Adversarial and latent data augmentation for sound event detection

David Perera, Slim Essid, Gaël Richard

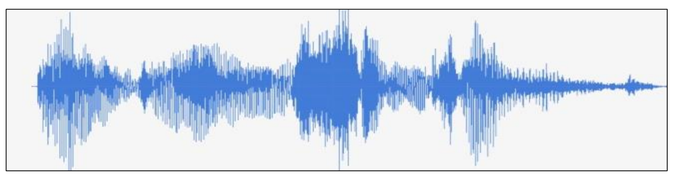
LTCI, Telecom Paris, France

1. Motivation

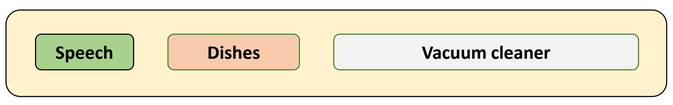
- Data augmentation mitigates the lack of training data
- Invariant based learning enforce regularity and interpretability

2. Task

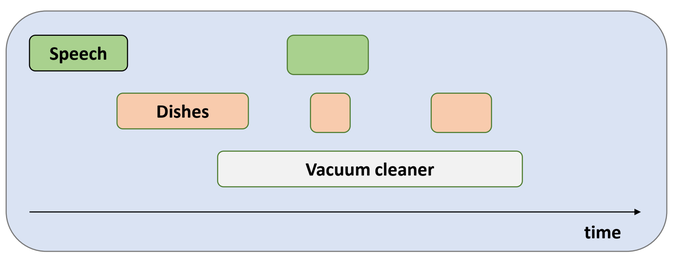
Audio input



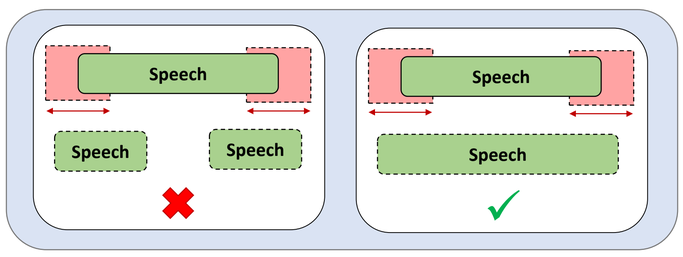
Weak annotations



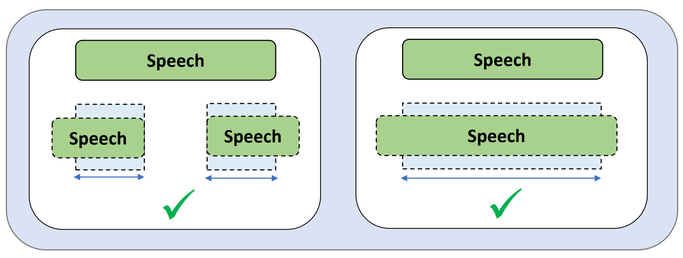
Strong annotations



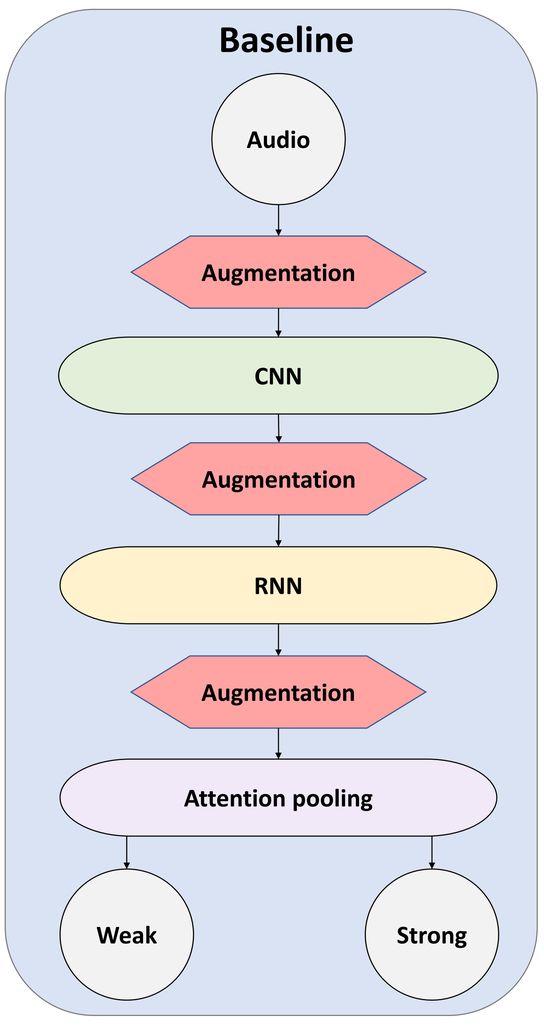
Segment based score

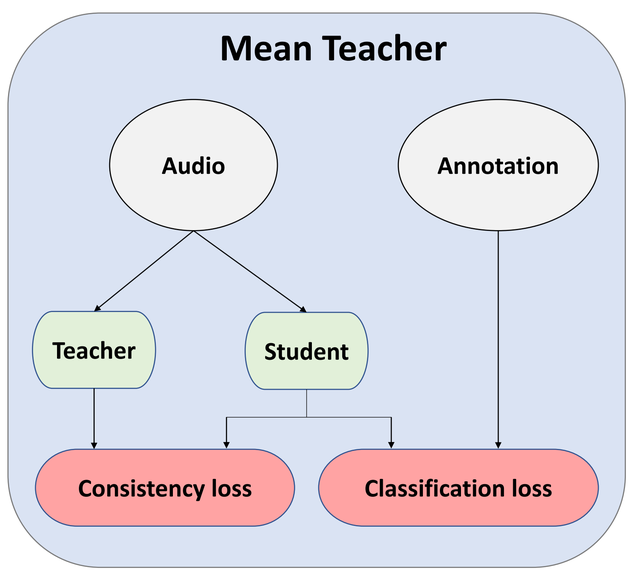


Intersection based score



3. Baseline model





Training objective

$$f_{teacher}^{(n+1)} = \alpha^{(n)} f_{student}^{(n)} + \left(1 - \alpha^{(n)}\right) \mathrm{f}_{\mathrm{teacher}}^{(n)}$$

$$\mathcal{L}_{consistency}(x) = \| f_{student}(x) - f_{teacher}(x + d) \|_2^2$$

$$\mathcal{L}_{classification}(x, y) = \mathcal{L}_{BCE}[\, f(x), y \,]$$

4. Data augmentation

$$\mathcal{L}_{total} = \mathcal{L}_{classification} + \mathcal{L}_{consistency} + \mathcal{L}_{regularization}$$

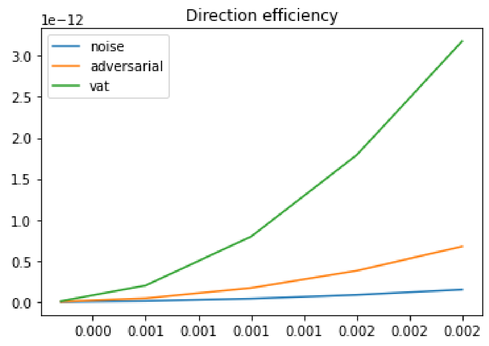
$$\mathcal{L}_{consistency} = \|f(x) - f(x + d)\|_2^2 \qquad \mathcal{L}_{regularization} = \|f\|_2^2$$

$$d_{random} \sim N(0, I)$$

$$d_{adversarial} \sim \nabla_d \|f(x) - f(x + d)\|_2^2$$

$$d_{VAT} \sim \underset{\|d\| \leq \epsilon}{\operatorname{argmax}} \|f(x) - f(x + d)\|_2^2$$

$$d_{mixup} = Mixup(x, x')$$



5. Results

Name	Augmentation	Location		Score	
		Input	Internal	Segment	Intersection
class	None			0.291	0.500
noise	Random	x		**0.397**	0.534
deep noise	Random		x	0.388	**0.565**
Deep adv	Adversarial		x	0.374	0.559
vat	VAT	x		0.358	0.539
Deep vat	VAT		x	0.362	0.542
deep mixup	Mixup		x	0.351	0.507
baseline	MT	x		0.381	0.552
deep vat mt	MT+VAT	x		0.311	0.461

6. Take-away

- We propose a training objective that is more flexible, requires less gpu memory, is easier to interpret, and leads to better performance.
- The performance gain is class invariant and resilient to a diminution of the volume of training data.
- Input and internal noise moves a classifier boundaries away from the training data, improving its performance.