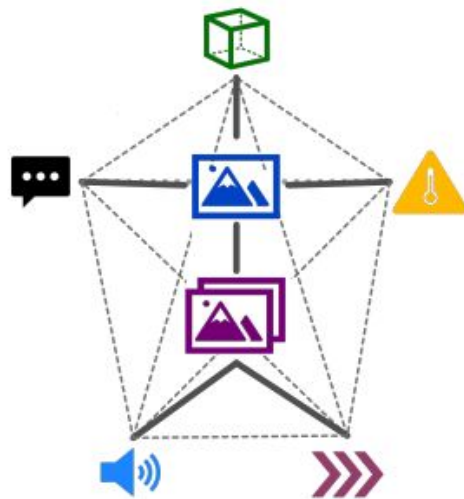


ImageBind: One Embedding Space To Bind Them All



Подготовил:
Федоров Никита, БПМИ202

nmfedorov@edu.hse.ru

Основная идея статьи

Сделать общее пространство эмбеддингов для картинок, видео, аудио, текста и других модальностей



План

1. Краткое описание всех модальностей
2. Особенность именно этой статьи
3. Детали процесса обучения
4. Результаты/эксперименты
 - 4.1. Emergent zero-shot classification
 - 4.2. Таблички и графики
 - 4.3. Прочие плюшки, которые дает общее признаковое пространство
 - 4.3.1. Cross-modal retrieval
 - 4.3.2. Multimodal embedding space arithmetic
 - 4.3.3. Upgrading text-based detectors to audio-based
 - 4.3.4. Upgrading text-based diffusion models to audio-based
5. Сильные и слабые стороны статьи

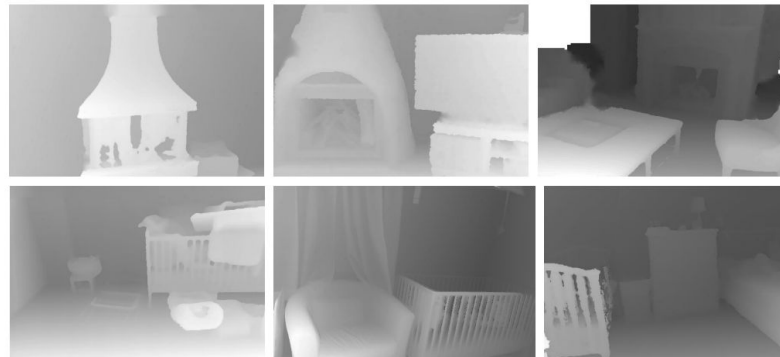
Модальности

IMU

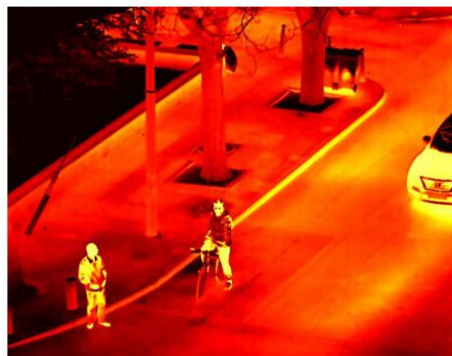
Egocentric Videos



Depth

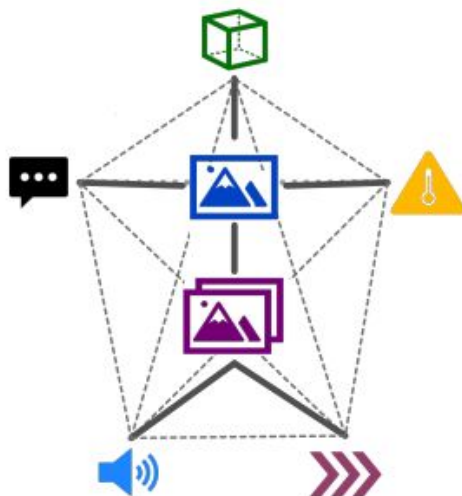


Thermal Data



Особенность этой статьи

Обучение велось только на сетах, содержащих пары
(картинка, другая модальность)



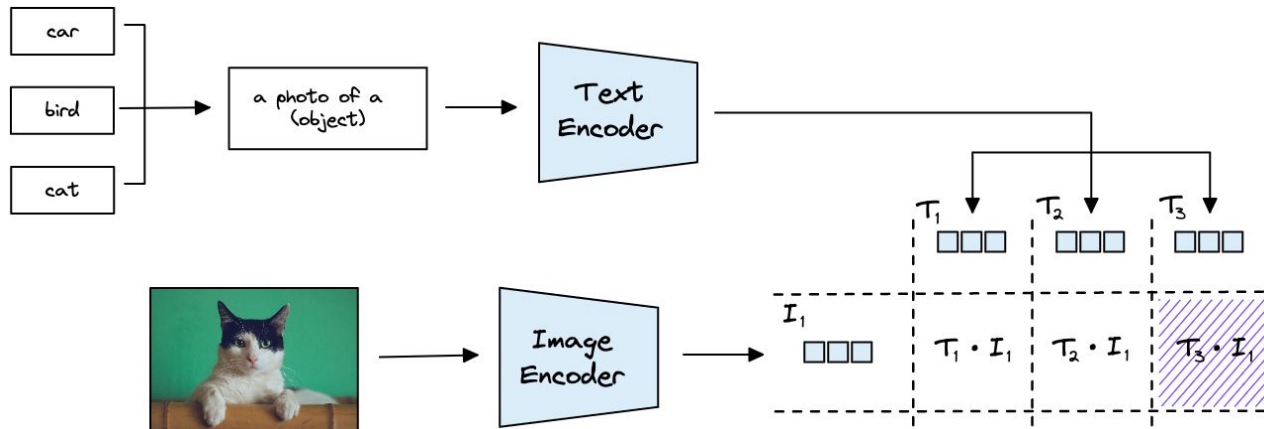
Процесс обучения

- для каждой модальности заводим энкодер (в данной статье все энкодеры трансформерные)
- энкодеры для **текста и картинок** берем из **CLIP** и **замораживаем** их веса
- для **текста и видео** использовался **один энкодер**
- функция потерь InfoNCE:







$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}$$

$$L_{\mathcal{I},\mathcal{M}} + L_{\mathcal{M},\mathcal{I}}$$

Emergent zero-shot classification



Результаты в zero-shot

| |  | |  | |  | |  | |  | |  |
|---------------|---|-----------|---|-----------|---|-----------|---|-----------|---|-------|---|
| | IN1K | P365 | K400 | MSR-VTT | NYU-D | SUN-D | AS-A | VGGS | ESC | LLVIP | Ego4D |
| Random | 0.1 | 0.27 | 0.25 | 0.1 | 10.0 | 5.26 | 0.62 | 0.32 | 2.75 | 50.0 | 0.9 |
| IMAGEBIND | 77.7 | 45.4 | 50.0 | 36.1 | 54.0 | 35.1 | 17.6 | 27.8 | 66.9 | 63.4 | 25.0 |
| Text Paired | - | - | - | - | 41.9* | 25.4* | 28.4 [†] [27] | - | 68.6 [†] [27] | - | - |
| Absolute SOTA | 91.0 [82] | 60.7 [67] | 89.9 [80] | 57.7 [79] | 76.7 [21] | 64.9 [21] | 49.6 [39] | 52.5 [36] | 97.0 [9] | - | - |

| | Emergent | Clotho | | AudioCaps | | ESC |
|--|----------|------------|-------------|------------|-------------|-------------|
| | | R@1 | R@10 | R@1 | R@10 | Top-1 |
| <i>Uses audio and text supervision</i> | | | | | | |
| AudioCLIP [27] | ✗ | - | - | - | - | 68.6 |
| <i>Uses audio and text loss</i> | | | | | | |
| AVFIC [51] | ✗ | 3.0 | 17.5 | 8.7 | 37.7 | - |
| <i>No audio and text supervision</i> | | | | | | |
| IMAGEBIND | ✓ | 6.0 | 28.4 | 9.3 | 42.3 | 66.9 |
| <i>Supervised</i> | | | | | | |
| AVFIC finetuned [51] | ✗ | 8.4 | 38.6 | - | - | - |
| ARNLQ [53] | ✗ | 12.6 | 45.4 | 24.3 | 72.1 | - |

Cross-modal retrieval

1) Cross-Modal Retrieval

Audio



Crackle of a Fire

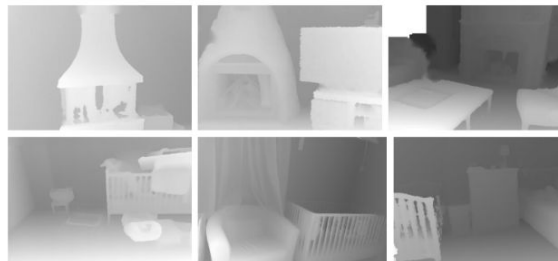


Baby Cooing

Images & Videos



Depth



Text

“A fire crackles while a pan of food is frying on the fire.”

“Fire is crackling then wind starts blowing.”

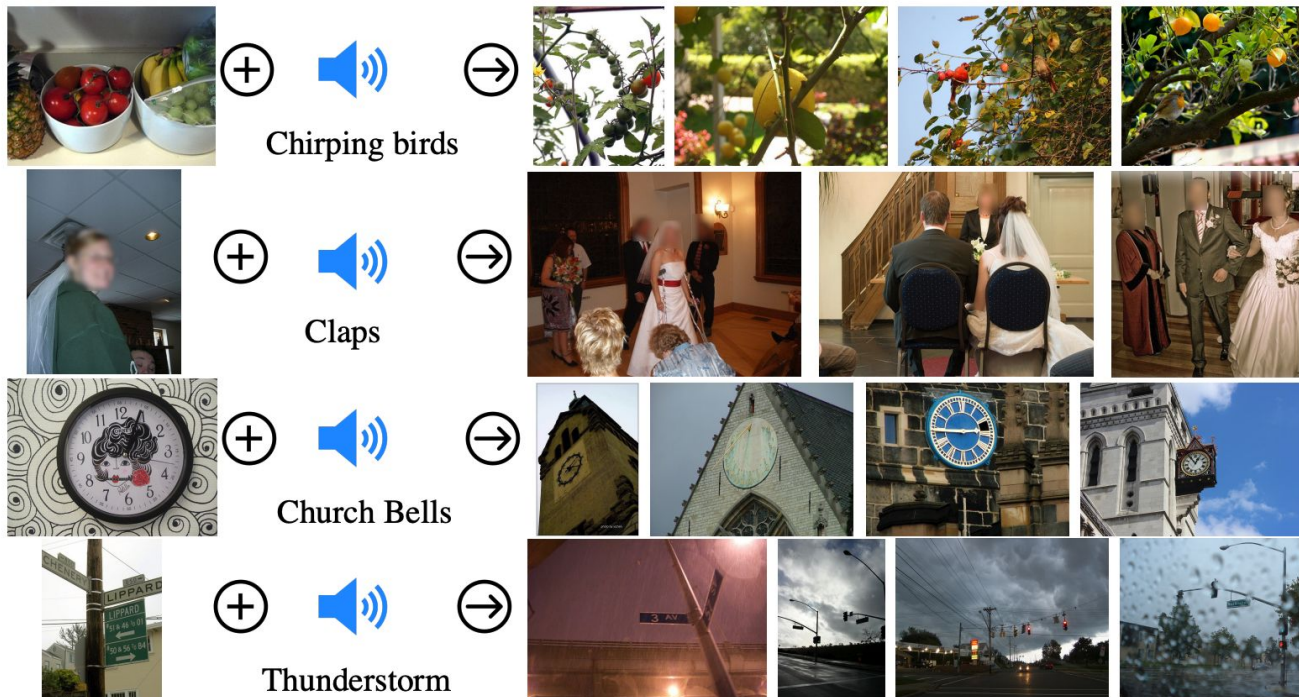
“Firewood crackles then music...”

“A baby is crying while a toddler is laughing.”

“A baby is laughing while an adult is laughing.”

“A baby laughs and something...”

Multimodal embedding space arithmetic



Upgrading text-based detectors and diffusion models to audio-based

3) Audio to Image Generation



Dog



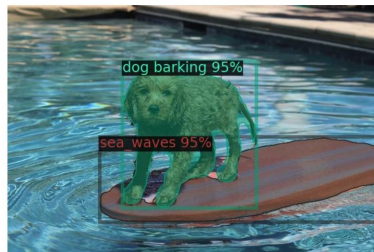
Engine



Fire

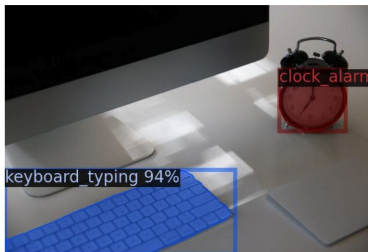


Rain



Dog barking

Sea waves



Keyboard typing

Clock alarm

Сильные стороны статьи

- можно натренировать по одному энкодеру для каждой модальности и потом решать очень много кросс модальных задач (экономия ресурсов)
- используют только naturally-paired данные, можно натренировать при отсутствии датасетов для каждой пары модальностей

Слабые стороны статьи

- идея довольно очевидная, никаких новых методов не предложено (сори за душноту)
- (чисто мой наброс) возможно пространства данных в разных модальностях могут отличаться структурно (например, некоторые предметы на картинках могут вовсе не ассоциироваться с каким-то звуком). Поэтому выравнивание эмбеддингов всех модальностей на картинки (как и в целом идея общего признакового пространства) может приводить к просадке качества по сравнению с узкоспециализированными алгоритмами