Understanding Contrastive Learning Requires Incorporating Inductive Biases

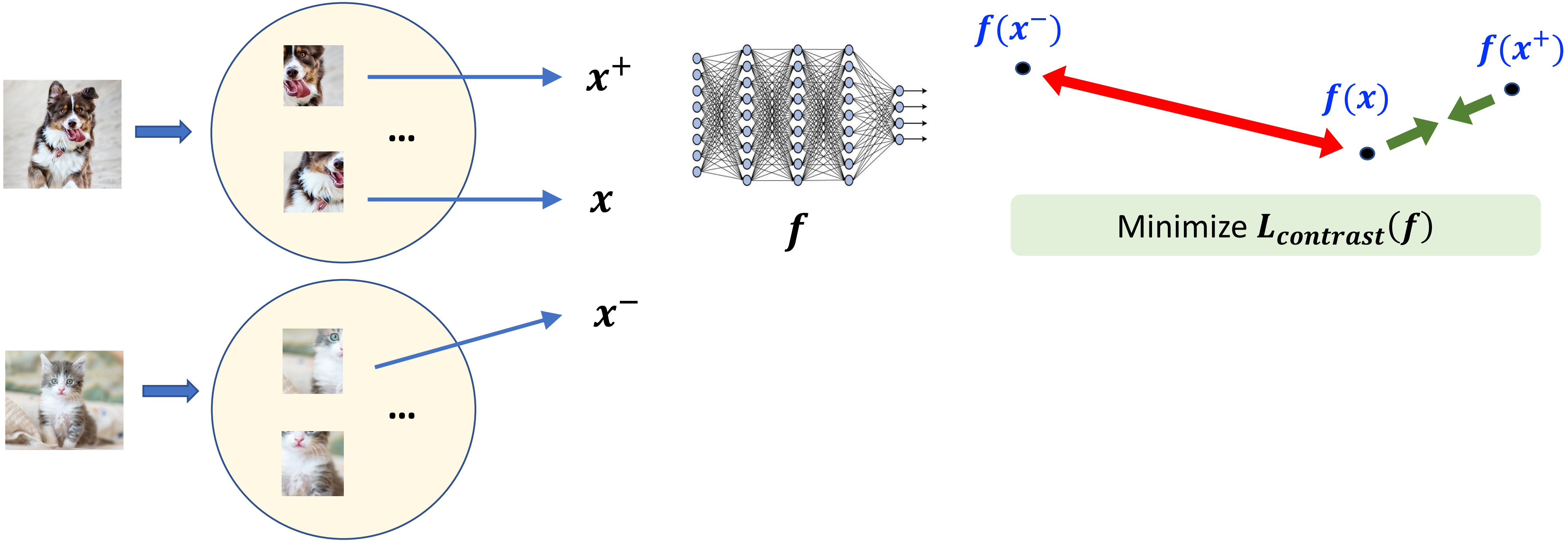
Nikunj Saunshi[1*], Jordan T. Ash[2], Surbhi Goel[2], Dipendra Misra[2], Cyril Zhang[2]
Sanjeev Arora[1], Sham Kakade[2 3], Akshay Krishnamurthy[2]

ICML 2022

[1]Princeton University, [2]Microsoft Research NYC, [3]Harvard University
* nsaunshi@cs.princeton.edu

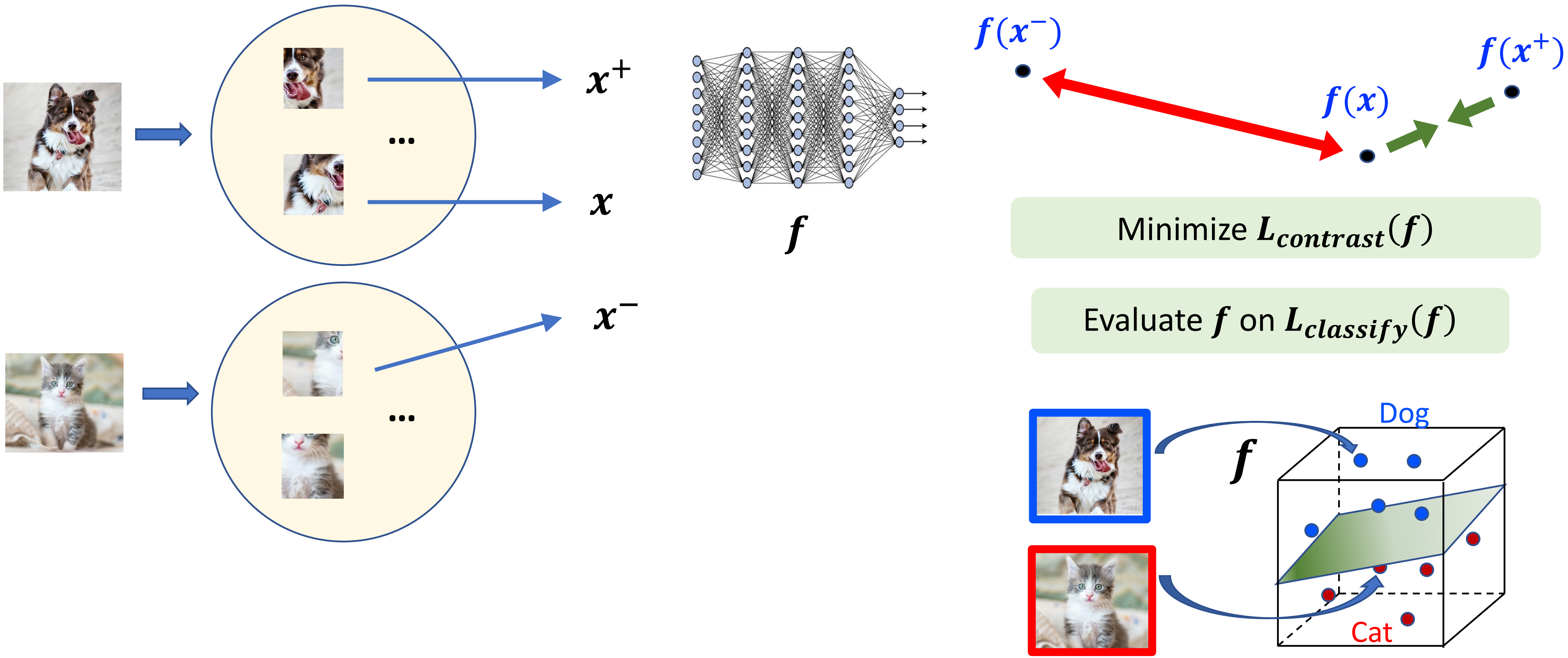
Contrastive learning with augmentations

Representations contrast **similar points** (data augmentations) against random points, e.g. SimCLR [CKNH 20]



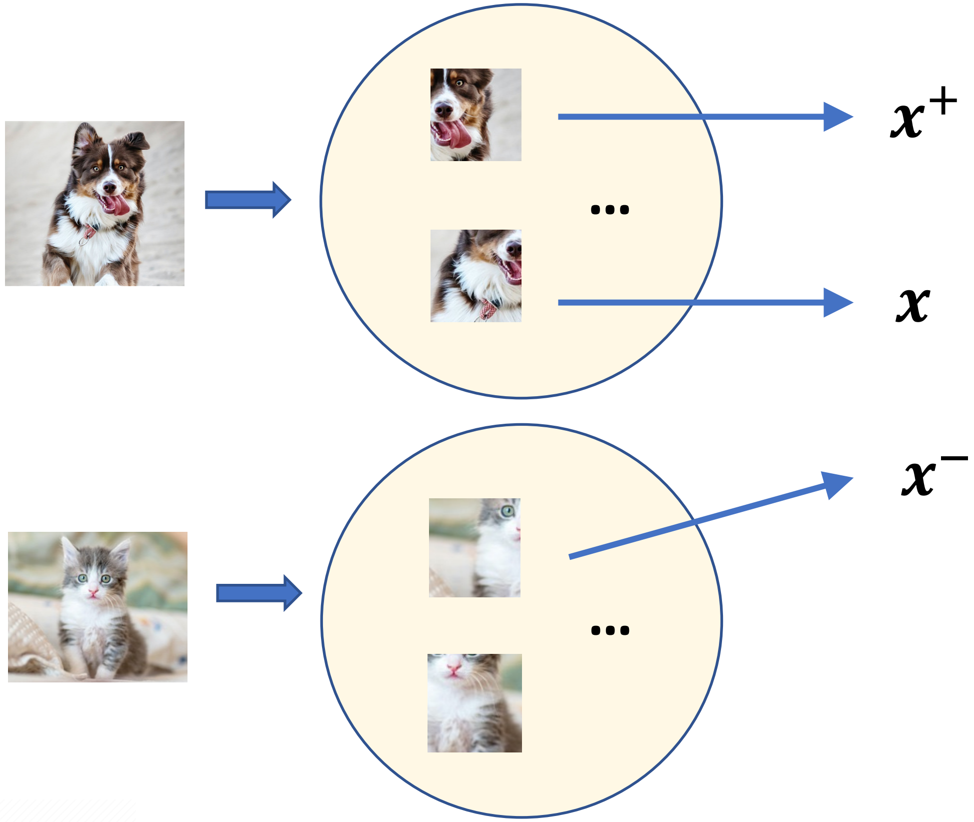
Contrastive learning with augmentations

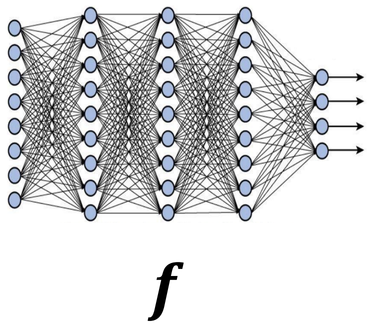
Representations contrast **similar points** (data augmentations) against random points, e.g. SimCLR [CKNH 20]

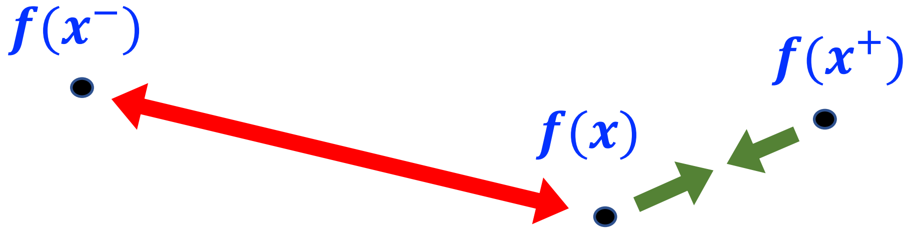


Contrastive learning with augmentations

Representations contrast **similar points** (data augmentations) against random points, e.g. SimCLR [CKNH 20]

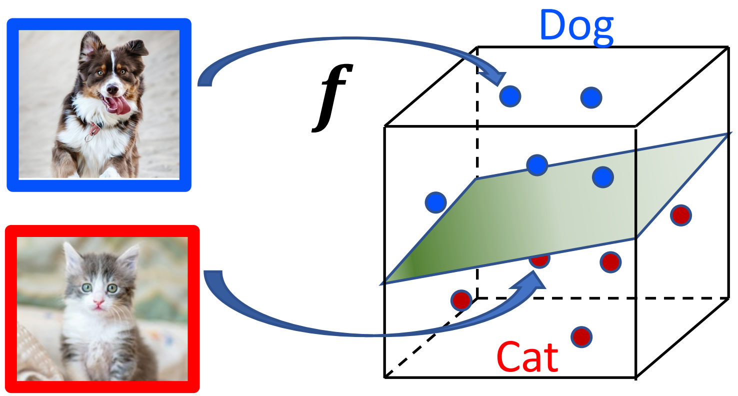


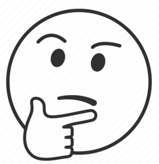




Minimize $L_{contrast}(f)$

Evaluate f on $L_{classify}(f)$





Why should small $L_{contrast}(f)$ ➔ small $L_{classify}(f)$

Theory for contrastive learning

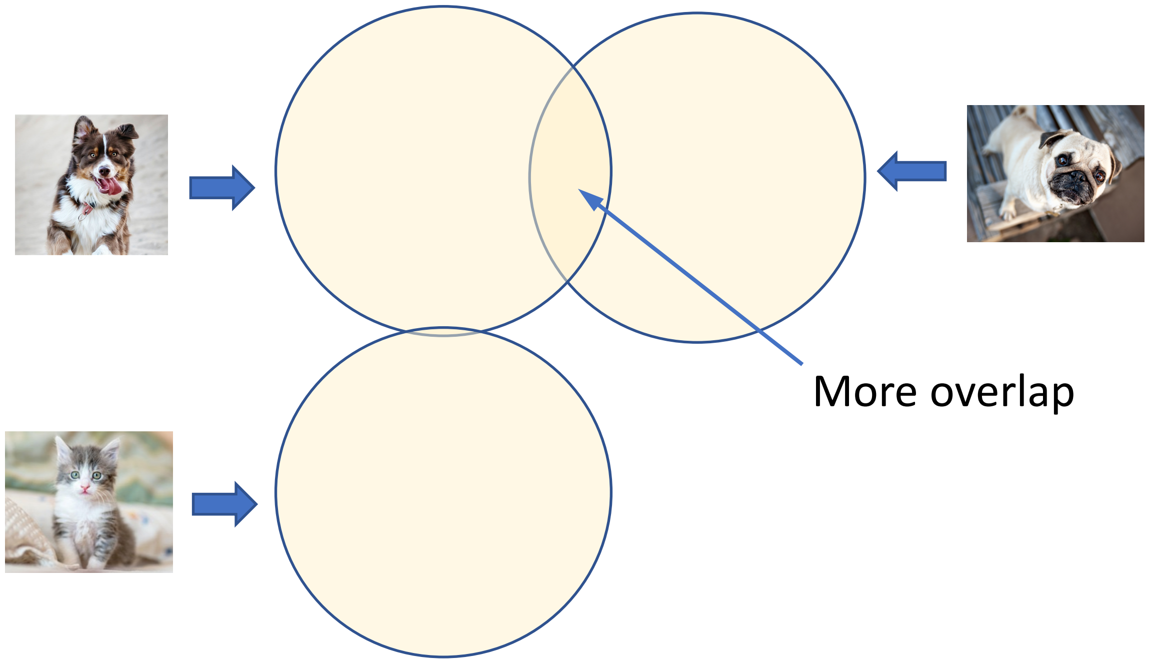
Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations

Theory for contrastive learning

Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations



Spectral CL [HWGM 21]

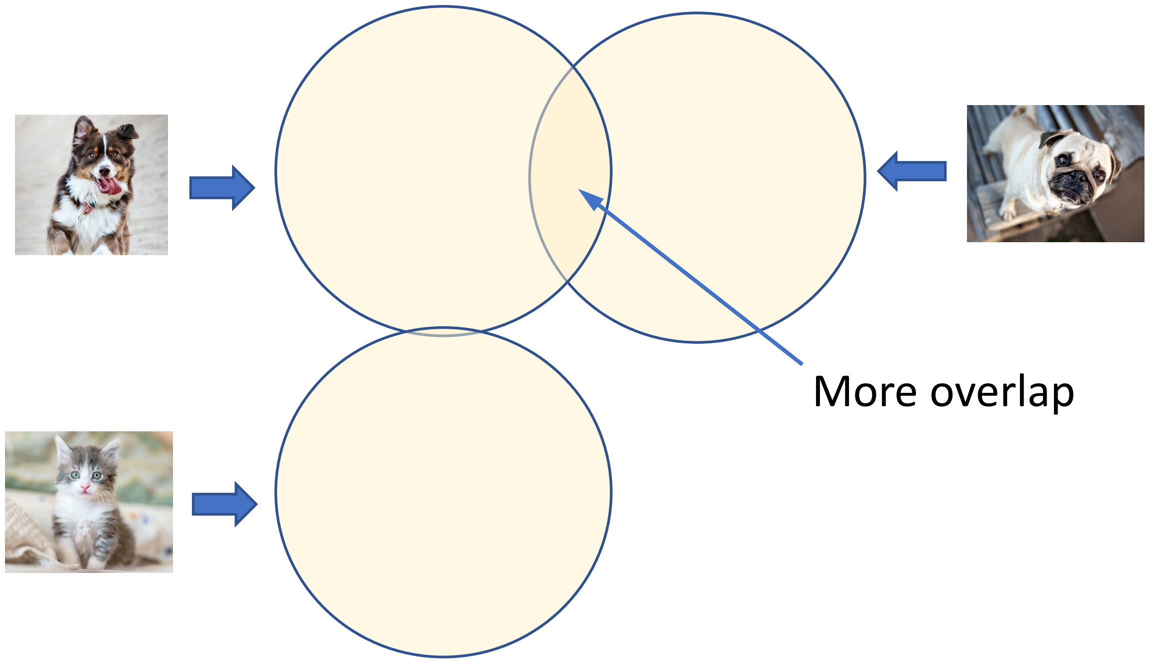
Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\le\; a\, L_{contrast}(f) + b, \;\forall f$

Theory for contrastive learning

Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations



Spectral CL [HWGM 21]

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \leq a\, L_{contrast}(f) + b, \; \forall f$

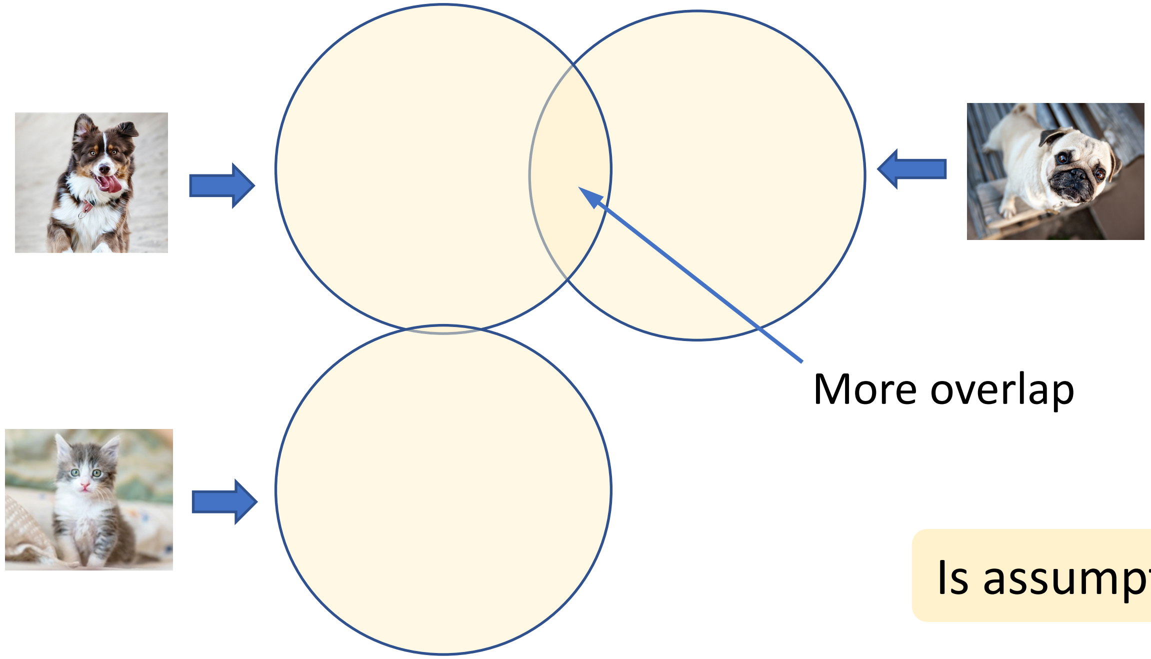
Treats f as "black-box"

Minimize $L_{contrast}$ any way possible

Theory for contrastive learning

Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations



Spectral CL [HWGM 21]

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\; \forall f$

Treats f as "black-box"

Minimize $L_{contrast}$ any way possible

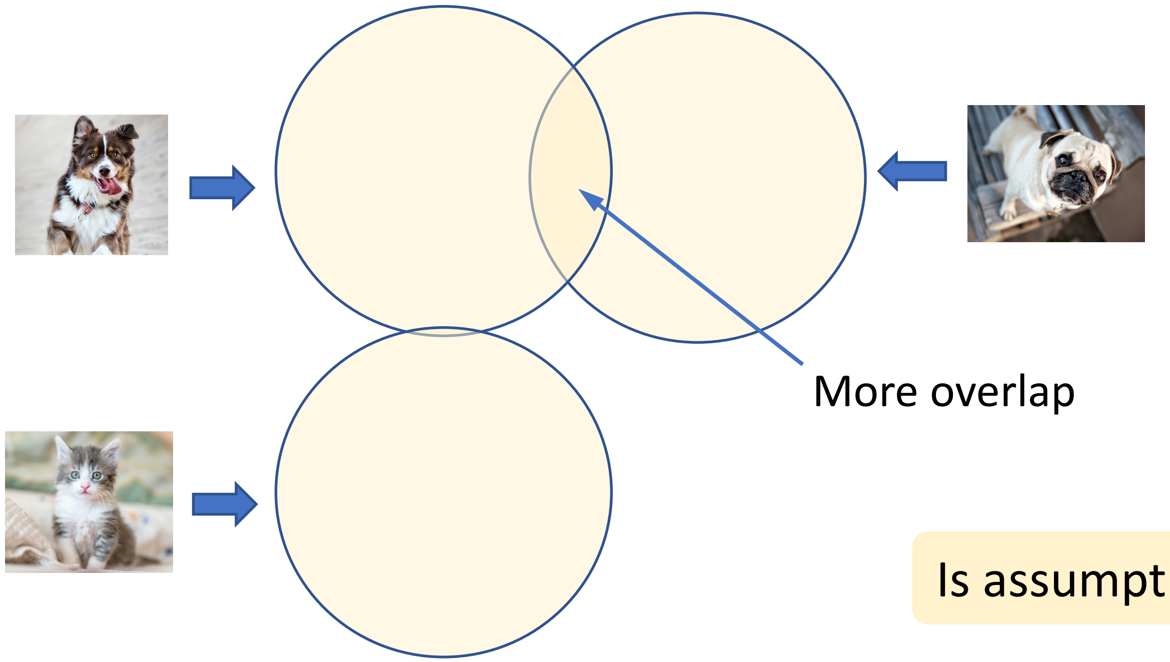
Is assumption satisfied in practice?

Maybe not

Theory for contrastive learning

Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations



Spectral CL [HWGM 21]

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \leq a\, L_{contrast}(f) + b, \ \forall f$

Treats f as "black-box"

Minimize $L_{contrast}$ any way possible

Is assumption satisfied in practice?

Maybe not

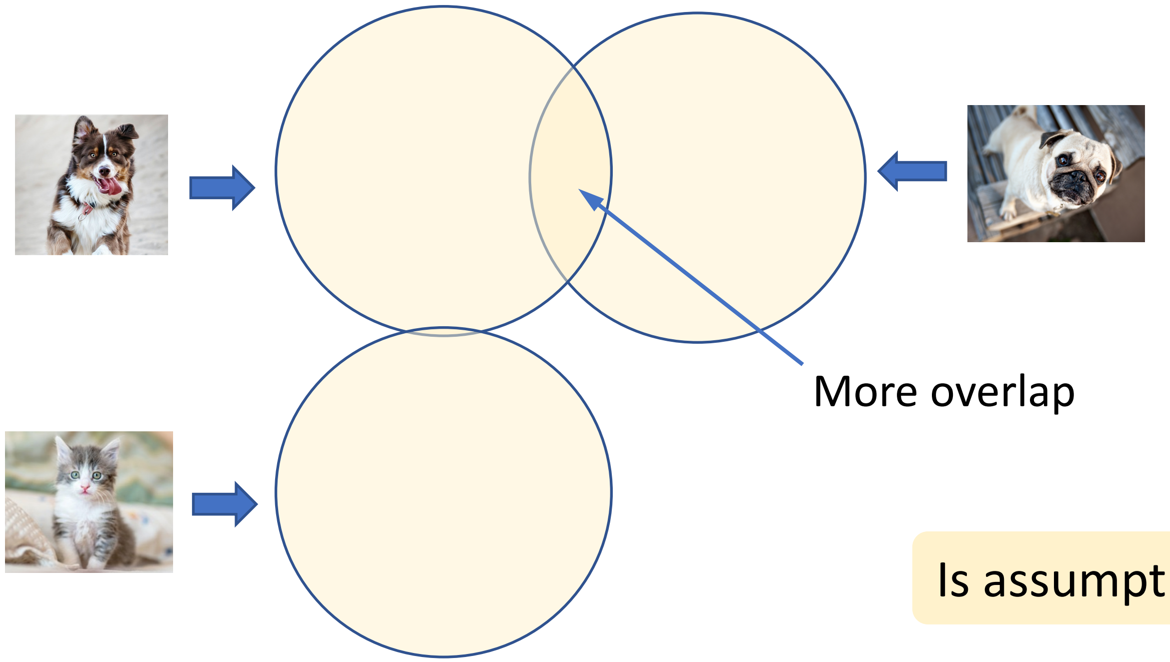
Can contrastive learning work without overlap?

Yes!

Theory for contrastive learning

Connect **similar point** distributions to downstream classes

Conditional independence [AKKPS 19] is unrealistic for augmentations



Spectral CL [HWGM 21]

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\ \forall f$

Treats f as "black-box"

Minimize $L_{contrast}$ any way possible

Is assumption satisfied in practice? **Maybe not**

Can contrastive learning work without overlap? **Yes!**

Can inductive bias ***agnostic*** analysis explain this success? **No!**

Questions

Assumption: Augmentation overlap within class

Guarantees: $\boldsymbol{L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\ \forall f}$

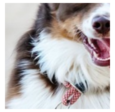
Is assumption satisfied in practice?

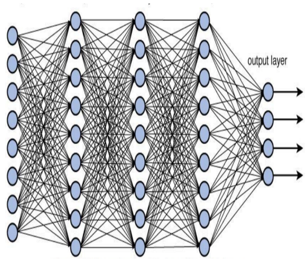
Is there overlap?

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b, \;\forall f$

Is assumption satisfied in practice?





Augmentation → original image classification

99.6% Accuracy on 5000-way classification!

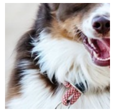
Is there overlap?

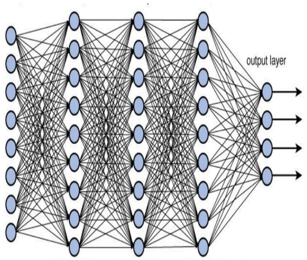
Assumption: Augmentation overlap within class

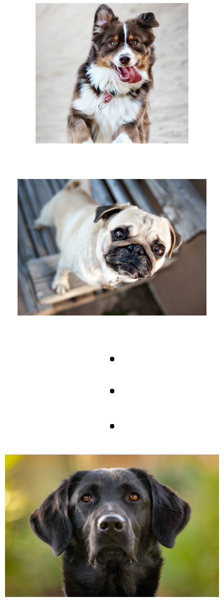
Guarantees: $L_{classify}(f) \;\le\; a\, L_{contrast}(f) + b,\ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)







Augmentation → original image classification

99.6% Accuracy on 5000-way classification!

Contrastive learning without overlap

Assumption: Augmentation overlap within class

Guarantees: $\boldsymbol{L_{classify}(f) \leq a\, L_{contrast}(f) + b,\ \forall f}$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Contrastive learning without overlap

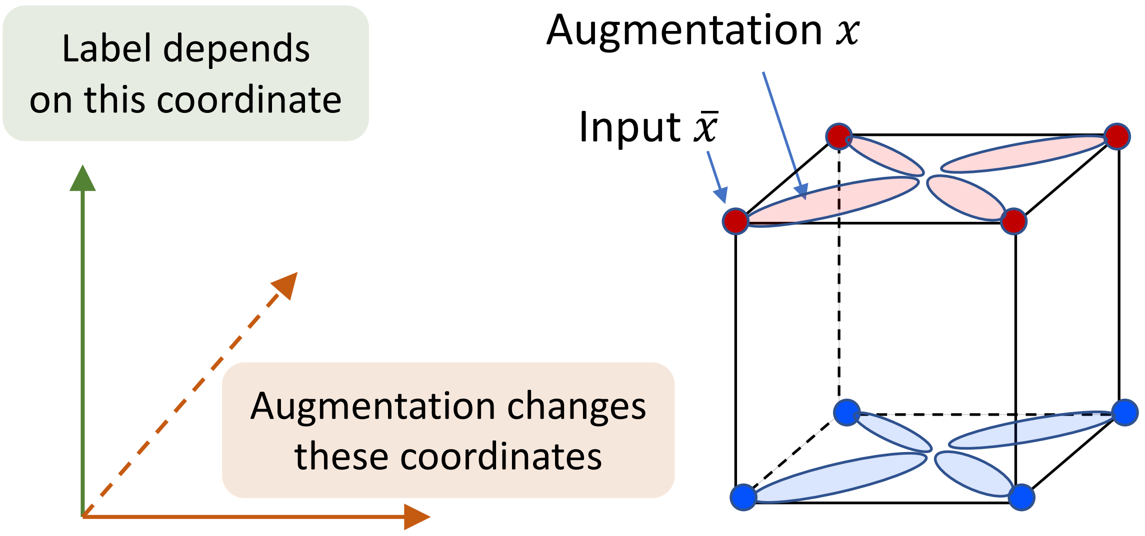
Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?



Contrastive learning without overlap

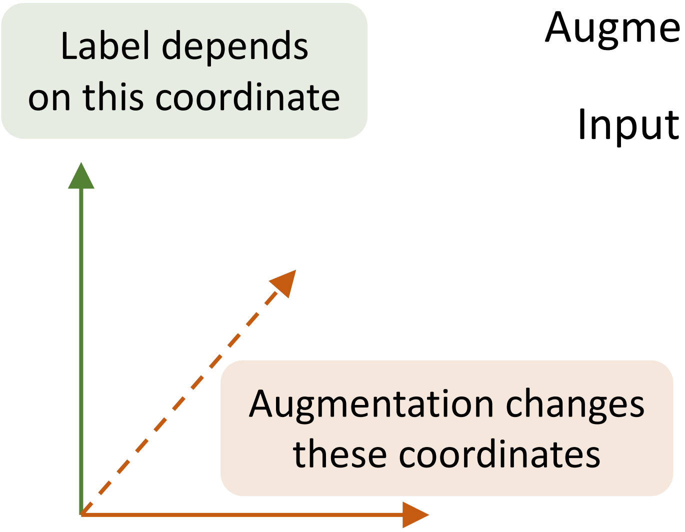
Assumption: Augmentation overlap within class

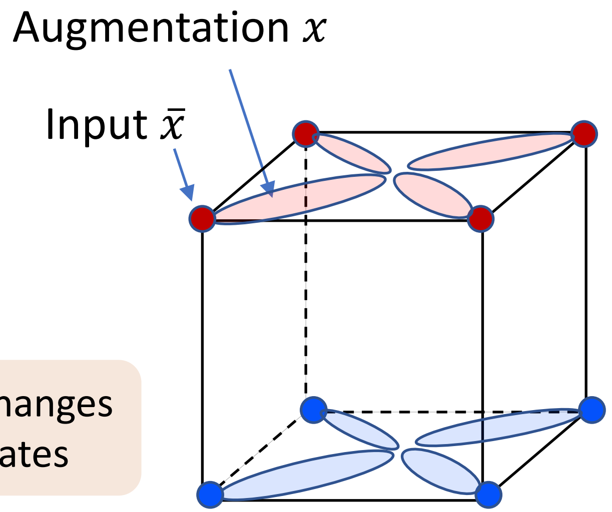
Guarantees: $L_{classify}(f) \leq a\, L_{contrast}(f) + b, \ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?





Representation	$L_{cont}(f)$	Acc (%)
Linear	5.13	99.5
MLP + Adam	5.04	74.1
MLP + Adam + wd		89.5
$\exists\, f$ (spurious)	4.94	50

Contrastive learning without overlap

Assumption: Augmentation overlap within class

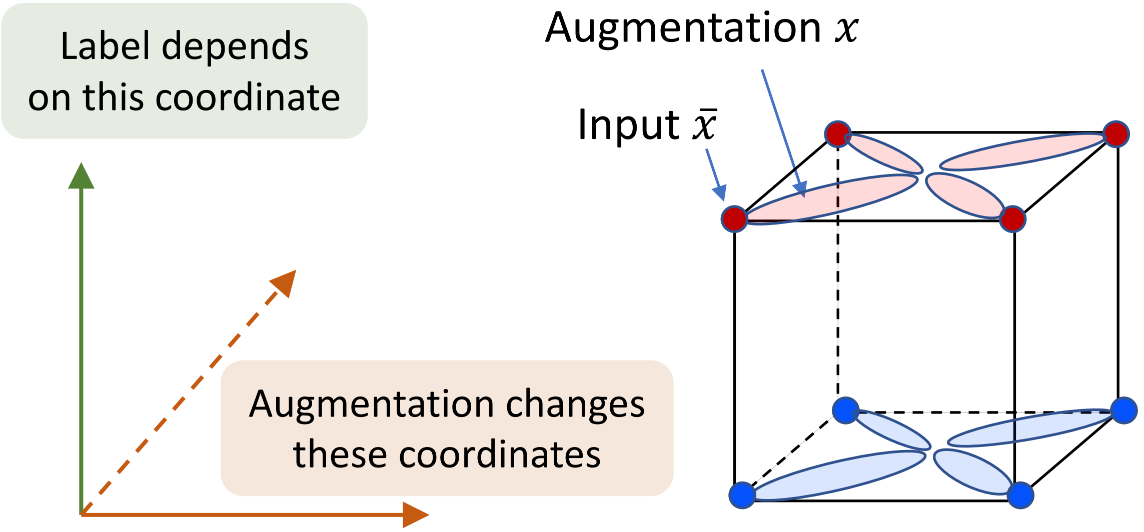
Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b, \;\; \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all



Representation	$L_{cont}(f)$	**Acc (%)**
Linear	5.13	**99.5**
MLP + Adam	5.04	74.1
MLP + Adam + wd		89.5
$\exists\, f$ (spurious)	4.94	**50**

Lower bound

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Lower bound

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b,\; \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the "right" function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Lower bound (general)

Theorem: If **augmentations do not overlap**, then any function class agnostic guarantee for contrastive learning will be vacuous.

Spurious minimizers of $L_{contrast}$ can be constructed

Upper bound

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \leq a\, L_{contrast}(f) + b, \ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Can inductive bias ***sensitive*** analysis explain this success?

Lower bound (general)

Theorem: If **augmentations do not overlap**, then any function class agnostic guarantee for contrastive learning will be vacuous.

Spurious minimizers of $L_{contrast}$ can be constructed

Upper bound

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \leq a\, L_{contrast}(f) + b, \ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Can inductive bias ***sensitive*** analysis explain this success?

Yes! For linear representation class

Lower bound (general)

Theorem: If **augmentations do not overlap**, then any function class agnostic guarantee for contrastive learning will be vacuous.

Spurious minimizers of $L_{contrast}$ can be constructed

Upper bound

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \ \leq \ a\, L_{contrast}(f) + b, \ \forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the "right" function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Can inductive bias ***sensitive*** analysis explain this success?

Yes! For linear representation class

Lower bound (general)

Theorem: If **augmentations do not overlap**, then any function class agnostic guarantee for contrastive learning will be vacuous.

Spurious minimizers of $L_{contrast}$ can be constructed

Function class sensitive guarantees

Theorem: For a linear representation function class, i.e. $\mathcal{F} = \{f(x) = W\, \phi(x)\}$, we have

$$L_{classify}(f) \ \leq \ a(\mathcal{F})\, L_{contrast}(f) + b(\mathcal{F}) \quad \forall f \in \mathcal{F}$$

Only need overlap in the view of $\mathcal{F}$

Inductive biases in practice

Assumption: Augmentation overlap within class

Guarantees: $\boldsymbol{L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b, \;\forall f}$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all

Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Can inductive bias ***sensitive*** analysis explain this success?

Yes! For linear representation class

Effect of inductive biases observable in practice

Inductive biases in practice

Assumption: Augmentation overlap within class

Guarantees: $L_{classify}(f) \;\leq\; a\, L_{contrast}(f) + b, \;\forall f$

Is assumption satisfied in practice?

Not in the train set. Overlap in population? (still open)

Can contrastive learning work without overlap?

Yes! For the ”right” function class but not all

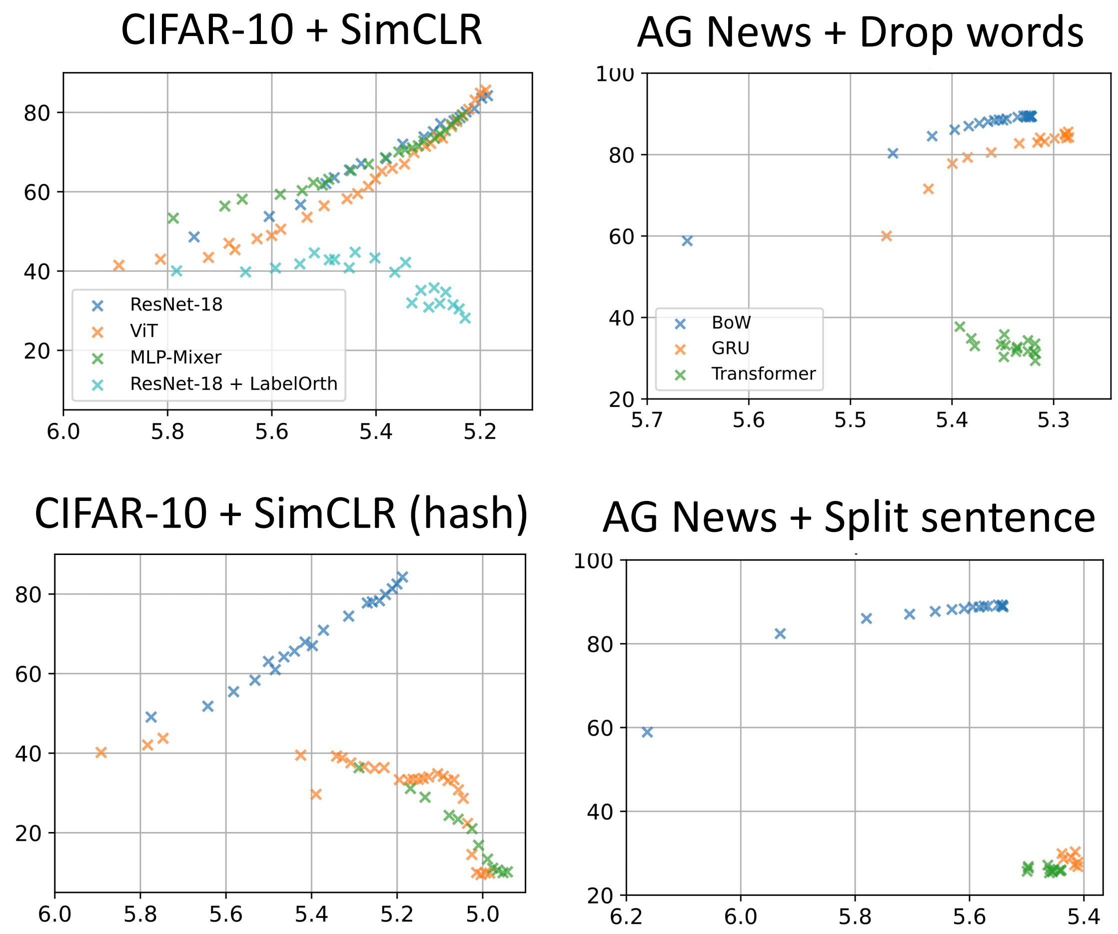
Can inductive bias ***agnostic*** analysis explain this success?

Provably no!

Can inductive bias ***sensitive*** analysis explain this success?

Yes! For linear representation class

Effect of inductive biases observable in practice



For the same augmentation some function classes/algorithms transfer well but others fail miserably