機械学習を用いたファズデータのチェックサム及びハッシュ値の推定

石浦研究室
27016627 藤本高史

背景

- **システムの脆弱性**

  → 徹底的なテストが必要

- **ファジング**
  - 大量のデータ入力でシステムをテスト
  - 変異ベース：既存データの一部を変異

  → 実装が容易かつ汎用性が高い

  → 入力データ通過率が低い

関連研究

➢ [難波2018]

機械学習を用いてランダム文字列に対するチェックサムの推定

– 変異させた文字列に対して正しいチェックサムに変換

– 入力データ通過率向上

→ 推定する文字列が8 byteのみ

→ 汎用性が低い

[難波2018]難波 学之: ``変異ベースファジングのためのチェックサムの機械学習,''
関西学院大学理工学部情報科学科卒業論文 (Mar. 2019)

本研究

8 byte以上の文字列に対する チェックサム及びハッシュ値を推定

- Encoder・Decoderモデルを採用

➢ **実験**

- ランダム, 規則性のある文字列
- 可変長データ

チェックサム

- **誤り検出符号の一種**
- **文字列の総和から特定の値の剰余**

Thank you very much. ➡ 28 + 88 + 129 + … + 223 = 12985

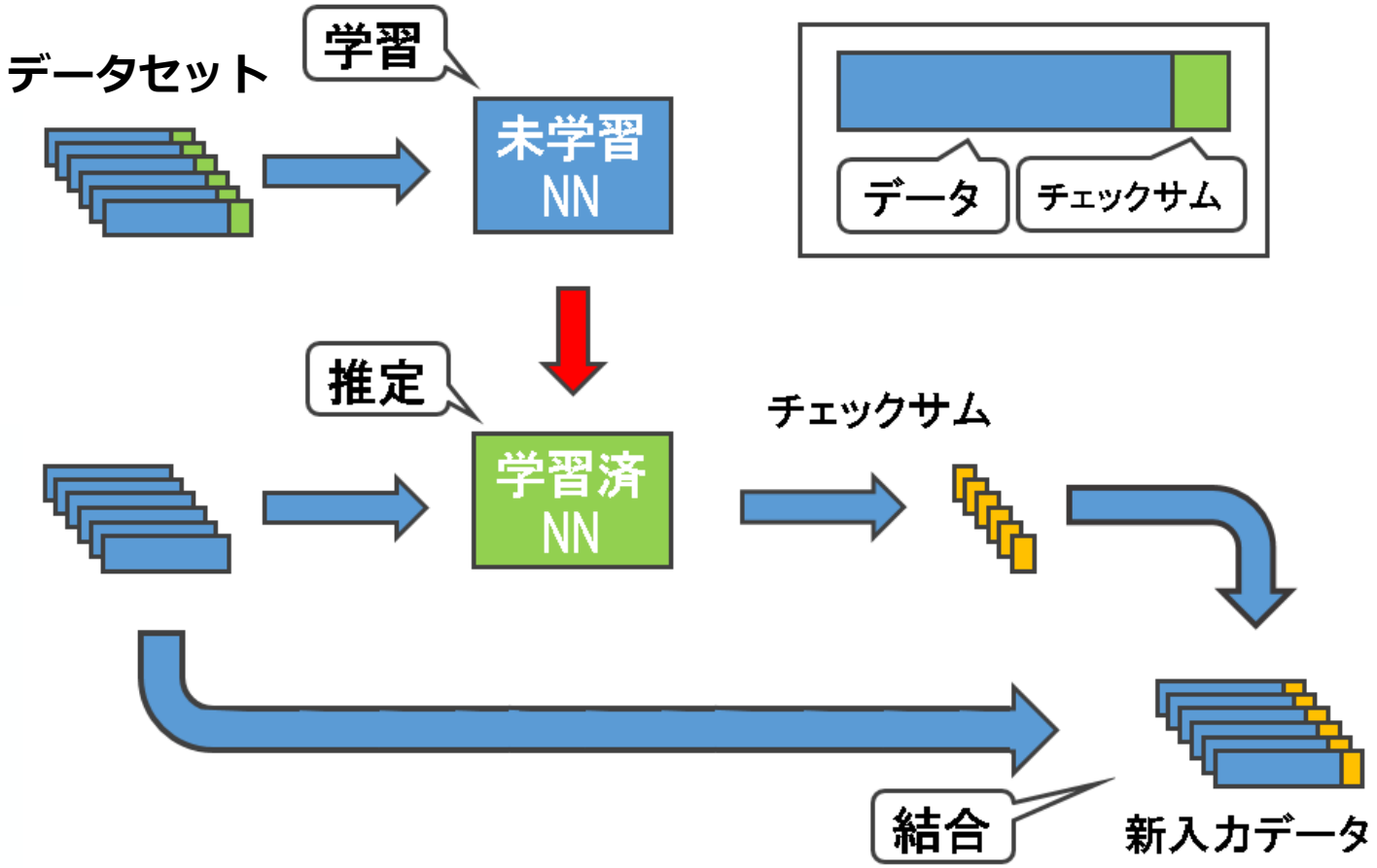
12985 mod 256 = 128

ハッシュ値

- **CRC :**2 進数の文字列, 長さは様々
- **MD5 :**16進数の文字列, 128 bit
- **SHA1:**16進数の文字列, 160 bit

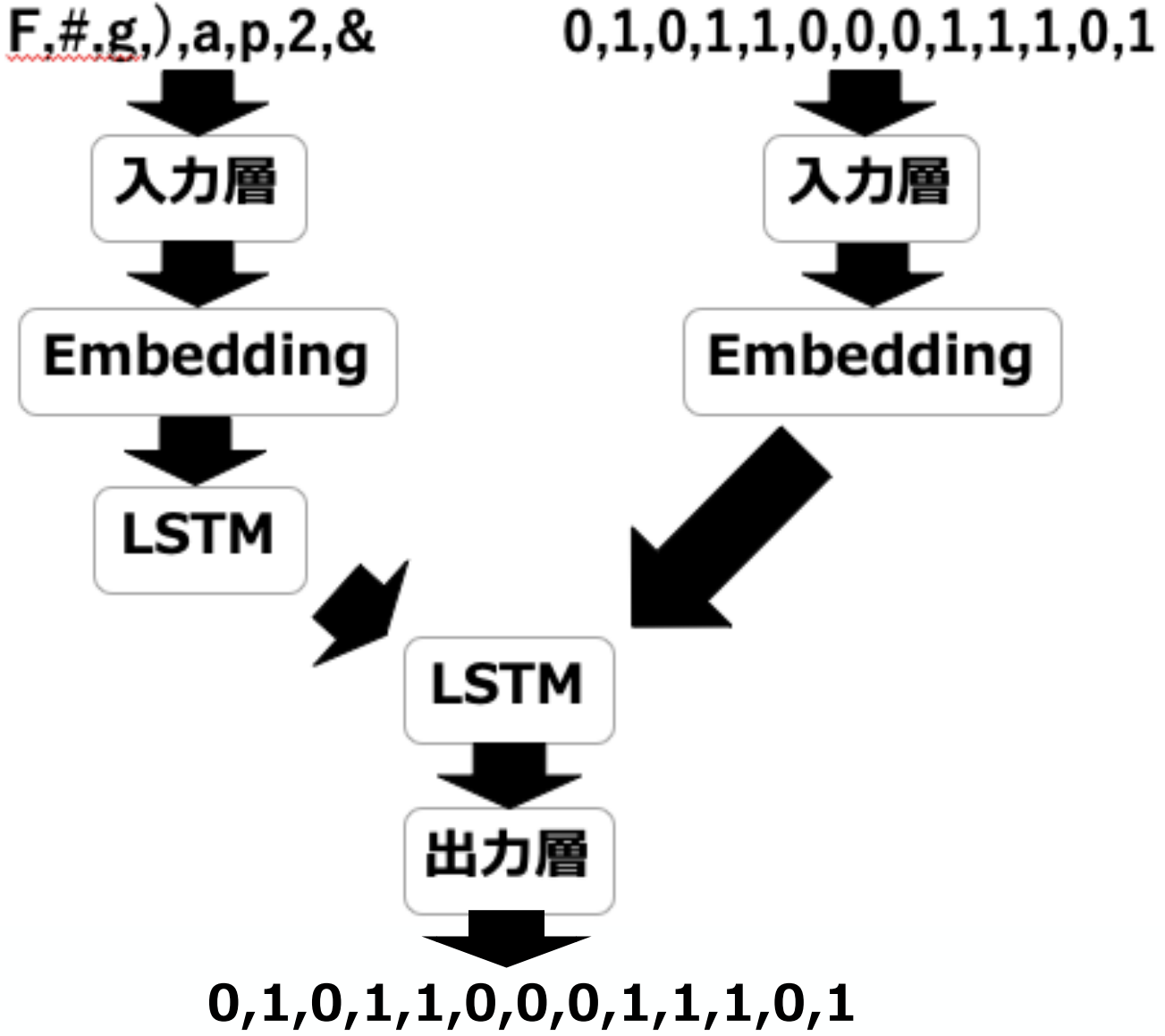
NNによるチェックサムの学習

- [難波2018]



→ 推定する文字列が8 byteのみ
→ 汎用性が低い

Encoder・Decoderモデル



実験

- **ランダム文字列と規則性のある文字列からチェックサム及びハッシュ値を推定**
  - チェックサム: 文字列の総和 mod 256
  - ハッシュ値: CRC16, CRC32, MD5, SHA1

- **データセット**
  - 総データ数: 10 ~ 20 万文書
  - ランダム文字列: 1 ~ 48 byte
  - 規則性の文字列(英文): 2 ~ 32 byte

実験結果

	Train Random	Train English	Random	English
cksum	52%	50%	20%	51%
CRC16	53%	47%	9%	9%
CRC32	54%	50%	12%	4%
MD5	11%	11%	12%	2%
SHA1	14%	19%	5%	11%

むすび

ランダム文字列, 英文に対して
チェックサム, CRC16, CRC32, MD5, SHA1を推定

➢ 今後の課題

- 学習精度向上
- 他のハッシュ値の推定
- 固定長による推定評価
- 実装評価

実験結果

	Train Random	Train English	Random	English
cksum	52%	50%	20%	51%
CRC16	53%	47%	9%	9%
CRC32	54%	50%	12%	4%
MD5	11%	11%	12%	2%
SHA1	14%	19%	5%	11%

Encoder・Decoderモデル

- **自然言語処理系ニューラルネットワーク**
  - 翻訳
  - 文章の要約
  - 対話作成

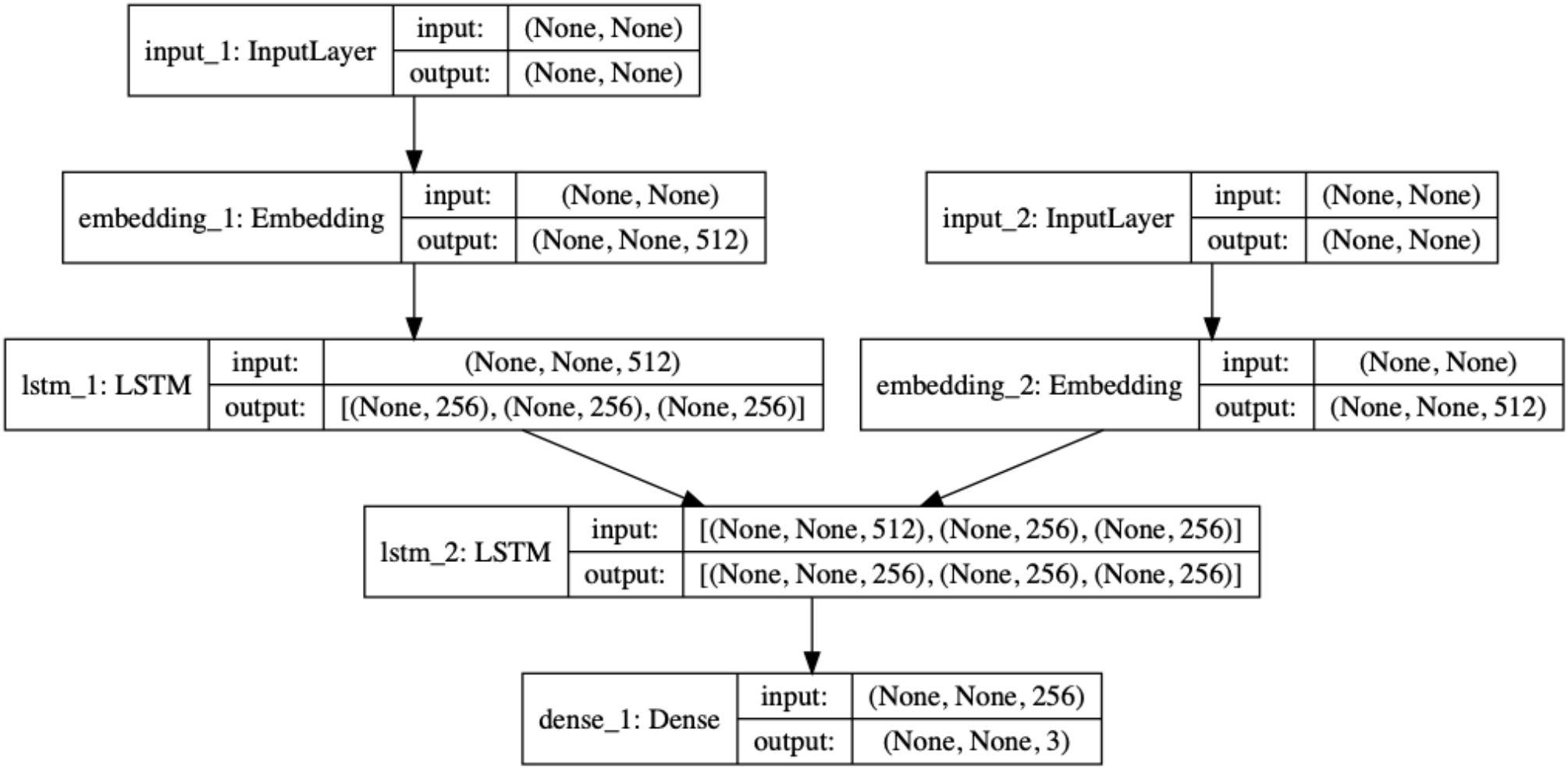
変異ベース

- **既存データの一部を変異**

変異に伴ってチェックサムや
ハッシュ値は変化しない

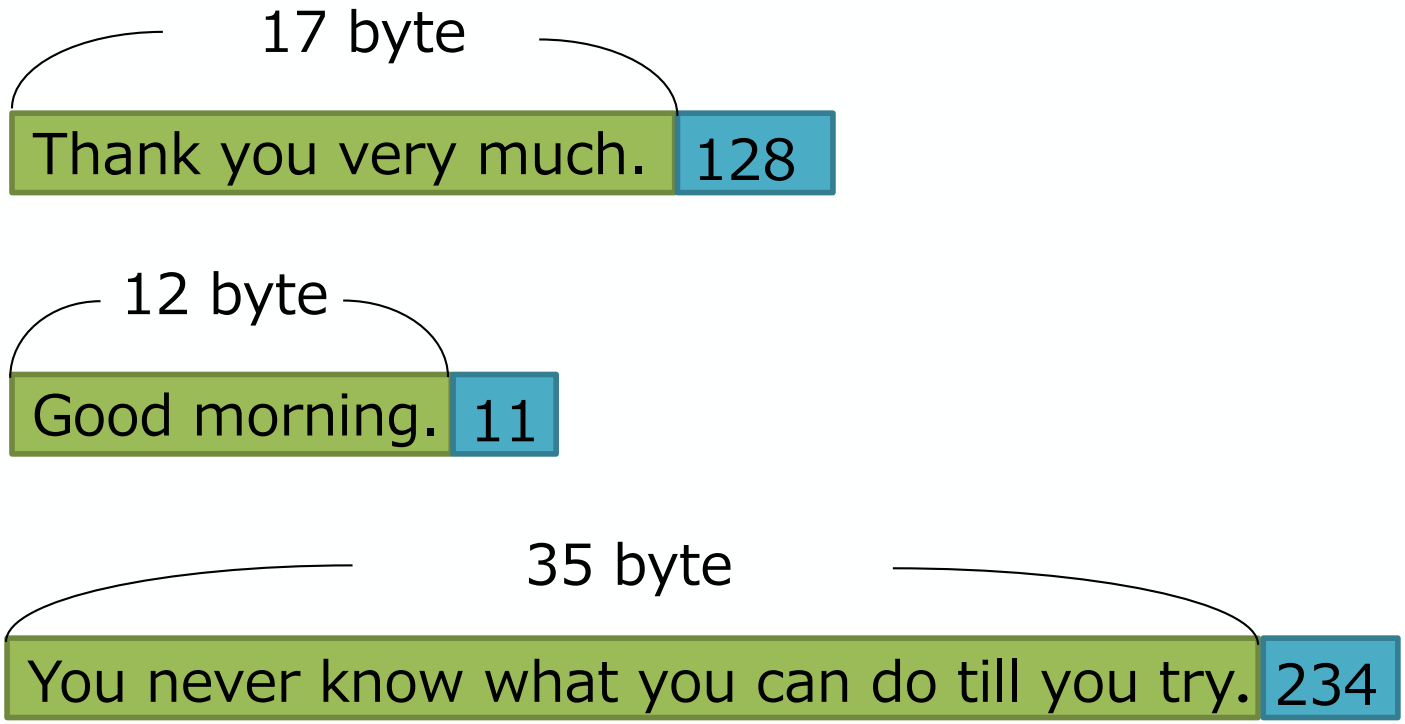
壊れたデータと扱われ，殆ど通過しない

Encoder・Decoderモデル



2. 可変長データ

- 文字列の長さが一定ではない



1. ランダム, 規則性のある文字列

➢ ランダム文字列

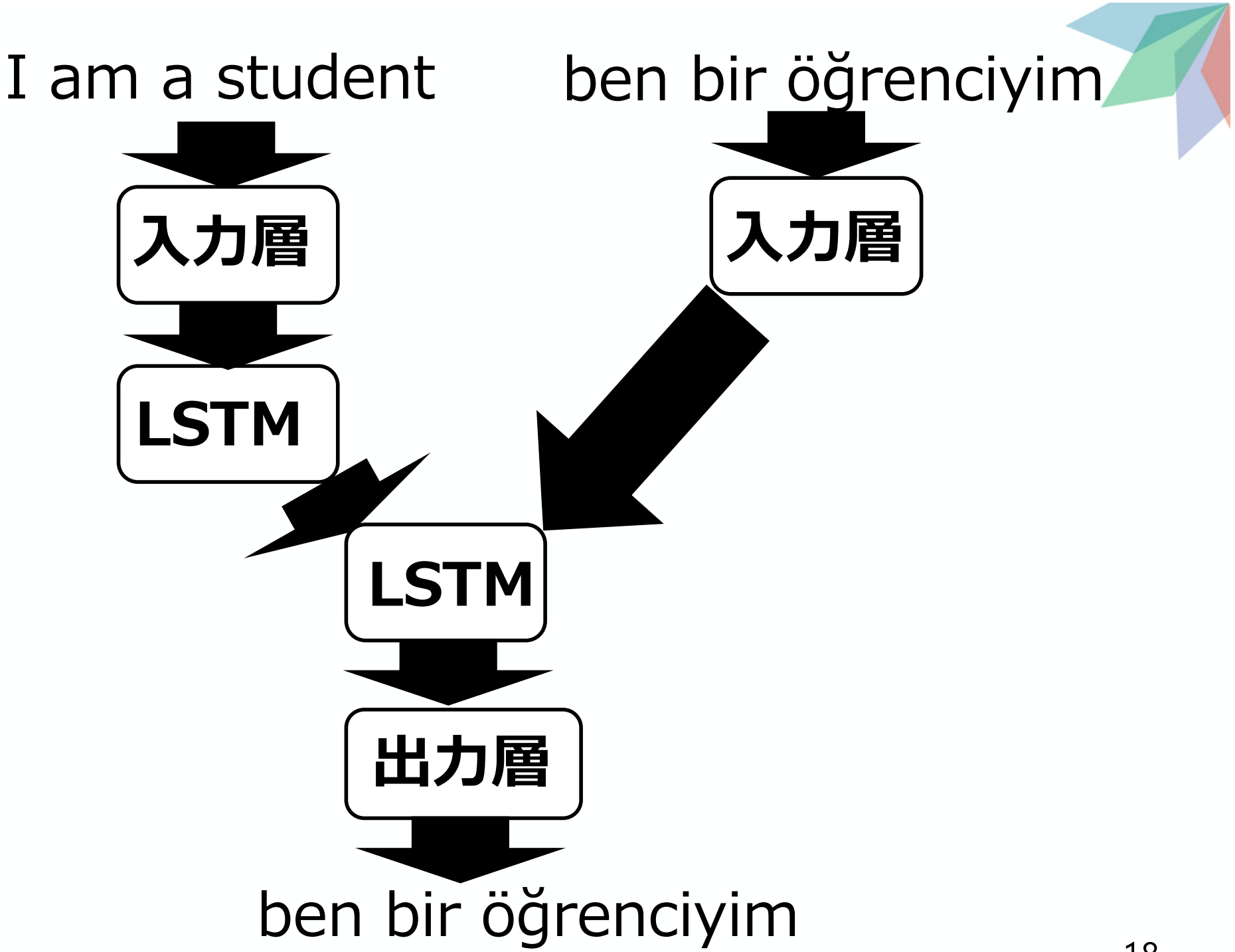
- アルファベット, 記号, 数字
- 関係性を持たない文字列

D21jhg(#”mfnkjea”	55

➢ 規則性のある文字列

- 英語など

Thank you very much.	128


I am a student
ben bir öğrenciyim
入力層
入力層
LSTM
LSTM
出力層
ben bir öğrenciyim

関連研究

- **[難波2018]**
  - 機械学習によるランダム文字列に対するチェックサムの推定
  - 入力データ通過率向上

→ 8 byteの固定長だけの学習

→汎用性が低い

[難波2018]難波 学之: ``変異ベースファジングのためのチェックサムの機械学習,'' 関西学院大学理工学部情報科学科卒業論文, (Mar. 2019)

本研究

- 8 byte以上の文字列に対するチェックサムの推定
- 可変長の入力データに対応
- ハッシュ値の推定

→ 汎用性を高める

Q & A

先行研究と比べて，NN規模はどんな感じなのか.

– NNは大方は同じだが, 大きく異なるのは中間層を変えたところである. また、BatchNormalizationとか、kerasを使ったところである.

• 中間層だけ変えるだけならアホでもできる. それ以外にどこに時間がかかったのか.

– 特に中間層のノード数を何度も変えて, 最適なノード数にするために実験を行ったこと.

• 規則性のある文字列の学習はさせないのか?.

– 規則性のない文字列である程度の学習精度が得られるなら, 規則性のある文字列でもいけるであろ うと自負している. でも実際にやらないとわからないから、やる価値は大いにあり

実験

- **文字列からチェックサム及びハッシュ値を機械学習により推定**

- **実験条件**
  - 入力データ: 8, 16 byteのランダム文字列
  - チェックサム: 文字列の総和 mod 256
  - ハッシュ値: CRC16
  - ツール: Keras
  - ニューラルネットワーク: LSTM
  - データ数: 学習データ120000, テストデータ30000