深層学習による灯謎問題システムの構築

1 はじめに

近年深層学習の発展により、人工知能による人類の創作物 (漫画、小説や) への理解といった分野の研究が盛んでいる。本研究では人類の創作物の一種類「クイズゲーム」に注目し、中国の伝統的クイズゲーム「灯謎 (トウメイ)」を深層学習の手法で解く方法について提案する.

2 灯謎 (トウメイ)

灯謎は, 中国の伝統的クイズ問題である. 質問者は問題を詩や熟語の形で出し, 回答者はそれに回答する. 答えは常に字または単語になる. 灯謎は質問に答えるための問題文以外の文書や知識など必要がないものが多く, 質問の文中から答えの情報を得ることが容易である. つまり, 質問を理解すれば回答できると言える. 灯謎を解くためには, 問題に隠された情報をもとに, 問われている内容を理解して抽出しなければならないので, 灯謎の研究は一種の情報抽出として考えることもできる.

灯謎のパターンは主に謎とヒントと答えで構成される. 謎は詩や熟語や普通の話し言葉で記述された文である. ヒントは答えの形を説明する文である. ヒントは 1 つ以上与えられ, 答えは字か単語である, 問題に隠された字の構成, 発音, 意味などの情報から解くことができる. 図 1 に灯謎の一つの例を示す.

本研究では灯謎問題のうち,「字謎」と呼ばれる, 答えが一つの漢字のみとなる種類のみについて考える. 字謎の答えは, 単語の意味に加えて漢字の形も強く関わるので, 単純に大量の問題の文の情報のみをニューラルネットワークで学習しても効果が薄いと予想される. そこで本研究では漢字の形の情報に着目し, 漢字と SUB 漢字成分を利用した Sequence to Sequence (seq2seq) モデルで灯謎問題を解くシステムを構築した.

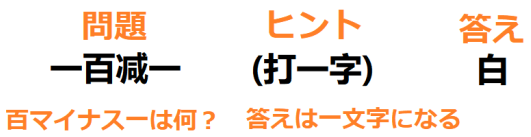


図 1: 灯謎の例

表 1: 問題の種類

問題の総数	文中の字の情報のみで解ける問題	文中の字の情報のみで解けない問題
79725	72937	6788

3 要素技術

3.1 Transformer

Transformer [?] は 2017 年に Google が発表した Encoder-Decoder 構造な言語モデルである. RNN や CNN などを使わず, Self-Attention mechanism のみ使用して, 位置埋め込みの情報から Token(単語や漢字等) の重要性を計算することで, 高速な並列計算の実現が可能になる. 図に Transformer のモデル構造を示す.

3.2 Focal Loss

Focal Loss は Facebook が 2017 に物体検出を対象に提案された損失関数である. 分類クラス間の不均衡である問題を解決するため, Focal Loss は分類が容易なサンプルの重みを下げることで, 分類が困難なサンプルにより焦点をあてる. この方法により, サンプル数が少ないクラスや分類が難しいサンプルに対して学習しやすくなる特徴がある.

4 データセット

4.1 中華灯謎ベース

中華灯謎ベース [?] は, 中国各地の灯謎ファン達が集めた灯謎問題 1,362,911 件を収録したデータセットである.

本研究では灯謎のヒントの文を使わないため, 答えが一文字である問題 79,725 件のみ利用し, 研究用の灯謎のデータセットを構築した.

研究用灯謎データセットについて, 文中の字の情報のみで解ける問題 72,937 件のみを扱った.

表 2 に問題の種類を示す.

表 2: 使用した漢字数と補足したデータ数

使用した漢字数	補足した漢字数
9285	782

4.2 IDS データセット

IDS (Ideographic Description Sequence)[**?**] とは, 中国語, 日本語, 韓国語の漢字データを 'unicode', '漢字', 'サブ漢字' の形で集まったデータセットである.

本研究の対象となる灯謎は中国語で作られたため,IDS の中国語データを利用した. しかし, IDS データセットには漢字の成分不足 (漢字「爽」の「メ」など) という問題がある, そこで本研究で使用したデータ 9285 件の中に漢字の成分が不足であるデータ 782 件のに対して, IDS データセットの「CDP コード」で補足した.

表 1 に本研究に使われている漢字数と手動で SUB 漢字成分を補足した数を示す.

4.3 表意文字記述文字

表意文字記述文字（Ideographic Description Characters）は、日中韓の表意文字を記述するために用いられる図形文字を含む 10 種類の Unicode ブロックである. これらの記述文字を利用することで, 漢字の構造を判別することが可能になる.

表に本研究で使用した表意文字記述文字件の件数を示す.

5 提案手法

データ不均衡の影響を解決するために, 本研究は Focal Loss を導入した Transformer の手法を提案する.

6 実験

6.1 データ処理

本実験では問題文中の字の情報のみで解ける問題で作成した研究用の灯謎データ 72,937 件を利用した. その中からランダムに Train Data 58,350 件, Valid Data 7,293 件, Test Data 7,294 件を抽出し, 実験に使用した.

表 3 に資料のデータ数を示す.

表 3: 資料のデータ数

Train Data	Valid Data	Test Data
58350	7293	7294

6.2 実験内容

本研究は Focal Loss が不均衡データセットに対する影響を確認するため, 問題を「構造＋成分」の Tokenize し, 損失関数が「Focal Loss」と「Cross Entropy」二つの状況で, 対照実験をした.

6.3 実験結果

実験中

7 まとめと今後の課題

データ分析

漢字構造	問題	答え
⿰	203411	50601
⿱	354559	66461
⿲	2954	553
⿳	14164	3725
⿴	11906	1891
⿵	23296	3121
⿶	4265	589
⿷	1322	400
⿸	36304	5759
⿹	13960	3482
⿺	23411	2350
⿻	39574	5435
分けられない	99228	1744

データ分析

Question

问题字数	计数	问题字数	计数	问题字数	计数	问题字数	计数	问题字数	计数
1	9	10	4325	19	19	28	75	40	7
2	72	11	496	20	133	29	5	42	3
3	349	12	742	21	15	30	8	44	2
4	2069	13	201	22	17	32	4	47	1
5	7915	14	1030	23	9	33	3	50	1
6	2894	15	62	24	56	34	2	57	1
7	45129	16	131	25	2	35	3	58	1
8	5665	17	32	26	6	36	4		
9	1387	18	48	27	2	39	2		

Trainformer による実験結果

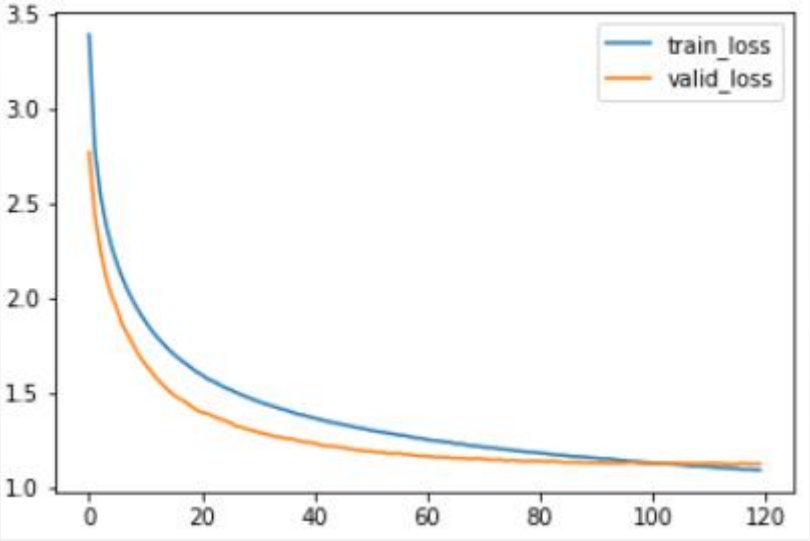
Loss

Loss Function	Data	
	Train	Valid
CE	1.089	1.121
Focal Loss	0.148	0.163

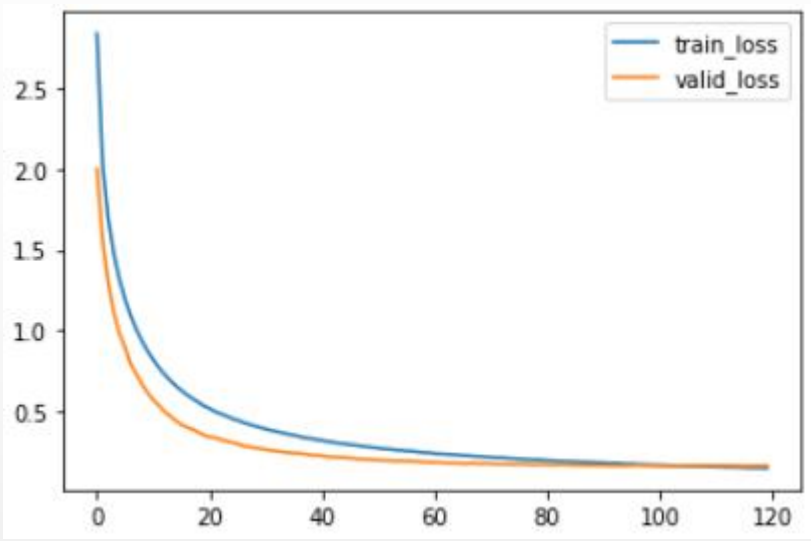
Precision

Loss Function	Precision		Recall		F1	
	Valid	Test	Valid	Test	Valid	Test
CE	0.264	0.267	0.199	0.241	0.203	0.239
Focal Loss	0.245	0.311	0.221	0.255	0.221	0.264

実験結果

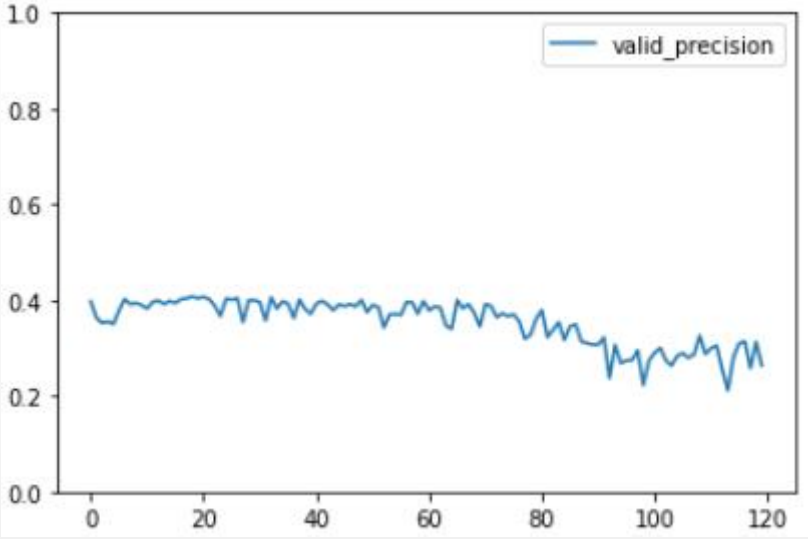


Cross Entropy

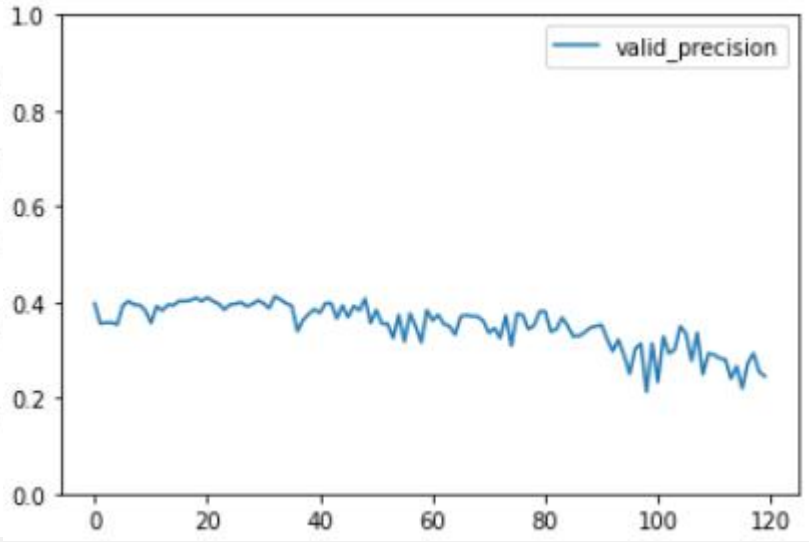


Focal Loss

実験結果

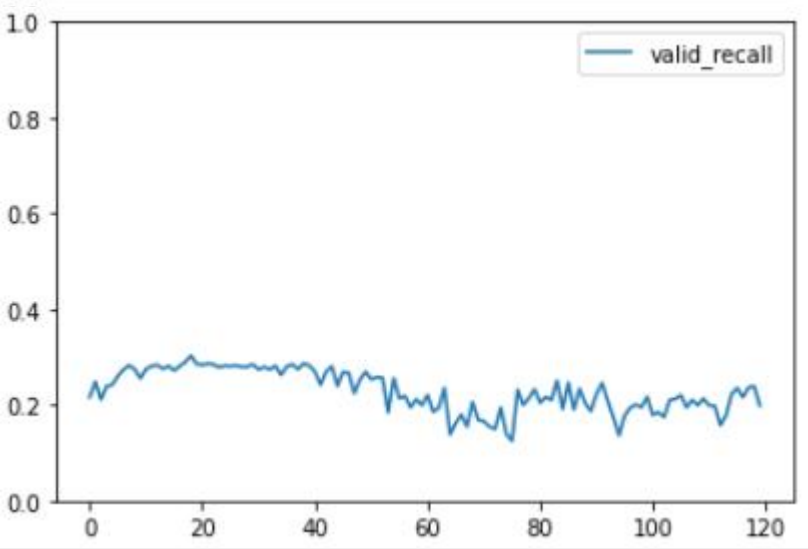


Cross Entropy

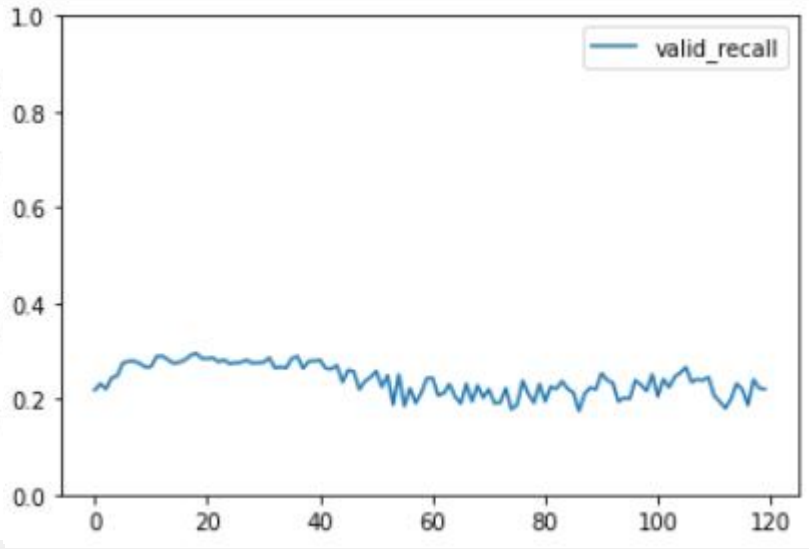


Focal Loss

実験結果

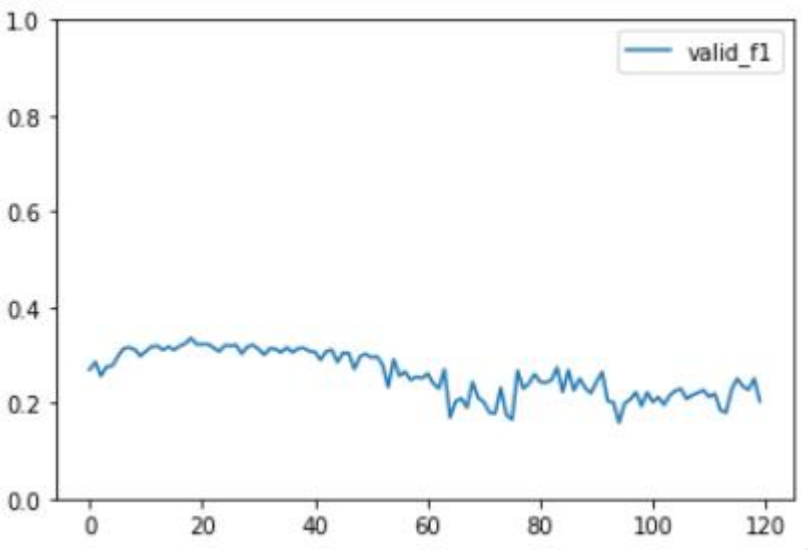


Cross Entropy

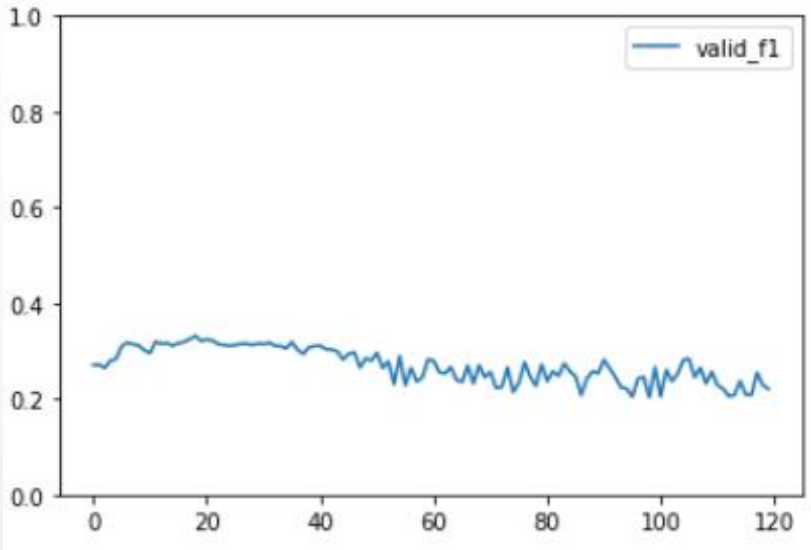


Focal Loss

実験結果



Cross Entropy



Focal Loss