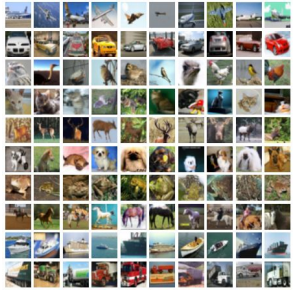
Dataset Condensation for Text

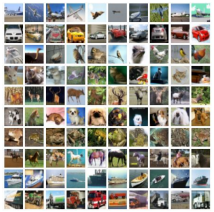
DeepTrojans
Vishesh Agrawal, Shivin Dass, Gaurav Nuti, Rafael Sanchez

What is Dataset Condensation?



cifar - 10

smaller set of
synthesized images



NN

NN

similar
performance

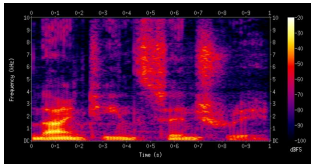
Our Objective: Extend to other Modalities

Text

1. my whole body feels itchy and like its on fire…
2. @Kenichan I dived many times for the ball. Man...

⋮

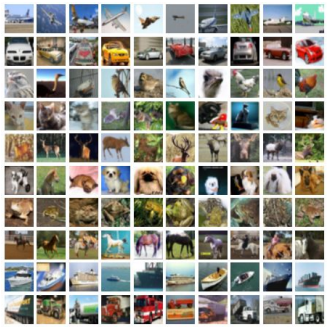
n. is upset that he can't update his Facebook by...



Speech

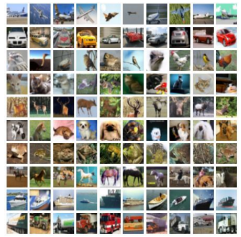
Video

What is **Dataset Condensation**?



cifar - 10

smaller set of
synthesized data



NN

similar
performance

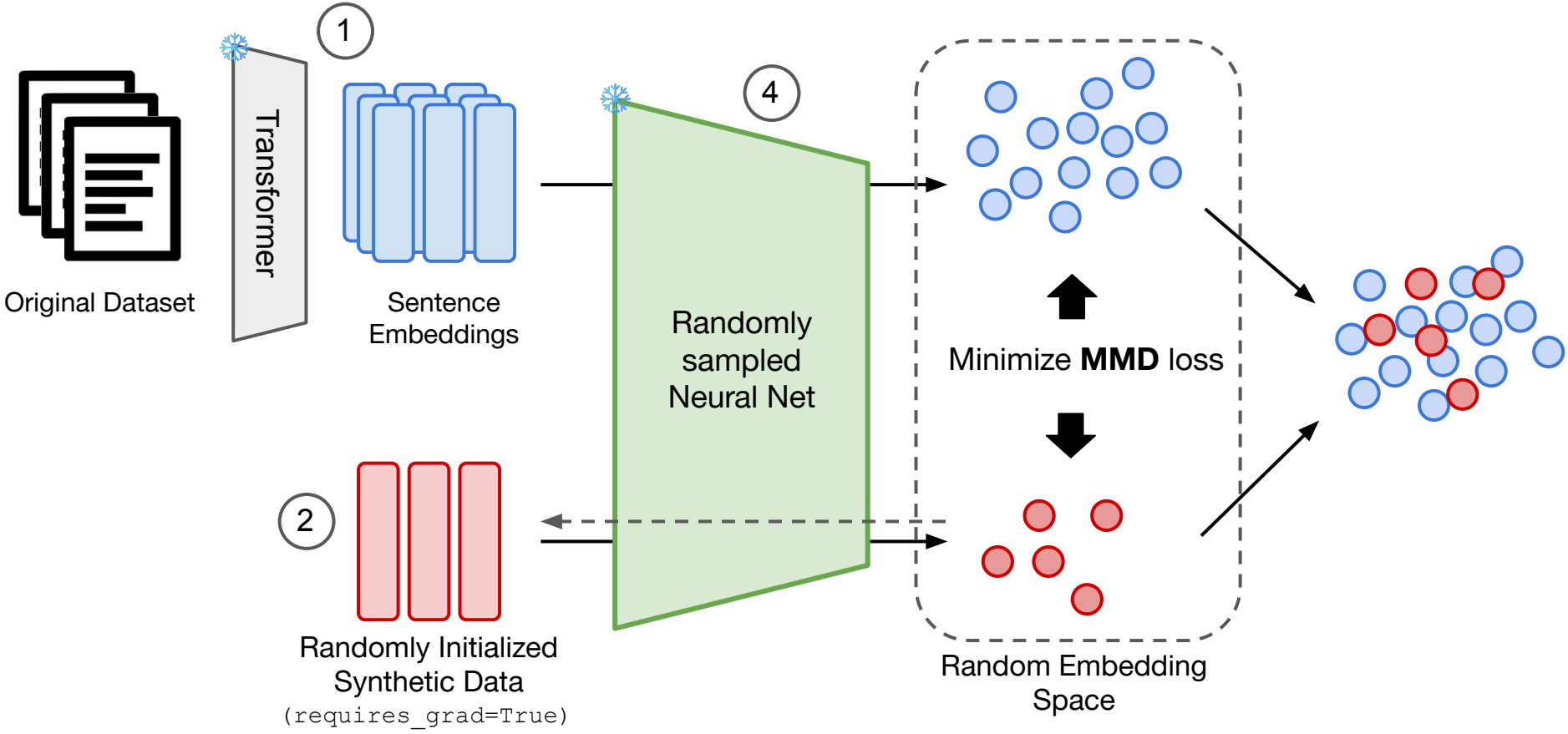
NN

Condense Data

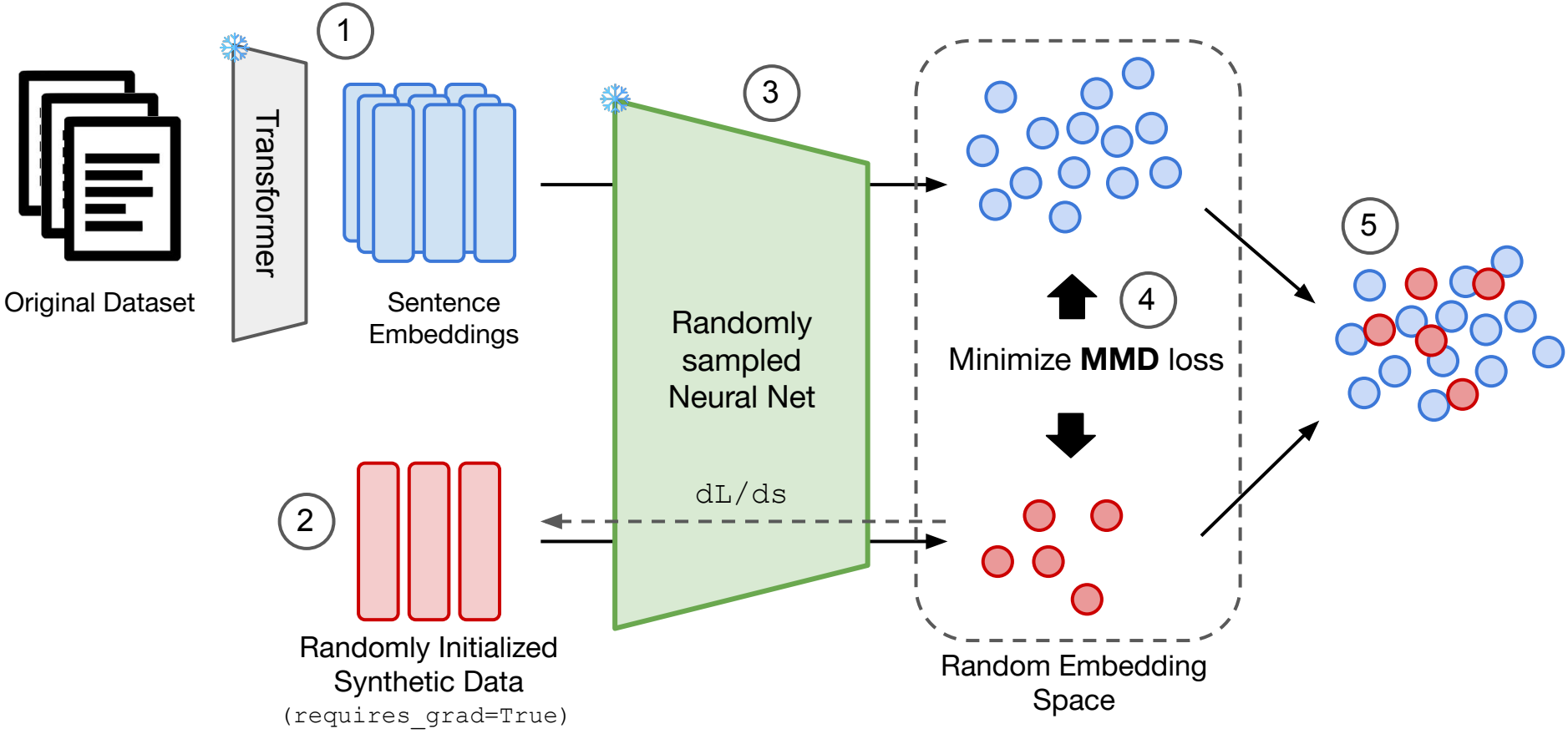
Downstream Task

Our Objective: Extend **condensation** to **text**

We use a **Distribution Matching** approach



We use a **Distribution Matching** approach



Bart Synthesized samples(First 10 neg, Second 10 positive)

'makes a joke out of car chases for an hour, and then gives us half',

"so mind numbingly awful that you hope britney won't do it one",

"a new book in manhattan proves that it's easier to change the sheets than",

'if you don’t have a car, go to the gas station today.',

'time of favor could have given audiences the time to day by concentrating on a culture that',

'a little bit potty training for a young girl in the neighborhood. Read more »',

"its tommy's job to clean the peep booths surrounding her, and after",

"it's so badly made on every level that i'm actually having a hard",

"doesn't get the job done, running off a lot of chemistry created by r",

'woefully pretentious. It’s all about the same thing: overeating',

'extremely well acted by the four primary actors, this is a seriously thought out movie that',

"it's probably worth catching solely on its own, but it's a good one",

"the film, while not exactly assured in its execution, is notable for it's sheer"

Results

Dataset	Samples/class	Synth. Samples/class	Synth. data accuracy (%)	Full data accuracy (%)
SST1	1710	1	47.05±0.47	52
SST2	3460	1	88.54±0.28	91
DBpedia	40000	50	88.71±0.38	98
Yahoo	140000	50	63.2±0.36	72

Synthetically Generated Data for SST2

Positive Synthetic Class

1. **extremely well acted** by the four primary actors, this is a **seriously thought out** movie that
2. **interesting** both as a historical study and an allegory for the relationship between two men
3. **intimate** and **panoramic** view of the world at a moment's notice

Negative Synthetic Class

1. **woefully pretentious**. It's all about the same thing: overeating
2. it's so **badly made on every level** that i'm actually having a hard
3. so **mind numbingly awful** that you hope britney won't do it one

Applications. Continual Learning

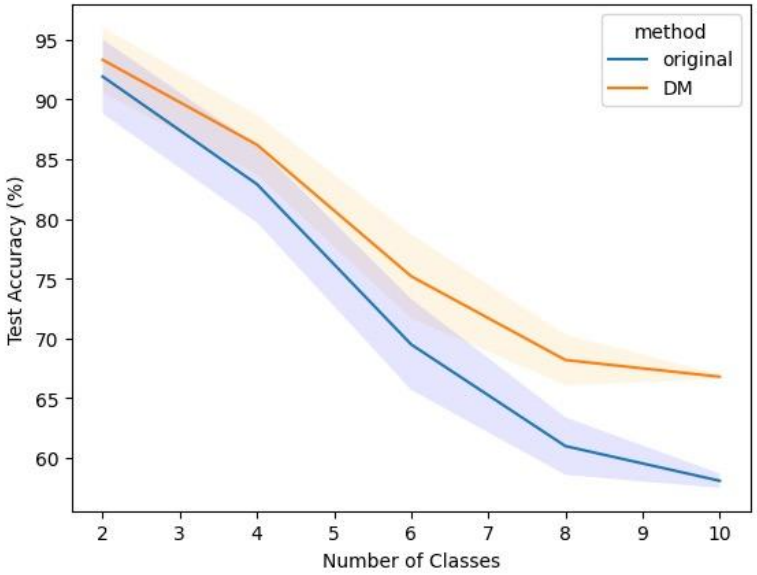


Fig 1**: 5-step class-incremental learning on **Yahoo Dataset

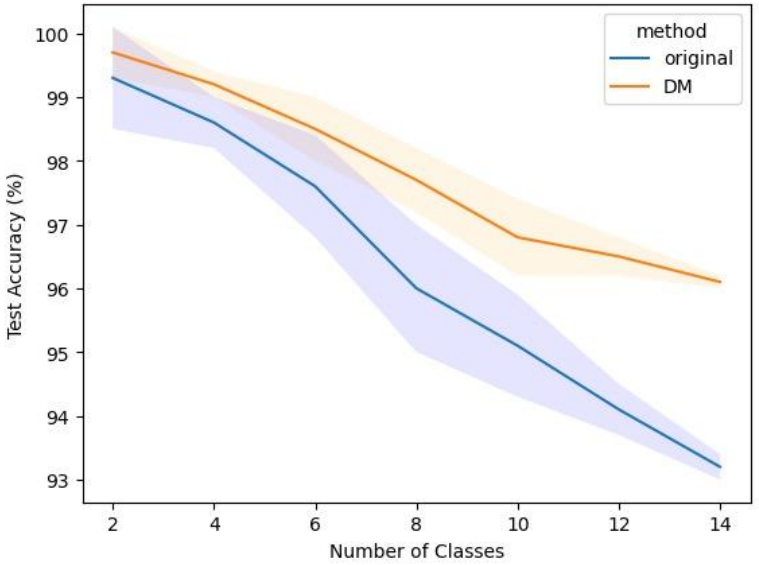
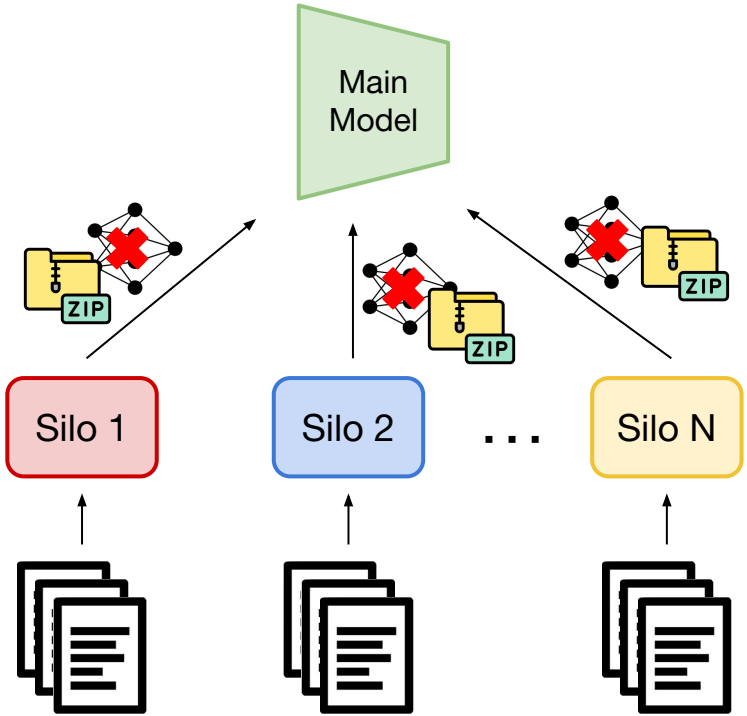
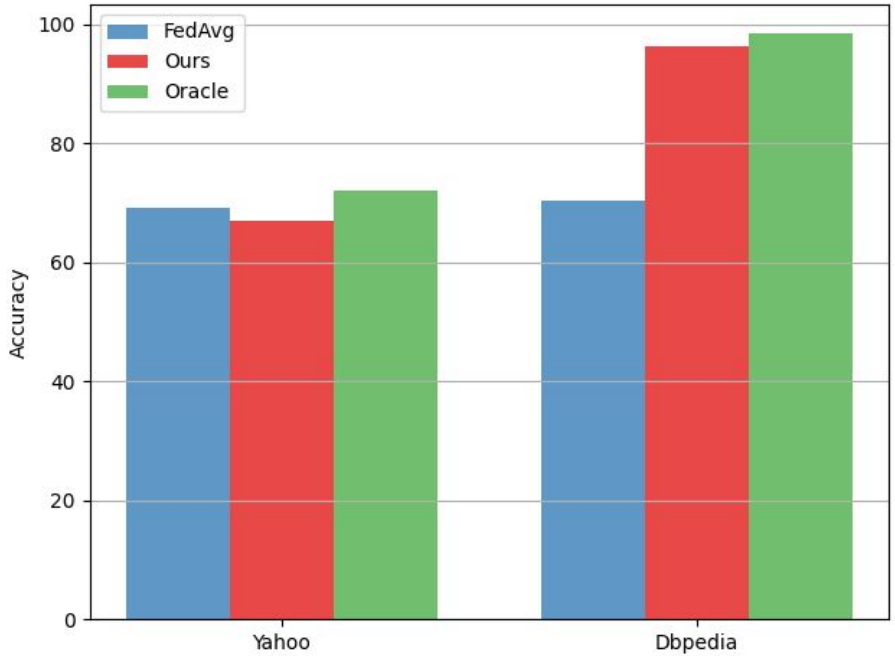


Fig 2**: 7-step class-incremental learning on **Dbpedia Dataset

Applications. Federated Learning





*standard error small so omitted

Future Work

- Evaluate Federated Learning with Dataset Condensation on non-iid data in each silo
- Compare with stronger federated learning baselines
- Work on explainability of the condensed data
- Combine different datasets for a harder continual learning setup