5** | 51.8 | 79.3 |
| Ours | **35.0** | **31.8** | 34.2 | **56.7** | 80.4 |

| Input word vectors | YR10 | MR10 |
|---|---|---|
| BERT wo. stop words | 19.0 | 19.7 |
| BERT w. stop words | 17.6 | 23.9 |

# Ablation Studies

- Importance of Multiple Instance Learning: trade-off between likelihood to align and noise
- Many symmetric negative candidates
- Simple language models

# Ablation Studies



# Comparison to SOTA

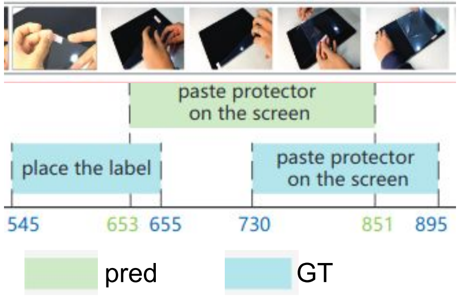
- **Video-only** representation:

  Action recognition


| Method | Dataset | MM | Model | Frozen | HMDB | UCF |
|---|---|---|---|---|---|---|
| OPN [45] | UCF | ✗ | VGG | ✗ | 23.8 | 59.6 |
| Shuffle & Learn [53]* | K600 | ✗ | S3D | ✗ | 35.8 | 68.7 |
| Wang *et al.* [77] | K400 | Flow | C3D | ✗ | 33.4 | 61.2 |
| CMC [73] | UCF | Flow | CaffeNet | ✗ | 26.7 | 59.1 |
| Geometry [25] | FC | Flow | FlowNet | ✗ | 23.3 | 55.1 |
| Fernando *et al.* [24] | UCF | ✗ | AlexNet | ✗ | 32.5 | 60.3 |
| ClipOrder [85] | UCF | ✗ | R(2+1)D | ✗ | 30.9 | 72.4 |
| 3DRotNet [37]* | K600 | ✗ | S3D | ✗ | 40.0 | 75.3 |
| DPC [30] | K400 | ✗ | 3D-R34 | ✗ | 35.7 | 75.7 |
| CBT [70] | K600 | ✗ | S3D | ✓ | 29.5 | 54.0 |
| CBT [70] | K600 | ✗ | S3D | ✗ | 44.6 | 79.5 |
| AVTS [42] | K600 | Audio | I3D | ✗ | 53.0 | 83.7 |
| AVTS [42] | Audioset | Audio | MC3 | ✗ | **61.6** | 89.0 |
| Ours | HTM | Text | I3D | ✓ | **54.8** | **83.4** |
| | | | | ✗ | 59.2 | **89.1** |
| | | | S3D | ✓ | **53.1** | **82.7** |
| | | | | ✗ | 61.0 | **91.3** |
| Fully-supervised SOTA [84] | | | S3D | ✗ | 75.9 | 96.8 |

# Comparison to SOTA

- **Video-only** representation:




Action Segmentation

| Method | Net | Pretraining Dataset | Pretraining Labels | FA |
|---|---|---|---|---|
| | R50 | ImNet | ✓ | 52.0 |
| Ours | I3D | K400 | ✓ | 52.9 |
| | I3D | K700 | ✓ | 54.2 |
| CBT [70] | S3D | K600+HTM | ✓ | 53.9 |
| Ours | I3D | HTM | ✗ | **59.4** |
| Ours | S3D | HTM | ✗ | **61.0** |

(a) **COIN**

# Comparison to SOTA


- **Joint text-video** representation:

Video-to-text retrieval for Action Step Localization

without fine-tuning!

| Method | Labels used | CTR |
|---|---|---|
| Alayrac *et al.* [2] | ImNet+K400 | 13.3 |
| CrossTask [90] | ImNet+K400 | 22.4 |
| CrossTask [90] | ImNet+K400+CT | 31.6 |
| Miech *et al.* [51] | ImNet+K400 | 33.6 |
| Ours (I3D) | **None** | **36.4** |
| Ours (S3D) | **None** | **40.5** |

(d) **CrossTask** (CT)

# Comparison to SOTA

- **Joint text-video** representation: Text-to-video retrieval

**Input:** cut tomato

**Output:**

| Method | Labeled dataset used | R@1↑ | R@5↑ | R@10↑ | MedR↓ |
|---|---|---|---|---|---|
| Random | None | 0.03 | 0.15 | 0.3 | 1675 |
| HGLMM FV CCA [41] | ImNet + K400 + YouCook2 | 4.6 | 14.3 | 21.6 | 75 |
| Miech *et al.* [51] | ImNet + K400 | 6.1 | 17.3 | 24.8 | 46 |
| Miech *et al.* [51] | ImNet + K400 + YouCook2 | 8.2 | 24.5 | 35.3 | 24 |
| Ours (I3D) | **None** | **11.4** | **30.6** | **42.0** | **16** |
| Ours (S3D) | **None** | **15.1** | **38.0** | **51.2** | **10** |

(a) **YouCook2**

without fine-tuning!

| Method | Labeled dataset used | R@1↑ | R@5↑ | R@10↑ | MedR↓ |
|---|---|---|---|---|---|
| Random | None | 0.01 | 0.05 | 0.1 | 500 |
| Miech *et al.* [51] | ImNet + K400 | 7.5 | 21.2 | 29.6 | 38 |
| Ours (I3D) | **None** | **9.4** | **22.2** | **30.0** | **35** |
| Ours (S3D) | **None** | **9.9** | **24.0** | **32.4** | **29.5** |

(b) **MSRVTT**

# Experiments

https://www.di.ens.fr/willow/research/mil-nce/

## Zero-shot Text-to-Video retrieval on YouCook2

crack eggs

**Ranked 1: IdEZ7LvLZPE -- Score: 9.83**


**Ranked 2: Vuy2nrJz0Zw -- Score: 9.73**

# Conclusion

“ Use the novel **MIL-NCE objective** to learn **video representations without annotations,** by dealing with **misalignments** from uncurated instructional videos ”

MIL
Multiple Instance Learning

**MIL-NCE**

NCE
Noise Contrastive Estimation

# Conclusion

“ Use the novel **MIL-NCE objective** to learn **video representations without annotations,** by dealing with **misalignments** in uncurated instructional videos ”

**Thank you!**

MIL
Multiple Instance Learning

**MIL-NCE**

NCE
Noise Contrastive Estimation

UNI FREIBURG

# Discussion impulses

- How does their self-supervised fine-tuned version compare to supervised SOTA?
- Further explanation about differences among datasets in ablation studies
- In which other applications could it be interesting to use MIL-NCE loss?