Ideas

Universal time-series representations

Time2Vec

(Rejected from ICLR'20)

https://arxiv.org/pdf/1907.05321.pdf

$$\mathbf{t2v}(\tau)[i] = \begin{cases} \omega_i \tau + \varphi_i, & \text{if } i = 0. \\ \mathcal{F}(\omega_i \tau + \varphi_i), & \text{if } 1 \leq i \leq k. \end{cases}$$

- General-purpose model-agnostic representation for time that can be potentially used in any architecture.
- Domain-agnostic.
- In contrast to static representations such as Fourier transforms of time-series, Time2Vec allows dynamic representation with learnable params.
- F here is either the sine or cosine function, which is intended to capture the periodicity of time

https://openreview.net/forum?id=rklklCVYvB

A word of encouragement

Eelco Hoogendoorn

12 Apr 2020 ICLR 2020 Conference Paper915 Public Comment Readers: Everyone

Comment: Id like to add a word of encouragement to the authors.

Even if we should assume there isnt a single 'novel' idea in this paper (and I am not saying there isnt), given how highly under-studied the subject of this paper is, any general reflection on the question of how to best encode time/position in a modern machine learning context, is a much more valuable contribution, than the vast majority of published pages, of authors contorting themselves to prove the 'novelty' of their architectural gizmos.

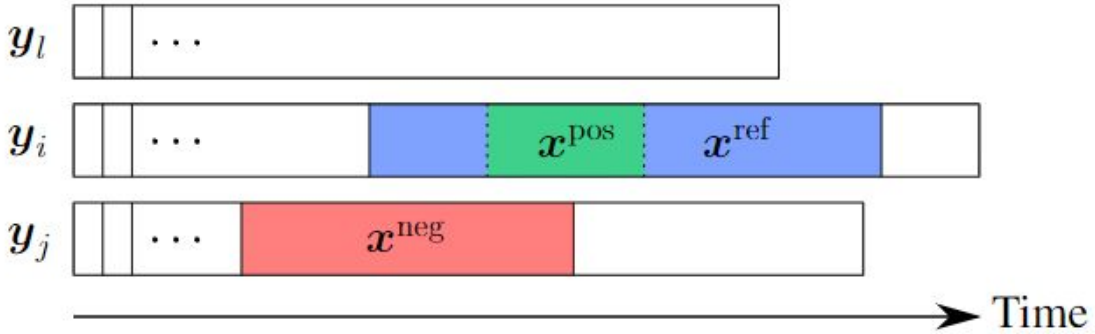
Yes, I can think of a million more experiments and theoretical questions to be pursued along the lines of this paper as well. But until people can point at other papers that already covered the ground this paper does, that seems like an exceedingly poor reason for rejection, given the complete lack of attention given in the literature, to this subject of tremendous practical importance.

Yes, maybe this is just an investigation of how a slight variant of positional encoding popularised by attention networks performs in other contexts. But if there was a single paper id read in the next year, that would probably be it.

Unsupervised Scalable Representation Learning for Multivariate Time Series

(NeurIPS ’19)

- Pre-training approach to utilize unlabeled time-series data
- Trained with a Triplet-loss
- A bit architecture-specific with dilated convolutions
- Their embeddings show great performance in complete unsupervised, as well as semi-supervised settings.
- Doesn’t work on univariate data.

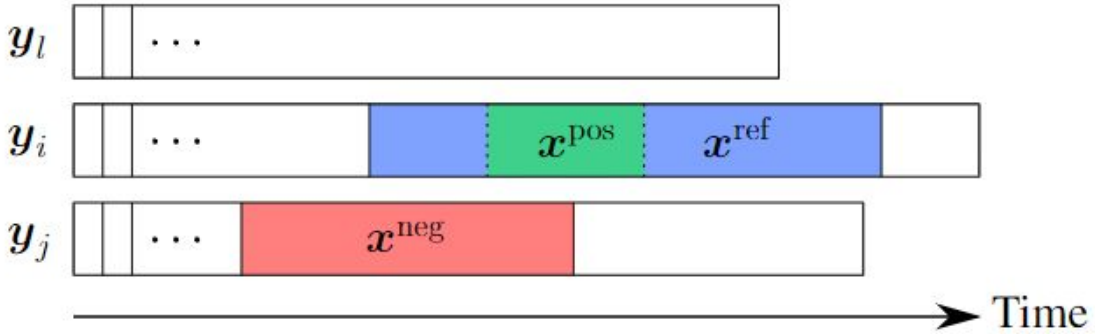


$$-\log\left(\sigma\left(f\left(x^{\text{ref}},\theta\right)^{\top} f\left(x^{\text{pos}},\theta\right)\right)\right)-\sum_{k=1}^{K}\log\left(\sigma\left(-f\left(x^{\text{ref}},\theta\right)^{\top} f\left(x_k^{\text{neg}},\theta\right)\right)\right)$$

Unsupervised Scalable Representation Learning for Multivariate Time Series

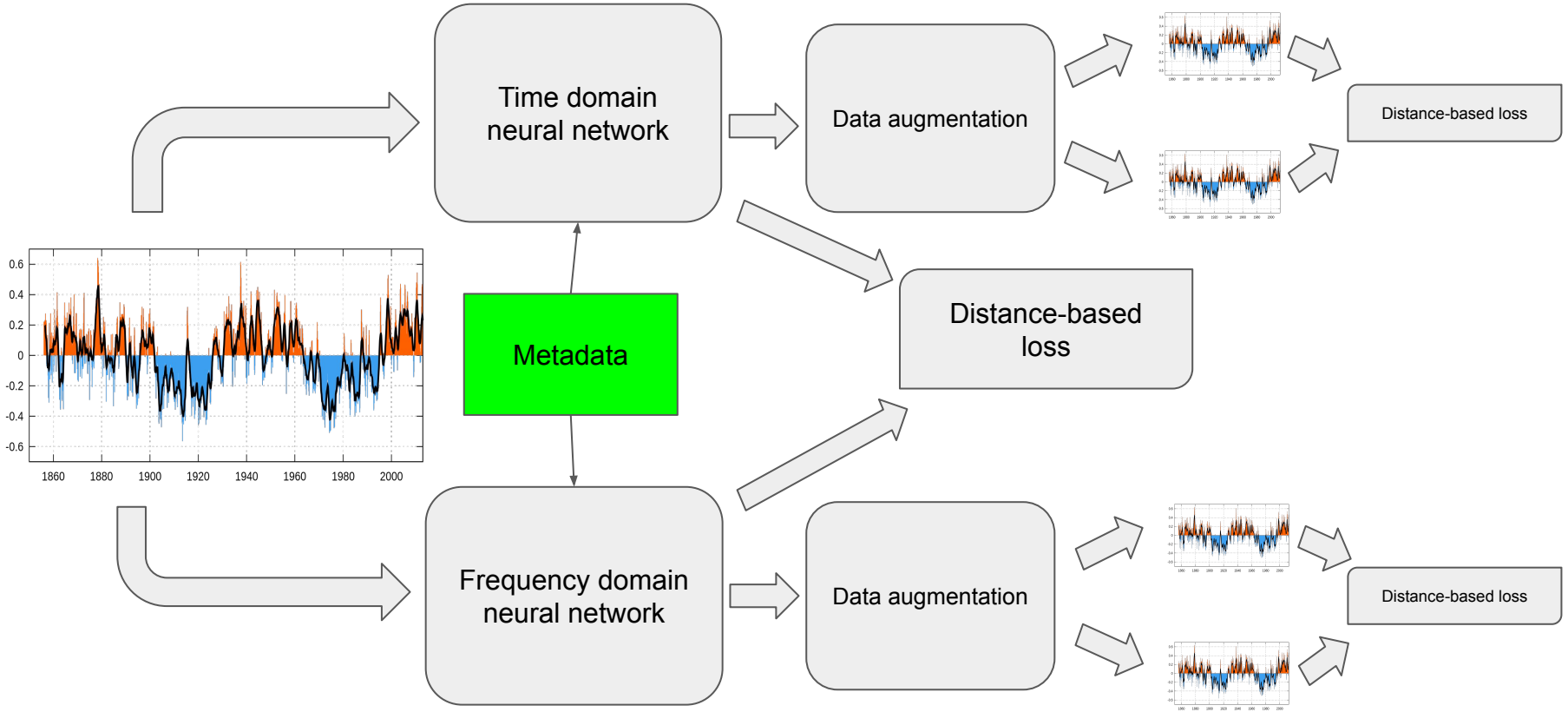
(NeurIPS ’19)

- Pre-training approach to utilize unlabeled time-series data
- Trained with a Triplet-loss
- A bit architecture-specific with dilated convolutions
- Their embeddings show great performance in complete unsupervised, as well as semi-supervised settings.
- Doesn’t work on univariate data.



$$-\log\left(\sigma\left(f\left(x^{\text{ref}},\theta\right)^{\top} f\left(x^{\text{pos}},\theta\right)\right)\right)-\sum_{k=1}^{K}\log\left(\sigma\left(-f\left(x^{\text{ref}},\theta\right)^{\top} f\left(x_k^{\text{neg}},\theta\right)\right)\right)$$

Idea 1: Can we build a universal pre-trained model that can be utilized for any-domain time-series data? (e.g. BERT in NLP)

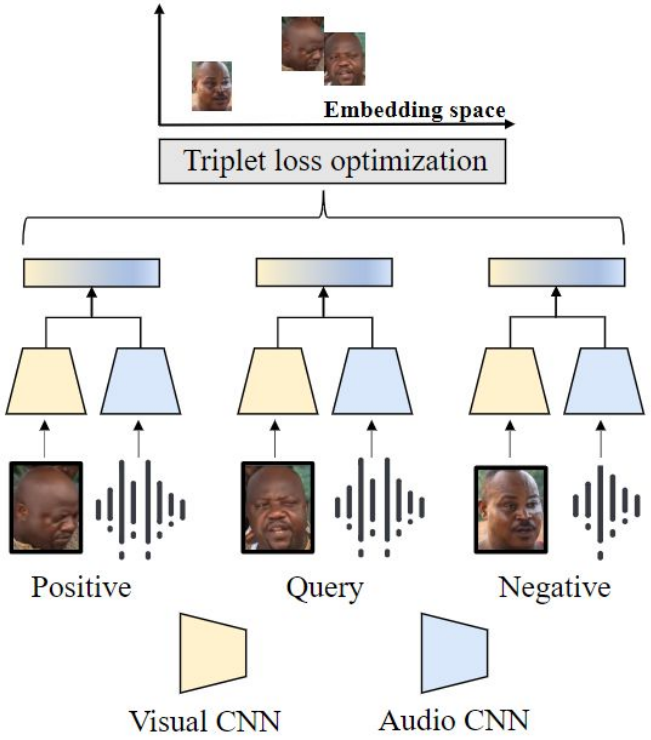


Cross-modal search / retrieval

APES: Audiovisual Person Search in Untrimmed Video

(CVPR MULA Workshop 2021)

- Proposed an audiovisual baseline and benchmark for person retrieval.
- Concluded that modeling audiovisual cues benefits the recognition of people's identities.
- Developed a two-stream model that predicts people's identities using audiovisual cues.
- Can we explore a generalizable cross-modal search model?

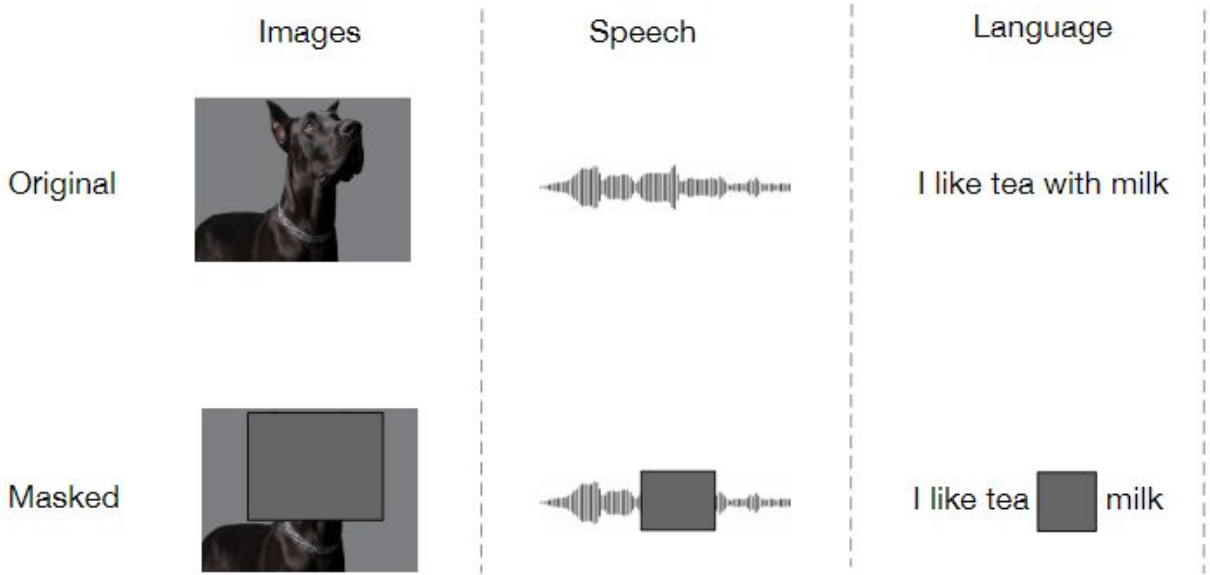


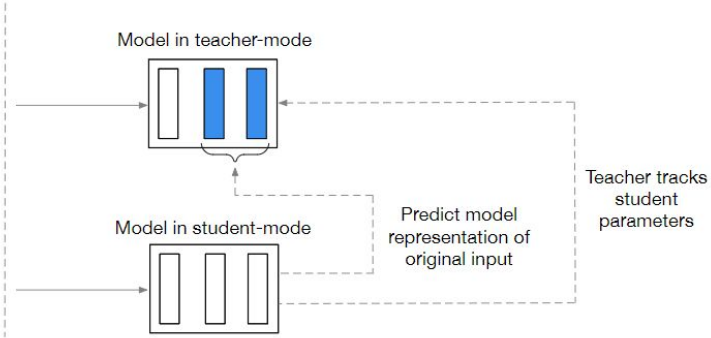
Multimodal foundation models

Data2Vec

(Meta AI research)

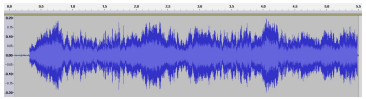
- Data2vec simplifies by training models to predict their own representations of the input data, regardless of the modality.
- The core idea is to predict latent representations of the full input data based on a masked view of the input in a self-distillation setup.
- Standard Transformer architecture.





Idea 2: Can we build a universal data encoder using cross-modal masking based generation?

Mask modalities one at a time during training



Audio encoder

Masked input

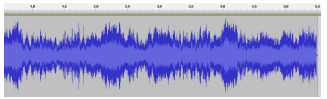
Vision encoder

A cat meowing

Text encoder

Model

Audio decoder



Vision decoder

Predicted missing input

Text decoder

A cat meowing